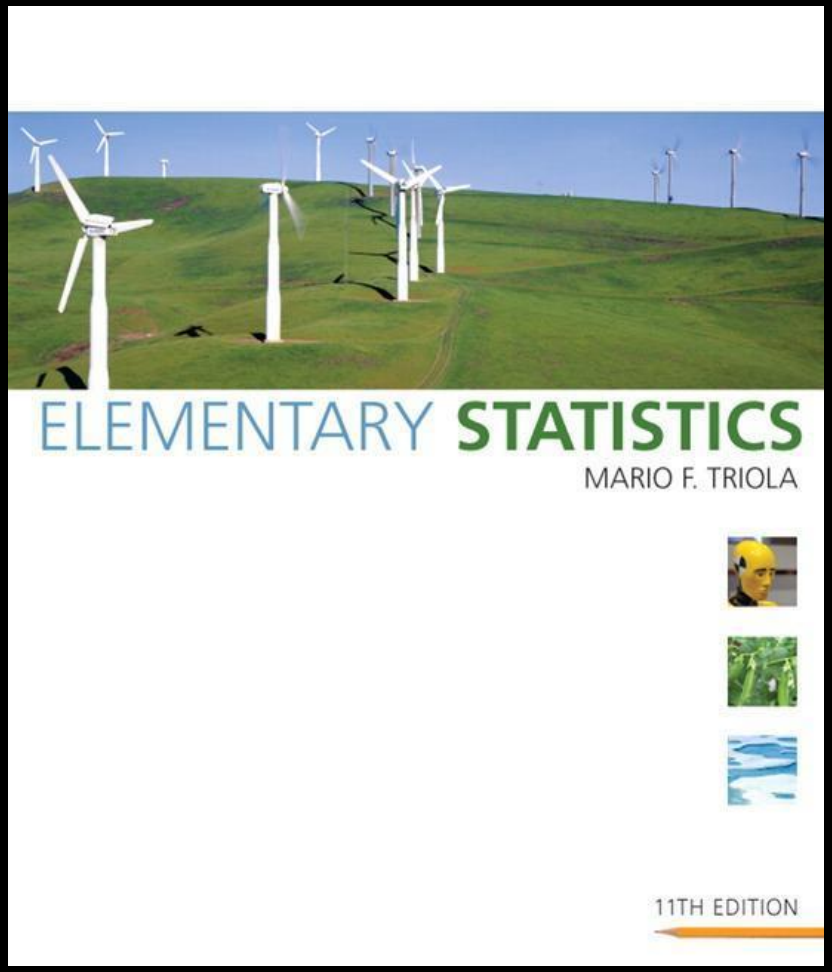
Lecture Slides



Elementary Statistics
Eleventh Edition

and the Triola Statistics Series

by Mario F. Triola

PEARSON

Chapter 11
Goodness-of-Fit and Contingency Tables

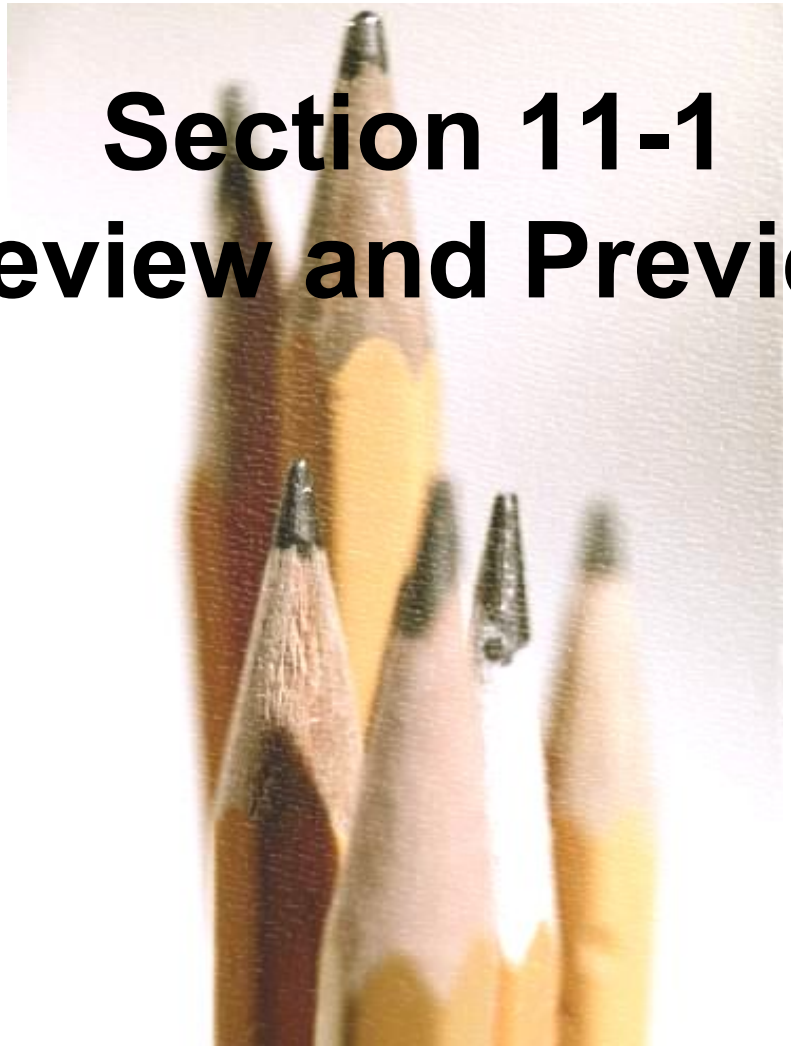
11-1 Review and Preview

11-2 Goodness-of-fit

11-3 Contingency Tables

11-4 McNemar's Test for Matched Pairs

Section 11-1
Review and Preview



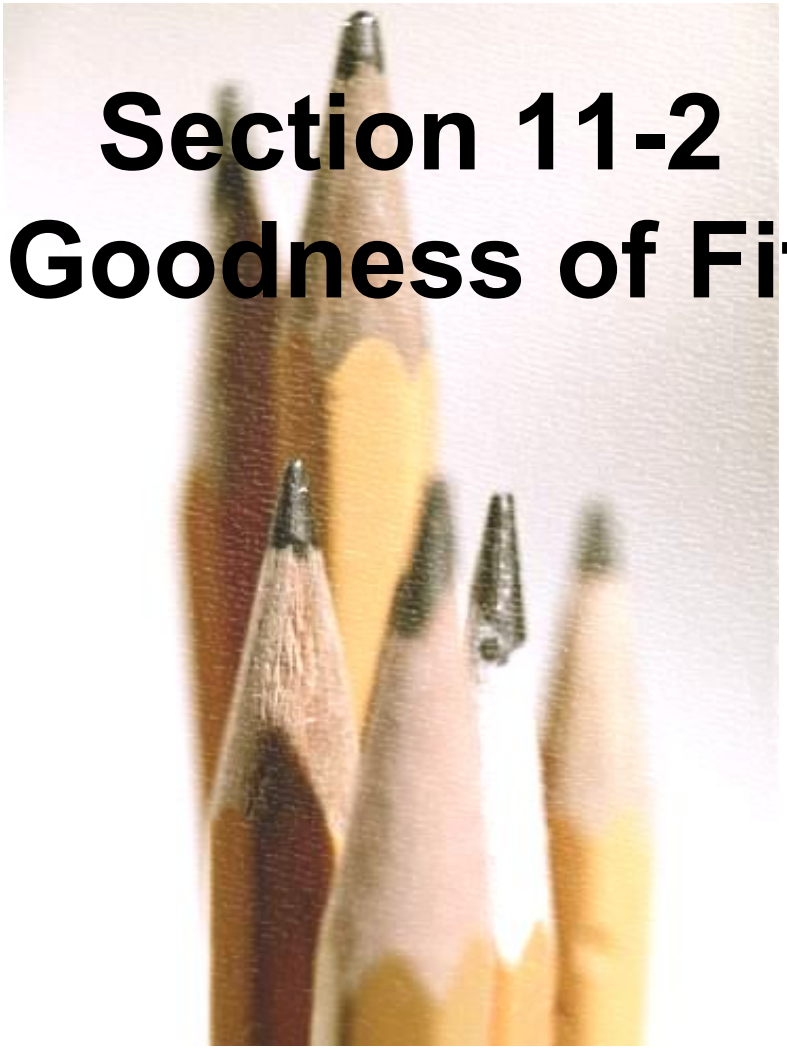
Review

We began a study of inferential statistics in Chapter 7 when we presented methods for estimating a parameter for a single population and in Chapter 8 when we presented methods of testing claims about a single population. In Chapter 9 we extended those methods to situations involving two populations. In Chapter 10 we considered methods of correlation and regression using paired sample data.

Preview

- We focus on analysis of categorical (qualitative or attribute) data that can be separated into different categories (often called cells).
- Hypothesis test: Observed counts agree with some claimed distribution.
- The contingency table or two-way frequency table (two or more rows and columns).
- Two-way tables involving matched pairs.
- Use the χ^2 (chi-square) distribution.

Section 11-2
Goodness of Fit



Key Concept

In this section we consider sample data consisting of observed frequency counts arranged in a single row or column (called a one-way frequency table). We will use a hypothesis test for the claim that the observed frequency counts agree with some claimed distribution, so that there is a good fit of the observed data with the claimed distribution.

Definition

A goodness-of-fit test is used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution.

Goodness-of-Fit Test

Notation

O represents the observed frequency of an outcome.

E represents the expected frequency of an outcome.

k represents the number of different categories or outcomes.

n represents the total number of trials.

Goodness-of-Fit Test

Requirements

1. **The data have been randomly selected.**
2. **The sample data consist of frequency counts for each of the different categories.**
3. **For each category, the expected frequency is at least 5. (The expected frequency for a category is the frequency that would occur if the data actually have the distribution that is being claimed. There is no requirement that the *observed* frequency for each category must be at least 5.)**

Goodness-of-Fit

Test Statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Goodness-of-Fit

Critical Values

1. Found in Table A- 4 using k – 1 degrees of freedom, where k = number of categories.
2. Goodness-of-fit hypothesis tests are always *right-tailed*.

Goodness-of-Fit

P-Values

***P*-values are typically provided by computer software, or a range of *P*-values can be found from Table A-4.**

Expected Frequencies

If all expected frequencies are <u>equal</u>:

$$E = \frac{n}{k}$$

the sum of all observed frequencies divided by the number of categories

Expected Frequencies

If expected frequencies are <u>not all equal</u>:

$$E = np$$

Each expected frequency is found by multiplying the sum of all observed frequencies by the probability for the category.

Goodness-of-Fit Test

- A close agreement between observed and expected values will lead to a small value of χ^2 and a large P-value.
- A large disagreement between observed and expected values will lead to a large value of χ^2 and a small P-value.
- A significantly large value of χ^2 will cause a rejection of the null hypothesis of no difference between the observed and the expected.

Goodness-of-Fit Test

“If the P is low, the null must go.”

(If the P-value is small, reject the null hypothesis that the distribution is as claimed.)

Relationships Among the χ^2 Test Statistic, P-Value, and Goodness-of-Fit

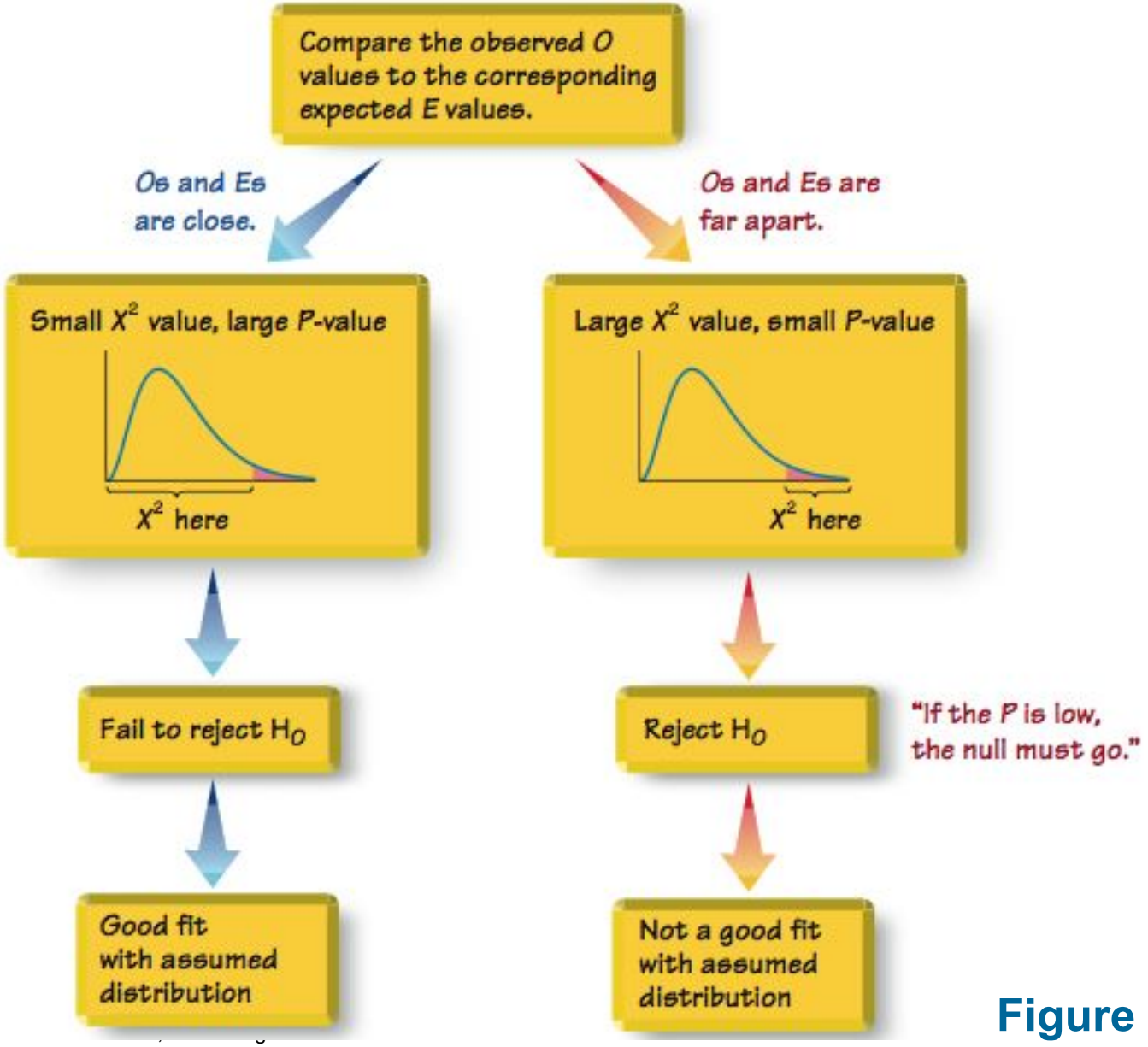


Figure 11-2

Example:

Data Set 1 in Appendix B includes weights from 40 randomly selected adult males and 40 randomly selected adult females. Those weights were obtained as part of the National Health Examination Survey. When obtaining weights of subjects, it is extremely important to actually weigh individuals instead of asking them to report their weights. By analyzing the last digits of weights, researchers can verify that weights were obtained through actual measurements instead of being reported.

Example:

When people report weights, they typically round to a whole number, so reported weights tend to have many last digits consisting of 0. In contrast, if people are actually weighed with a scale having precision to the nearest 0.1 pound, the weights tend to have last digits that are uniformly distributed, with 0, 1, 2, … , 9 all occurring with roughly the same frequencies. Table 11-2 shows the frequency distribution of the last digits from 80 weights listed in Data Set 1 in Appendix B.

Example:

(For example, the weight of 201.5 lb has a last digit of 5, and this is one of the data values included in Table 11-2.)
Test the claim that the sample is from a population of weights in which the last digits do not occur with the same frequency. Based on the results, what can we conclude about the procedure used to obtain the weights?

Example:

Table 11-2 Last Digits of Weights

Last Digit	Frequency
0	7
1	14
2	6
3	10
4	8
5	4
6	5
7	6
8	12
9	8

Example:

Requirements are satisfied: randomly selected subjects, frequency counts, expected frequency is 8 (> 5)

Step 1: at least one of the probabilities p_0, p_1, … p_9, is different from the others

Step 2: at least one of the probabilities are the same:

$$p_0 = p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9$$

Step 3: null hypothesis contains equality

H_0: $p_0 = p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9$
H_1: At least one probability is different

Example:

Step 4: no significance specified, use $\alpha = 0.05$

Step 5: testing whether a uniform distribution so use goodness-of-fit test: χ^2

Step 6: see the next slide for the computation of the χ^2 test statistic. The test statistic $\chi^2 = 11.250$, using $\alpha = 0.05$ and $k - 1 = 9$ degrees of freedom, the critical value is $\chi^2 = 16.919$

Example:

Table 11-3 Calculating the χ^2 Test Statistic for the Last Digits of Weights

Last Digit	Observed Frequency O	Expected Frequency E	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
0	7	8	−1	1	0.125
1	14	8	6	36	4.500
2	6	8	−2	4	0.500
3	10	8	2	4	0.500
4	8	8	0	0	0.000
5	4	8	−4	16	2.000
6	5	8	−3	9	1.125
7	6	8	−2	4	0.500
8	12	8	4	16	2.000
9	8	8	0	0	0.000
	80	80	$\chi^2 = \sum \frac{(O - E)^2}{E} = 11.250$		

Example:

Step 7: Because the test statistic does not fall in the critical region, there is not sufficient evidence to reject the null hypothesis.

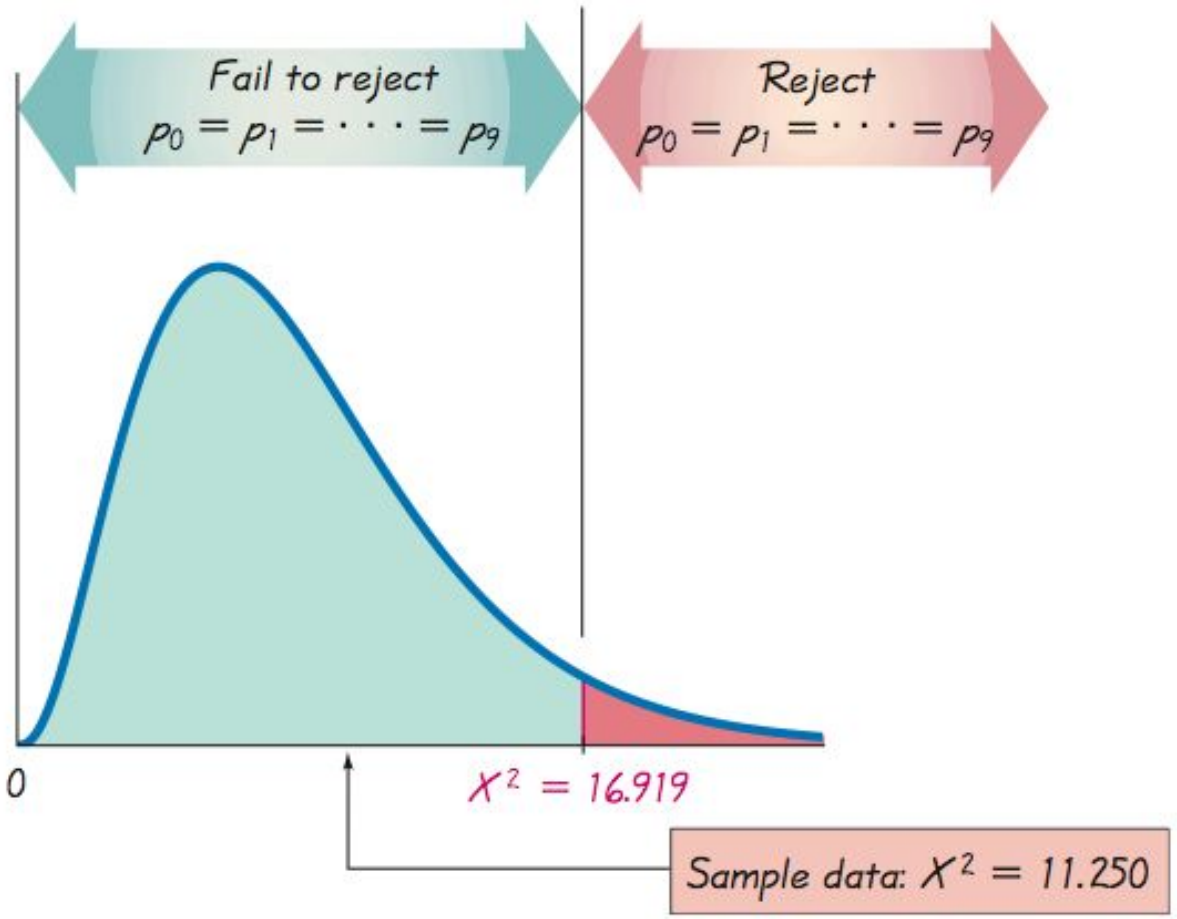


Figure 11-3 Test of $p_0 = p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9$

Example:

Step 8: There is not sufficient evidence to support the claim that the last digits do not occur with the same relative frequency.

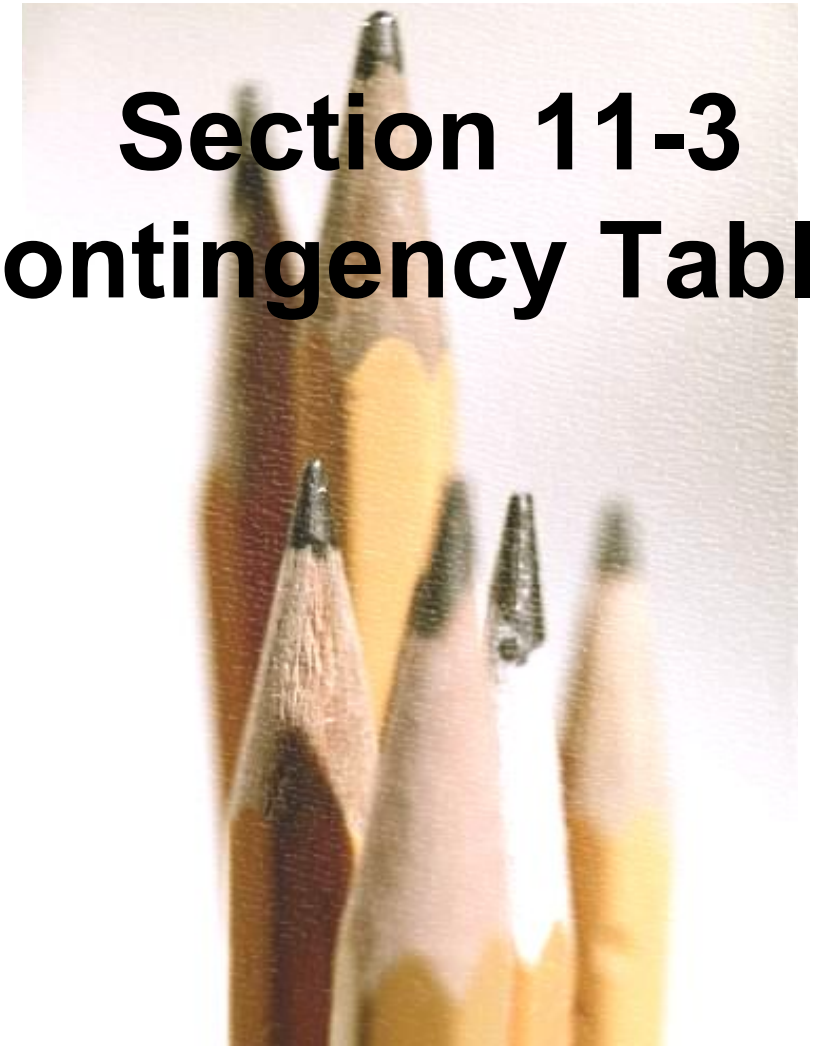
This goodness-of-fit test suggests that the last digits provide a reasonably good fit with the claimed distribution of equally likely frequencies. Instead of asking the subjects how much they weigh, it appears that their weights were actually measured as they should have been.

Recap

In this section we have discussed:

- **Goodness-of-Fit**
- **Equal Expected Frequencies**
- **Unequal Expected Frequencies**
- **Test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution.**

Section 11-3
Contingency Tables



Key Concept

In this section we consider *contingency tables* (or *two-way frequency tables*), which include frequency counts for categorical data arranged in a table with a least two rows and at least two columns.

We present a method for testing the claim that the row and column variables are independent of each other.

We will use the same method for a test of homogeneity, whereby we test the claim that different populations have the same proportion of some characteristics.

Part 1: Basic Concepts of Testing for Independence

Definition

A contingency table (or two-way frequency table) is a table in which frequencies correspond to two variables.

(One variable is used to categorize rows, and a second variable is used to categorize columns.)

Contingency tables have at least two rows and at least two columns.

Definition

Test of Independence

A test of independence tests the null hypothesis that in a contingency table, the row and column variables are independent.

Notation

O **represents the *observed frequency* in a cell of a contingency table.**

E **represents the *expected frequency* in a cell, found by assuming that the row and column variables are independent**

r **represents the number of rows in a contingency table (not including labels).**

c **represents the number of columns in a contingency table (not including labels).**

Requirements

1. **The sample data are randomly selected.**

2. **The sample data are represented as frequency counts in a two-way table.**

3. **For every cell in the contingency table, the expected frequency E is at least 5. (There is no requirement that every observed frequency must be at least 5. Also, there is no requirement that the population must have a normal distribution or any other specific distribution.)**

Null and Alternative Hypotheses

H_0: The row and column variables are *independent*.

H_1: The row and column variables are dependent.

Test of Independence

Test Statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where *O* is the observed frequency in a cell and *E* is the expected frequency found by evaluating

$$E = \frac{(\text{row total})\,(\text{column total})}{(\text{grand total})}$$

Test of Independence

Critical Values

1. **Found in Table A-4 using**

   degrees of freedom = $(r - 1)(c - 1)$

   r is the number of rows and c is the number of columns

2. **Tests of Independence are always *right-tailed*.**

Test of Independence

P-Values

***P*-values are typically provided by computer software, or a range of *P*-values can be found from Table A-4.**

Test of Independence

This procedure cannot be used to establish a direct cause-and-effect link between variables in question.

Dependence means only there is a relationship between the two variables.

Expected Frequency for Contingency Tables

$$E = \text{grand total} \cdot \frac{\text{row total}}{\text{grand total}} \cdot \frac{\text{column total}}{\text{grand total}}$$

$$E = n \cdot p$$

(probability of a cell)

$$E = \frac{(\text{row total})\,(\text{column total})}{(\text{grand total})}$$

Example:

Refer to Table 11-6 and find the expected frequency for the first cell, where the observed frequency is 88.

Table 11-6 Results from Experiment with Echinacea

	Treatment Group		
	Placebo	Echinacea: 20% extract	Echinacea: 60% extract
Infected	88	48	42
Not infected	15	4	10

The first cell lies in the first row (with a total frequency of 178) and the first column (with total frequency of 103). The "grand total" is the sum of all frequencies in the table, which is 207. The expected frequency of the first cell is

Example:

$$E = \frac{\text{(row total) (column total)}}{\text{(grand total)}} = \frac{(178)(103)}{207} = 88.570$$

We know that the first cell has an observed frequency of $O = 88$ and an expected frequency of $E = 88.570$. We can interpret the expected value by stating that if we assume that getting an infection is independent of the treatment, then we expect to find that 88.570 of the subjects would be given a placebo and would get an infection. There is a discrepancy between $O = 88$ and $E = 88.570$, and such discrepancies are key components of the test statistic.

Example:

Common colds are typically caused by a rhinovirus. In a test of the effectiveness of echinacea, some test subjects were treated with echinacea extracted with 20% ethanol, some were treated with echinacea extracted with 60% ethanol, and others were given a placebo. All of the test subjects were then exposed to rhinovirus. Results are summarized in Table 11-6 (next slide). Use a 0.05 significance level to test the claim that getting an infection (cold) is independent of the treatment group. What does the result indicated about the effectiveness of echinacea as a treatment for colds?

Example:

Table 11-6 Results from Experiment with Echinacea

	Treatment Group		
	Placebo	Echinacea: 20% extract	Echinacea: 60% extract
Infected	88	48	42
Not infected	15	4	10

Requirements are satisfied: randomly assigned to treatment groups, frequency counts, expected frequencies are all at least 5

H_0: Getting an infection is independent of the treatment

H_1: Getting an infection and the treatment are dependent

Example:

Significance level is $\alpha = 0.05$.

Contingency table: use χ^2 distribution

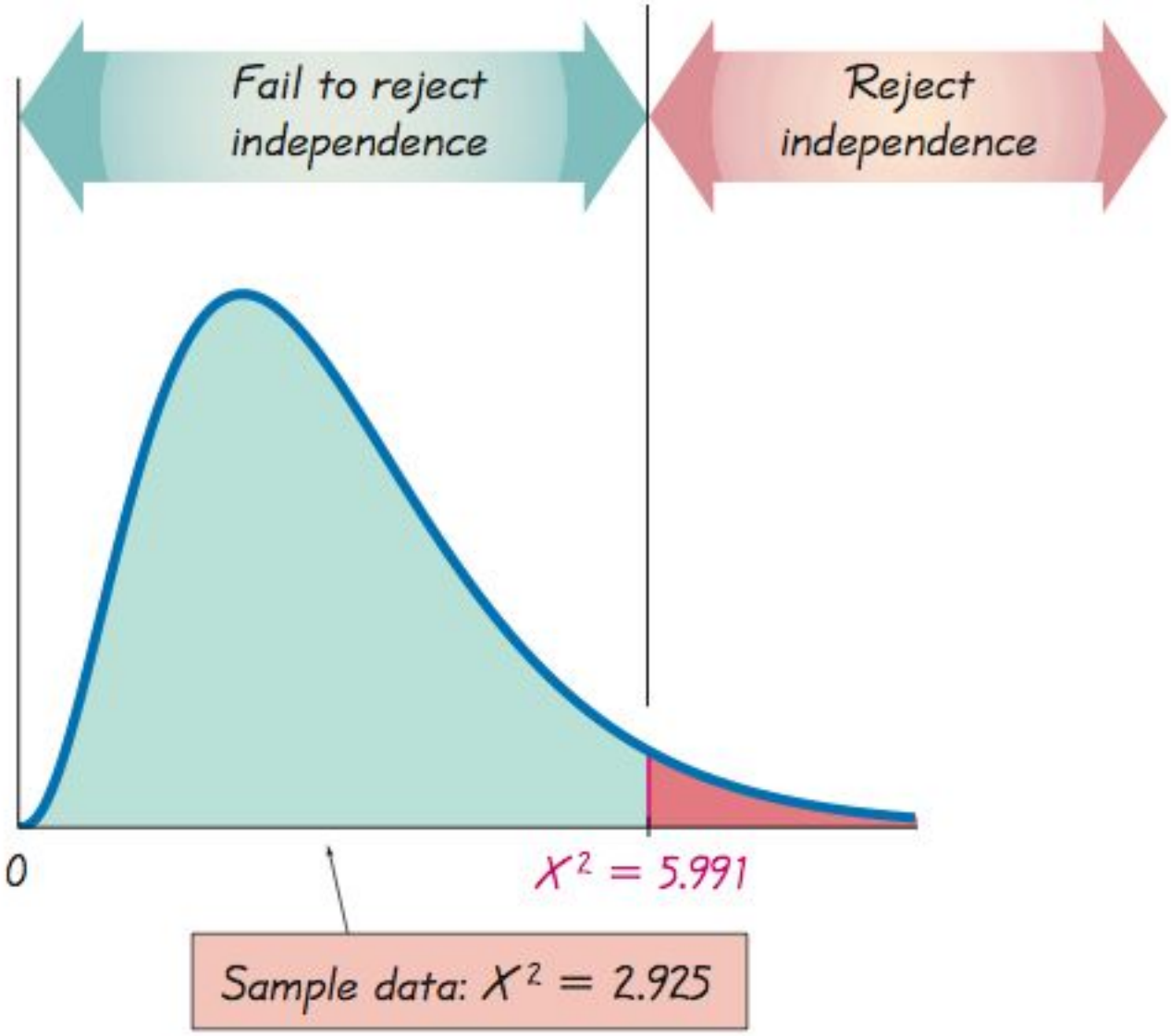
$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(88-88.570)^2}{88.570} + \ldots + \frac{(10-7.285)^2}{7.285}$$

$$= 2.925$$

The critical value of $\chi^2 = 5.991$ is found from Table A-4 with $\alpha = 0.05$ in the right tail and the number of degrees of freedom given by $(r-1)(c-1) = (2-1)(3-1) = 2$.

Example:



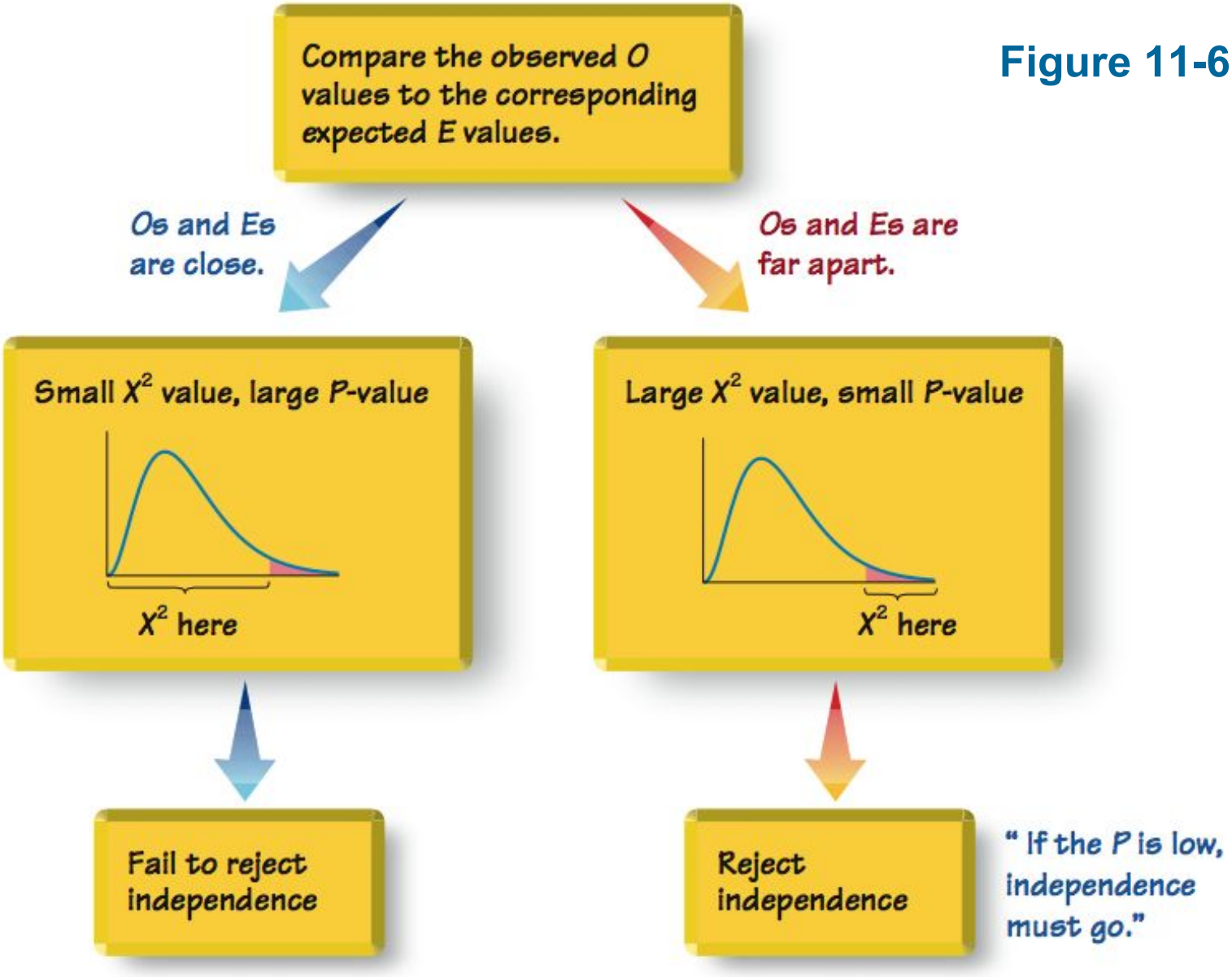
Example:

Because the test statistic does not fall within the critical region, we fail to reject the null hypothesis of independence between getting an infection and treatment.

It appears that getting an infection is independent of the treatment group. This suggests that echinacea is not an effective treatment for colds.

Relationships Among Key Components in Test of Independence

Figure 11-6



Part 2: Test of Homogeneity and the Fisher Exact Test

Definition

Test of Homogeneity

In a test of homogeneity, we test the claim that *different populations* have the same proportions of some characteristics.

How to Distinguish Between a Test of Homogeneity and a Test for Independence:

Were *predetermined* sample sizes used for different populations (test of homogeneity), or was one big sample drawn so both row and column totals were determined randomly (test of independence)?

Example:

Does a pollster's gender have an effect on poll responses by men? A U.S. News & World Report article about polls stated: "On sensitive issues, people tend to give 'acceptable' rather than honest responses; their answers may depend on the gender or race of the interviewer." To support that claim, data were provided for an Eagleton Institute poll in which surveyed men were asked if they agreed with this statement: "Abortion is a private matter that should be left to the woman to decide without government intervention."

Example:

We will analyze the effect of gender on male survey subjects only. Table 11-8 is based on the responses of surveyed men. Assume that the survey was designed so that male interviewers were instructed to obtain 800 responses from male subjects, and female interviewers were instructed to obtain 400 responses from male subjects. Using a 0.05 significance level, test the claim that the proportions of agree/disagree responses are the same for the subjects interviewed by men and the subjects interviewed by women.

Example:

Table 11-6 Gender and Survey Responses

	Gender of Interviewer	
	Man	Woman
Men who agree	560	308
Men who disagree	240	92

Example:

Requirements are satisfied: data are random, frequency counts in a two-way table, expected frequencies are all at least 5

Test for homogeneity.

H_0: The proportions of agree/disagree responses are the same for the subjects interviewed by men and the subjects interviewed by women.

H_1: The proportions are different.

Example:

Significance level is *a* = 0.05.
This time we'll use MINITAB.

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

	C1	C2	Total
1	**560**	**308**	868
	578.67	289.33	
	0.602	1.204	
2	**240**	**92**	332
	221.33	110.67	
	1.574	3.149	
Total	800	400	1200

Chi-Sq = 6.529, DF = 1, P-Value = 0.011

Example:

The Minitab display shows the expected frequencies of 578.67, 289.33, 221.33, and 110.67. It also includes the test statistic of $\chi^2 = 6.529$ and the *P*-value of 0.011.
Using the *P*-value approach to hypothesis testing, we reject the null hypothesis of equal (homogeneous) proportions (because the *P*-value of 0.011 is less than 0.05).
There is sufficient evidence to warrant rejection of the claim that the proportions are the same.

Example:

It appears that response and the gender of the interviewer are dependent. Although this statistical analysis cannot be used to justify any statement about causality, it does appear that men are influenced by the gender of the interviewer.

Fisher Exact Test

The procedures for testing hypotheses with contingency tables with two rows and two columns (2 × 2) have the requirement that every cell must have an expected frequency of at least 5. This requirement is necessary for the χ^2 distribution to be a suitable approximation to the exact distribution of the test statistic.

Fisher Exact Test

The Fisher exact test is often used for a 2 × 2 contingency table with one or more expected frequencies that are below 5. The Fisher exact test provides an *exact P*-value and does not require an approximation technique. Because the calculations are quite complex, it's a good idea to use computer software when using the Fisher exact test. STATDISK and Minitab both have the ability to perform the Fisher exact test.

Recap

In this section we have discussed:

- Contingency tables where categorical data is arranged in a table with a least two rows and at least two columns.
- Test of Independence tests the claim that the row and column variables are independent of each other.
- Test of Homogeneity tests the claim that different populations have the same proportion of some characteristics.
- Fisher Exact Test

Section 11-4
McNemar's Test for Matched Pairs

Key Concept

The Contingency table procedures in Section 11-3 are based on independent data. For 2 x 2 tables consisting of frequency counts that result from matched pairs, we do not have independence, and for such cases, we can use McNemar's test for matched pairs.

Key Concept

In this section we present the method of using McNemar's test for testing the null hypothesis that the frequencies from the discordant (different) categories occur in the same proportion.

Table 11-9 is a general table summarizing the frequency counts that result from matched pairs.

Table 11-9 2 × 2 Table with Frequency Counts from Matched Pairs

		Treatment X	
		Cured	Not Cured
Treatment Y	Cured	*a*	*b*
	Not cured	*c*	*d*

Definition

McNemar's Test uses frequency counts from matched pairs of nominal data from two categories to test the null hypothesis that for a 2 × 2 table such as Table 11-9, the frequencies *b* and *c* occur in the same proportion.

Notation

***a*, *b*, *c*, and *d* represent the frequency counts from a 2 × 2 table consisting of frequency counts from matched pairs.**

(The total number of subjects is *a* + *b* + *c* + *d*.)

Requirements

1. **The sample data have been randomly selected.**
2. **The sample data consist of matched pairs of frequency counts.**
3. **The data are at the nominal level of measurement, and each observation can be classified two ways: (1) According to the category distinguishing values with each matched pair, and (2) according to another category with two possible values.**
4. **For tables such as Table 11-9, the frequencies are such that $b + c \geq 10$.**

Null and Alternative Hypotheses

H_0: The proportions of the frequencies *b* and *c* (as in Table 11-9) are the same.

H_1: The proportions of the frequencies *b* and *c* (as in Table 11-9) are different.

Test Statistic

Test Statistic (for testing the null hypothesis that for tables such as Table 11-9, the frequencies *b* and *c* occur in the same proportion):

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c}$$

where the frequencies of *b* and *c* are obtained from the 2 × 2 table with a format similar to Table 11-9.

Critical Values

1. **The critical region is located in the right tail only.**

2. **The critical values are found in Table A-4 by using degrees of freedom = 1.**

P-Values

***P*-values are typically provided by computer software, or a range of *P*-values can be found from Table A-4.**

Example:

A randomized controlled trial was designed to test the effectiveness of hip protectors in preventing hip fractures in the elderly. Nursing home residents each wore protection on one hip, but not the other. Results are summarized in Table 11-10.
Using a 0.05 significance level, apply McNemar's test to test the null hypothesis that the following two proportions are the same:

Example:

- The proportion of subjects with no hip fracture on the protected hip and a hip fracture on the unprotected hip.

- The proportion of subjects with a hip fracture on the protected hip and no hip fracture on the unprotected hip.

Based on the results, do the hip protectors appear to be effective in preventing hip fractures?

Example:

Requirements are satisfied: randomly selected subject; matched pairs of frequency counts; nominal level of measurement, categorized according to two variables, one is "hip protector worn" or "not", the other is "hip fractured" or "not"; $b + c = 10 + 15 = 25$, which is at least 10

Data comes from matched pairs so use McNemar's test: $b = 10$ and $c = 15$

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} = \frac{(|10-15|-1)^2}{10+15} = 0.640$$

Example:

Table A-4 with a 0.05 significance level and degrees of freedom = 1, right-tailed test:

$$\chi^2 = 3.841$$

The test statistic of $\chi^2 = 0.640$ does not exceed the critical value of $\chi^2 = 3.841$, so fail to reject the null hypothesis. Note the *P*-value is 0.424, which is greater than 0.05, so reject the null.

It appears that the proportion of hip fractures with the protectors worn is not significantly different from the proportion of hip fractures with- out the protectors worn. The hip protectors do not appear to be effective in preventing hip fractures.

Definition

Discordant pairs of results come from matched pairs of results in which two categories are different (as in the frequencies *b* and *c* in Table 11-9).

Caution

When applying McNemar's test, be careful to use only the frequencies from the pairs of categories that are *different*. Do not blindly use the frequencies in the upper right and lower left corners, because they do not necessarily represent the discordant pairs. If Table 11-10 were reconfigured as shown on the next slide, it would be inconsistent in its format, but it would be technically correct in summarizing the same results as Table 11-10; however, blind use of the frequencies of 0 and 309 would result in the wrong test statistic.

Caution

		No Hip Protector Worn	
		No Hip Fracture	Hip Fracture
Hip Protector Worn	Hip Fracture	15	0
	No Hip Fracture	309	10

The discordant pairs of frequencies are:

Hip fracture / No hip fracture: 15

No hip fracture / Hip fracture: 10

Still use 15 and 10 (not 0 and 309).

Caution

In addition to comparing treatments given to matched pairs McNemar's test is often used to test a null hypothesis of no change in before/after types of experiments.

Recap

In this section we have discussed:

McNemar's test for matched pairs.

- **Data are place in a 2 x 2 table where each observation is classified in two ways.**
- **The test only compares categories that are different (discordant pairs).**