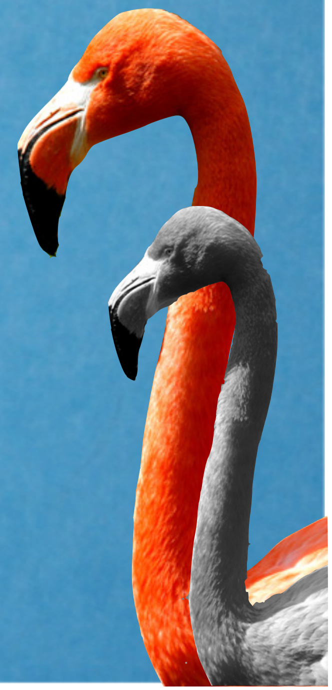
CHAPTER 11
Inference for Distributions of Categorical Data

11.2b

Inference for Two-Way Tables



The Practice of Statistics, 5th Edition
Starnes, Tabor, Yates, Moore

Bedford Freeman Worth Publishers

Inference for Two-Way Tables

Learning Objectives

After this section, you should be able to:

- ✓ COMPARE conditional distributions for data in a two-way table.
- ✓ STATE appropriate hypotheses and COMPUTE expected counts for a chi-square test based on data in a two-way table.
- ✓ CALCULATE the chi-square statistic, degrees of freedom, and *P*-value for a chi-square test based on data in a two-way table.
- ✓ PERFORM a chi-square test for homogeneity.
- ✓ **PERFORM a chi-square test for independence.**
- ✓ **CHOOSE the appropriate chi-square test.**

Chi-Square Test for Homogeneity

Chi-Square Test for Homogeneity

Suppose the 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 count for each category assuming that H_0 is true. 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 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 based on two categorical variables.

In that case, our goal is to analyze the relationship between the variables.

Our null hypothesis is that there is no association between the two categorical variables in the population of interest.

The alternative hypothesis is that there is an association between the variables.

Finger length

Is there a relationship between gender and relative finger length? In Chapter 5, we looked at a sample of 452 U.S. high school students who completed a survey. The two-way table shows the gender of each student and which finger was longer on their left hand (index finger or ring finger).

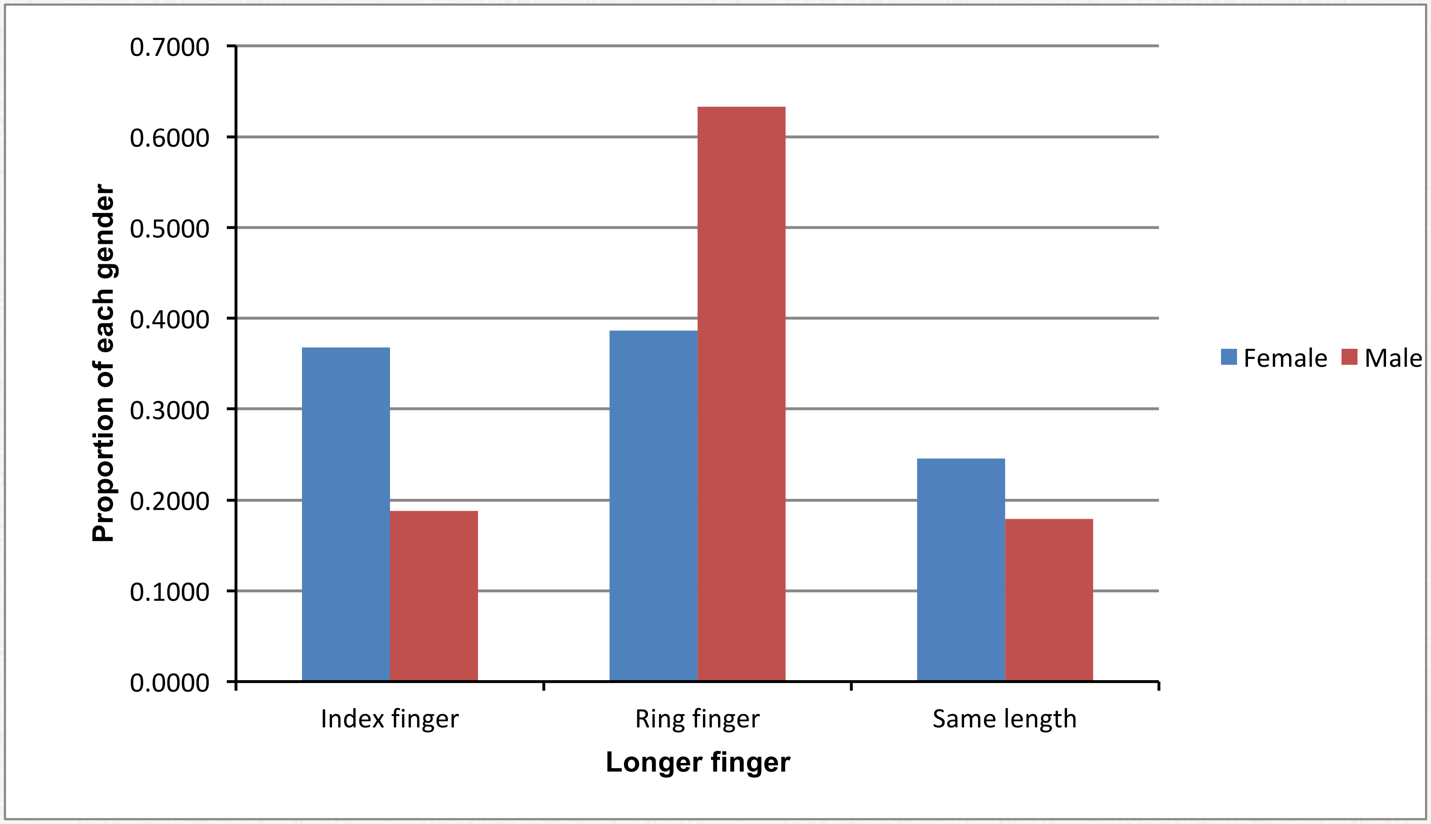
	Female	Male	**Total**
Index finger	78	45	**123**
Ring finger	82	152	**234**
Same length	52	43	**95**
Total	**212**	**240**	**452**

a) Is this an observational study or an experiment? Justify your answer.

This is an observational study. Gender was not randomly assigned to the members of the sample.

b) Make a well-labeled bar graph that compares the distribution of finger length for females and males. Describe what you see.

Finger length



A higher proportion of females had longer index fingers compared to males, while a higher proportion of males had longer ring fingers. A slightly higher proportion of females had index fingers and ring fingers of the same length.

Finger length

We want to test hypotheses

H_0: There is no association between gender and relative finger length in the population of U.S. high school students who filled out the CensusAtSchool survey. *(instead of no association we can say are independent)*

H_a: There is an association between gender and relative finger length in the population of U.S. high school students who filled out the CensusAtSchool survey. *(instead of association we can say are not independent)*

If the null hypothesis is true, then P(longer index finger | female) = P(longer index finger | male) = P(longer index finger).

P(longer index finger) = 123/452 = 0.272,

the expected counts for females is then

212(.272) = 57.7

the expected count for males is then

240(.272) = 65.3

	Female	Male	**Total**
Index finger	78	45	**123**
Ring finger	82	152	**234**
Same length	52	43	**95**
Total	**212**	**240**	**452**

Find all expected counts using $\frac{\text{row total} \cdot \text{column total}}{\text{table total}}$

The Chi-Square Test for Independence

The 10% and Large Counts conditions for the chi-square test for independence are the same as for the homogeneity test.

There is a slight difference in the Random condition for the two tests: a test for independence uses data from one sample but a test for homogeneity uses data from two or more samples/groups.

Conditions for Performing a Chi-Square Test for Independence

- **Random**: The data come a well-designed random sample or from a randomized experiment.
  - **10%**: When sampling without replacement, check that

    $n \leq (1/10)N$.
- **Large Counts**: All *expected* counts are greater than 5

Chi-Square Test for Independence

Chi-Square Test for Independence

Suppose the conditions are met. You can use the **chi-square test for 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.

Start by finding the expected count for each category assuming that H_0 is true. 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.

Finger length

Here is a complete table of observed counts and expected counts (in parentheses):

	Female	Male	**Total**
Index finger	78 (57.7)	45 (65.3)	**123**
Ring finger	82 (109.8)	152 (124.3)	**234**
Same length	52 (44.5)	43 (50.4)	**95**
Total	**212**	**240**	**452**

Do the data provide convincing evidence of an association between gender and relative finger length for U.S. high school students who filled out the CensusAtSchool survey?

STATE

We want to perform a test of the following hypotheses using $\alpha = 0.05$:

H_0: There is no association between gender and relative finger length in the population of U.S. high school students who filled out the CensusAtSchool survey.

H_a: There is an association between gender and relative finger length in the population of U.S. high school students who filled out the CensusAtSchool survey.

Finger length

PLAN

If conditions are met, we will perform a chi-square test for independence.

- Random: The sample was randomly selected.
  - 10%: The sample of 452 students is less than 10% of all U.S. students who filled out the survey.
- Large Counts: The expected counts (see table) are all at least 5.

	Female	Male	**Total**
Index finger	78 (57.7)	45 (65.3)	**123**
Ring finger	82 (109.8)	152 (124.3)	**234**
Same length	52 (44.5)	43 (50.4)	**95**
Total	**212**	**240**	**452**

DO

- Test statistic: $\chi^2 = \frac{(78-57.7)^2}{57.7} + \cdots 29.0$
- P-value: Using (3 – 1)(2 – 1) = 2 degrees of freedom, the P-value is less than 0.0005. Using technology: P-value ≈ 0.

Finger length

CONCLUDE

Because the *P*-value of approximately 0 is less than $\alpha = 0.05$, we reject H_0. We have convincing evidence that there is an association between gender and relative finger length in the population of U.S. high school students who filled out the CensusAtSchool survey.

Using Chi-Square Tests Wisely

Three different chi-square tests—think about how the data was collected

- Goodness of fit: one variable in one population
  - M&M's (compared to a specified distribution)
- Homogeneity: one variable in two or more populations or groups (two or more independent random samples or treatment groups in a randomized experiment)
  - music and entrée choice
  - Note that either the row or column totals were determined by the researcher collecting data
- Independence: two variables in one population
  - gender and finger length
  - Note that both the row totals and the column totals are random

If you really don't know…just say "chi-square test" rather than choosing the wrong one, BUT you are expected to recognize the difference.

Example: Online social networking

An article in the *Arizona Daily Star* (April 9, 2009) included the following table:

Age (years):	18–24	25–34	35–44	45–54	55–64	65+	Total
Use online social networks:	137	126	61	38	15	9	**386**
Do not use online social networks:	46	95	143	160	130	124	**698**
Total:	**183**	**221**	**204**	**198**	**145**	**133**	**1084**

Suppose that you decide to analyze these data using a chi-square test. However, without any additional information about how the data were collected, it isn't possible to know which chi-square test is appropriate.

(a) Explain how you know that a test for goodness of fit is *not* appropriate for analyzing these data.
(b) Describe how these data could have been collected so that a test for homogeneity is appropriate.
(c) Describe how these data could have been collected so that a test for independence is appropriate.

Example: Online social networking

(a) Because there are either two variables or two or more populations, a test for goodness of fit is not appropriate. Tests for goodness of fit are appropriate only when analyzing the distribution of one variable in one population.

(b) To make a test for homogeneity appropriate, we would need to take six independent random samples, one from each age category, and then ask every person whether or not they use online social networks. Or we could take two independent random samples, one of online social network users and one of people who do not use online social networks, and ask every member of each sample how old they are.

(c) To make a test for independence appropriate, we would take one random sample from the population and ask every member about their age and whether or not they use online social networks. This seems like the most reasonable method for collecting the data, so a test of independence is probably the best choice. But we can't know for sure unless we know how the data were collected.

Using Chi-Square Tests Wisely

What if we want to compare two proportions?

- A chi-square test for a 2 by 2 table, you can also use a two sample z test for difference in proportions.
- Chi-square test is always two-sided (so only checks for a difference in proportions rather than greater or less)
- If you want to estimate the difference between proportions, use a two-sample z interval. There are no confidence intervals that correspond to chi-square tests
- If comparing more than two treatments or the response variable has more than two categories, you must use chi-square test
- Ch 10 methods for comparing two proportions when given the choice is recommended (ability to perform one-sided tests and construct confidence intervals)

Using Chi-Square Tests Wisely

Grouping quantitative data into categories

- Grouping together intervals of values
- Imagine two schools with a mean AP Statistics score of 3.
- Rather than comparing means it may provide more information to consider the distributions:

Score	School A	School B
5	10	1
4	5	5
3	1	10
2	5	5
1	10	1

- Be careful not to use too few categories when converting.

Using Chi-Square Tests Wisely

What can we do if the expected cell counts aren't all at least 5?

- Combine two or more rows or columns

Be able to interpret computer output.

Chi-Square Test: Male, Female

Expected counts are printed below observed counts

Chi-Square contributions are printed below expected counts

	Male	Female	Total
1	7	29	36
	11.69	24.31	
	1.883	0.906	
2	31	50	81
	26.31	54.69	
	0.837	0.403	
Total	38	79	117

Chi-Sq = 4.028, DF = 1, P-Value = 0.045

Inference for Two-Way Tables

Section Summary

In this section, we learned how to…

- ✓ COMPARE conditional distributions for data in a two-way table.
- ✓ STATE appropriate hypotheses and COMPUTE expected counts for a chi-square test based on data in a two-way table.
- ✓ CALCULATE the chi-square statistic, degrees of freedom, and *P*-value for a chi-square test based on data in a two-way table.
- ✓ PERFORM a chi-square test for homogeneity.
- ✓ PERFORM a chi-square test for independence.
- ✓ CHOOSE the appropriate chi-square test.

- ✓ Read p. 711-721 ccc 41, 43, 45, 47, 49, 51-55