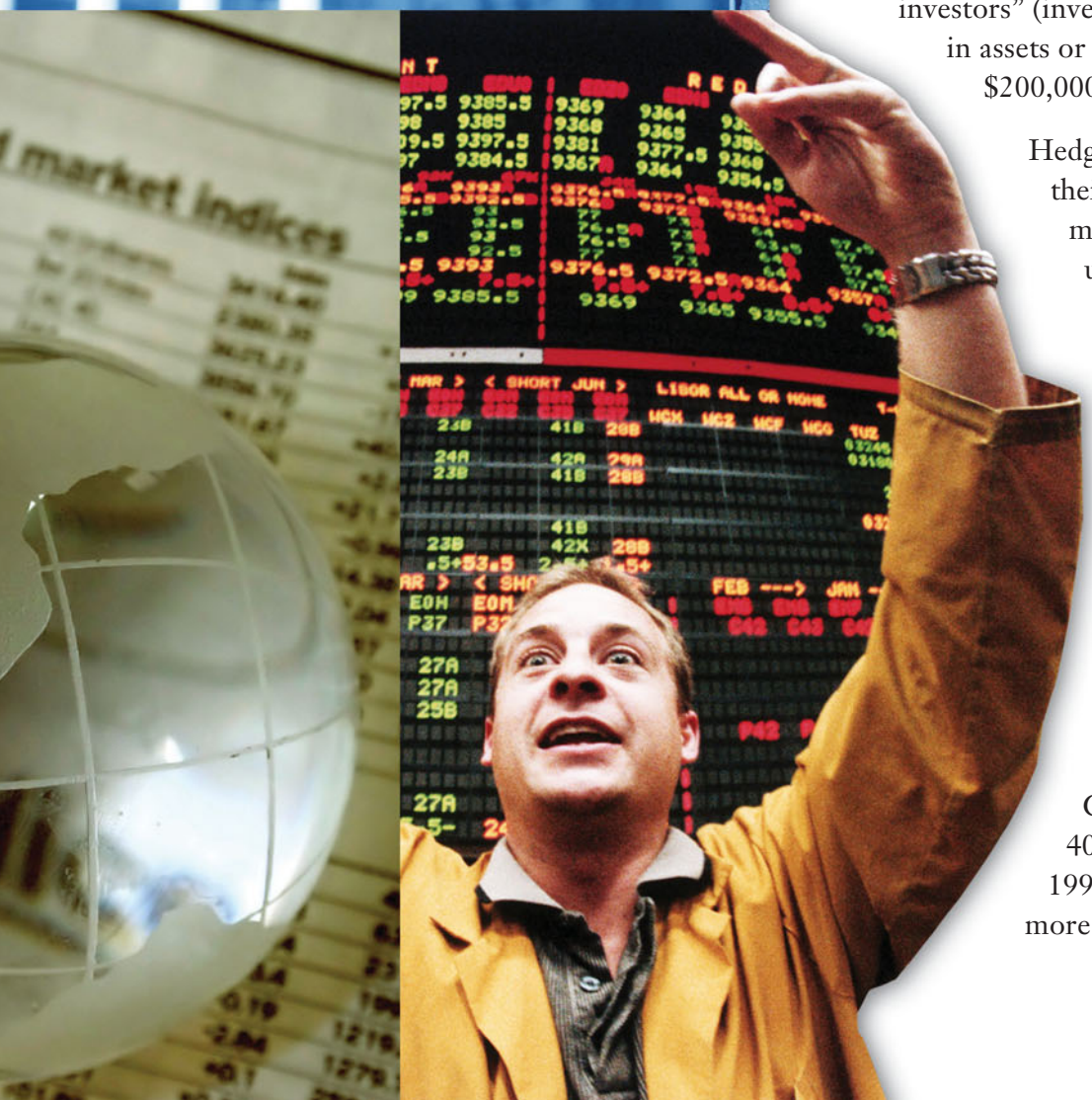
Inference for Counts: Chi-Square Tests

CHAPTER 15



SAC Capital

Hedge funds, like mutual funds and pension funds, pool investors' money in an attempt to make profits. Unlike these other funds, however, hedge funds are not required to register with the U.S. Securities and Exchange Commission (SEC) because they issue securities in "private offerings" only to "qualified investors" (investors with either $1 million in assets or annual income of at least $200,000).

Hedge funds don't necessarily "hedge" their investments against market moves. But typically these funds use multiple, often complex, strategies to exploit inefficiencies in the market. For these reasons, hedge fund managers have the reputation for being obsessive traders.

One of the most successful hedge funds is SAC Capital, which was founded by Steven (Stevie) A. Cohen in 1992 with nine employees and $25 million in assets under management (AUM). SAC Capital returned annual gains of 40% or more through much of the 1990s and is now reported to have more than 800 employees and nearly

$14 billion in assets under management. According to *Forbes*, Cohen's $6.4 billion fortune ranks him as the 36th wealthiest American.

Cohen, a legendary figure on Wall Street, is known for taking advantage of any information he can find and for turning that information into profit. SAC Capital is one of the most active trading organizations in the world. According to *Business Week* (7/21/2003), Cohen's firm "routinely accounts for as much as 3% of the NYSE's average daily trading, plus up to 1% of the NASDAQ's—a total of at least 20 million shares a day."

In a business as competitive as hedge fund management, information is gold. Being the first to have information and knowing how to act on it can mean the difference between success and failure. Hedge fund managers look for small advantages everywhere, hoping to exploit inefficiencies in the market and to turn those inefficiencies into profit.

Wall Street has plenty of "wisdom" about market patterns. For example, investors are advised to watch for "calendar effects," certain times of year or days of the week that are particularly good or bad: "As goes January, so goes the year" and "Sell in May and go away." Some analysts claim that the "bad period" for holding stocks is from the sixth trading day of June to the fifth-to-last trading day of October. Of course, there is also Mark Twain's advice:

> October. This is one of the peculiarly dangerous months to speculate in stocks. The others are July, January, September, April, November, May, March, June, December, August, and February.
>
> —Pudd'nhead Wilson's Calendar

One common claim is that stocks show a weekly pattern. For example, some argue that there is a *weekend effect* in which stock returns on Mondays are often lower than those of the immediately preceding Friday. Are patterns such as this real? We have the data, so we can check. Between October 1, 1928 and June 6, 2007, there were 19,755 trading sessions. Let's first see how many trading days fell on each day of the week. It's not exactly 20% for each day because of holidays. The distribution of days is shown in Table 15.1.

Day of Week	Count	% of days
Monday	3820	19.3369%
Tuesday	4002	20.2582
Wednesday	4024	20.3695
Thursday	3963	20.0607
Friday	3946	19.9747

Table 15.1 The distribution of days of the week among the 19,755 trading days from October 1, 1928 to June 6, 2007. We expect about 20% to fall in each day, with minor variations due to holidays and other events.

Of these 19,755 trading sessions, 10,272, or about 52% of the days, saw a gain in the Dow Jones Industrial Average (DJIA). To test for a pattern, we need a model. The model comes from the supposition that any day is as likely to show a gain as any other. In any sample of positive or "up" days, we should expect to see the same distribution of days as in Table 15.1—in other words, about 19.34% of "up" days would be Mondays, 20.26% would be Tuesdays, and so on. Here is the distribution of days in one such random sample of 1000 "up" days.

Day of Week	Count	% of days in the sample of "up" days
Monday	192	19.2%
Tuesday	189	18.9
Wednesday	202	20.2
Thursday	199	19.9
Friday	218	21.8

Table 15.2 The distribution of days of the week for a sample of 1000 "up" trading days selected at random from October 1, 1928 to June 6, 2007. If there is no pattern, we would expect the proportions here to match fairly closely the proportions observed among all trading days in Table 15.1.

Of course, we expect some variation. We wouldn't expect the proportions of days in the two tables to match exactly. In our sample, the percentage of Mondays in Table 15.2 is slightly lower than in Table 15.1, and the proportion of Fridays is a little higher. Are these deviations enough for us to declare that there is a recognizable pattern?

15.1 Goodness-of-Fit Tests

To address this question, we test the table's **goodness-of-fit**, where *fit* refers to the null model proposed. Here, the null model is that there is no pattern, that the distribution of *up* days should be the same as the distribution of trading days overall. (If there were no holidays or other closings, that would just be 20% for each day of the week.)

Assumptions and Conditions

Data for a goodness-of-fit test are organized in tables, and the assumptions and conditions reflect that. Rather than having an observation for each individual, we typically work with summary counts in categories. Here, the individuals are trading days, but rather than list all 1000 trading days in the sample, we have totals for each weekday.

Counted Data Condition. The data must be counts for the categories of a categorical variable. This might seem a silly condition to check. But many kinds of values can be assigned to categories, and it is unfortunately common to find the methods of this chapter applied incorrectly (even by business professionals) to proportions or quantities just because they happen to be organized in a two-way table. So check to be sure that you really have counts.

Independence Assumption

Independence Assumption. The counts in the cells should be independent of each other. You should think about whether that's reasonable. If the data are a random sample you can simply check the randomization condition.

Randomization Condition. The individuals counted in the table should be a random sample from some population. We need this condition if we want to generalize our conclusions to that population. We took a random sample of 1000 trading days on which the DJIA rose. That lets us assume that the market's performance on any one day is independent of performance on another. If we had selected 1000 consecutive trading days, there would be a risk that market performance on one day could affect performance on the next, or that an external event could affect performance for several consecutive days.

Expected Cell Frequencies

Companies often want to assess the relative successes of their products in different regions. However, a company whose sales regions had 100, 200, 300, and 400 representatives might not expect equal sales in all regions. They might expect observed sales to be proportional to the size of the sales force. The null hypothesis in that case would be that the proportions of sales were 1/10, 2/10, 3/10, and 4/10, respectively. With 500 total sales, their expected counts would be 50, 100, 150, and 200.

Sample Size Assumption

Sample Size Assumption. We must have enough data for the methods to work. We usually just check the following condition:

Expected Cell Frequency Condition. We should expect to see at least 5 individuals in each cell. The expected cell frequency condition should remind you of—and is, in fact, quite similar to—the condition that np and nq be at least 10 when we test proportions.

Chi-Square Model

We have observed a count in each category (weekday). We can compute the number of up days we'd *expect* to see for each weekday if the null model were true. For the trading days example, the expected counts come from the null hypothesis that the up days are distributed among weekdays just as trading days are. Of course, we could imagine almost any kind of model and base a null hypothesis on that model.

To decide whether the null model is plausible, we look at the differences between the expected values from the model and the counts we observe. We wonder: Are these differences so large that they call the model into question, or could they have arisen from natural sampling variability? We denote the *differences* between these observed and expected counts, $(Obs - Exp)$. As we did with variance, we square them. That gives us positive values and focuses attention on any cells with large differences. Because the differences between observed and expected counts generally get larger the more data we have, we also need to get an idea of the *relative* sizes of the differences. To do that, we divide each squared difference by the expected count for that cell.

The test statistic, called the **chi-square (or chi-squared) statistic**, is found by adding up the sum of the squares of the deviations between the observed and expected counts divided by the expected counts:

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp}.$$

The chi-square statistic is denoted χ^2, where χ is the Greek letter chi (pronounced kī). The resulting family of sampling distribution models is called the **chi-square models**.

Notation Alert!

We compare the counts *observed* in each cell with the counts we *expect* to find. The usual notation uses *Obs* and *Exp* as we've used here. The expected counts are found from the null model.

Notation Alert!

The only use of the Greek letter χ in Statistics is to represent the chi-square statistic and the associated sampling distribution. This violates the general rule that Greek letters represent population parameters. Here we are using a Greek letter simply to name a family of distribution models and a statistic.

The members of this family of models differ in the number of degrees of freedom. The number of degrees of freedom for a goodness-of-fit test is $k - 1$, where k is the number of cells—in this example, 5 weekdays.

We will use the chi-square statistic only for testing hypotheses, not for constructing confidence intervals. A small chi-square statistic means that our model fits the data well, so a small value gives us no reason to doubt the null hypothesis. If the observed counts don't match the expected counts, the statistic will be large. If the calculated statistic value is large enough, we'll reject the null hypothesis. So the chi-square test is always one-sided. What could be simpler? Let's see how it works.

For Example — Goodness of fit test

Atara manages 8 call center operators at a telecommunications company. To develop new business, she gives each operator a list of randomly selected phone numbers of rival phone company customers. She also provides the operators with a script that tries to convince the customers to switch providers. Atara notices that some operators have found more than twice as many new customers as others, so she suspects that some of the operators are performing better than others.

The 120 new customer acquisitions are distributed as follows:

Operator	1	2	3	4	5	6	7	8
New customers	11	17	9	12	19	18	13	21

Question: Is there evidence to suggest that some of the operators are more successful than others?

Answer: Atara has randomized the potential new customers to the operators so the Randomization Condition is satisfied. The data are counts and there are at least 5 in each cell, so we can apply a chi-square goodness-of-fit test to the null hypothesis that the operator performance is uniform and that each of the operators will convince the same number of customers. Specifically we expect each operator to have converted 1/8 of the 120 customers that switched providers.

Operator	1	2	3	4	5	6	7	8
Observed	11	17	9	12	19	18	13	21
Expected	15	15	15	15	15	15	15	15
Observed-Expected	−4	2	−6	−3	4	3	−2	6
(Obs-Exp)^2	16	4	36	9	16	9	4	36
(Obs-Exp)^2/Exp	16/15 = 1.07	4/15 = 0.27	36/15 = 2.40	9/15 = 0.60	16/15 = 1.07	9/15 = 0.60	4/15 = 0.27	36/15 = 2.40

$$\sum \frac{(Obs - Exp)^2}{Exp} = 1.07 + 0.27 + 2.40 + \cdots + 2.40 = 8.67$$

The number of degrees of freedom is $k - 1 = 7$.

$$P(\chi^2_7 > 8.67) = 0.2772.$$

8.67 is not a surprising value for a Chi-square statistic with 7 degrees of freedom. So, we fail to reject the null hypothesis that the operators actually find new customers at different rates.

By Hand

The chi-square calculation

Here are the steps to calculate the chi-square statistic:

1. **Find the expected values.** These come from the null hypothesis model. Every null model gives a hypothesized proportion for each cell. The expected value is the product of the total number of observations times this proportion. (The result need not be an integer.)
2. **Compute the residuals.** Once you have expected values for each cell, find the residuals, $Obs - Exp$.
3. **Square the residuals.** $(Obs - Exp)^2$
4. **Compute the components.** Find $\frac{(Obs - Exp)^2}{Exp}$ for each cell.

(continued)

5. **Find the sum of the components.** That's the chi-square statistic,
$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp}.$$
6. **Find the degrees of freedom.** It's equal to the number of cells minus one.
7. **Test the hypothesis.** Large chi-square values mean lots of deviation from the hypothesized model, so they give small P-values. Look up the critical value from a table of chi-square values such as Table X in Appendix D, or use technology to find the P-value directly.

The steps of the chi-square calculations are often laid out in tables. Use one row for each category, and columns for observed counts, expected counts, residuals, squared residuals, and the contributions to the chi-square total:

	A	B	C	D	E	F
1		Observed	Expected	Residual = (Obs - Exp)	(Obs-Exp)^2	Component = (Obs - Exp)^2 / Exp
2	Monday	192	193.369	-1.369	1.879	0.0097
3	Tuesday	189	202.582	-13.582	184.461	0.9105
4	Wednesday	202	203.695	-1.695	2.874	0.0141
5	Thursday	199	200.607	-1.607	2.584	0.0129
6	Friday	218	199.747	18.253	333.176	1.668

Table 15.3 Calculations for the chi-square statistic in the trading days example, can be performed conveniently in **Excel**. Set up the calculation in the first row and Fill Down, then find the sum of the rightmost column. The CHIDIST function looks up the chi square total to find the P-value.

Guided Example Stock Market Patterns

We have counts of the "up" days for each day of the week. The economic theory we want to investigate is whether there is a pattern in "up" days. So, our null hypothesis is that across all days in which the DJIA rose, the days of the week are distributed as they are across all trading days. (As we saw, the trading days are not quite *evenly* distributed because of holidays, so we use the *trading days* percentages as the null model.) We refer to this as *uniform*, accounting for holidays. The alternative hypothesis is that the observed percentages are *not* uniform. The test statistic looks at how closely the observed data match this idealized situation.

PLAN

Setup State what you want to know.

Identify the variables and context.

We want to know whether the distribution for "up" days differs from the null model (the trading days distribution). We have the number of times each weekday appeared among a random sample of 1000 "up" days.

Hypotheses State the null and alternative hypotheses. For χ^2 tests, it's usually easier to state the hypotheses in words than in symbols.

H_O: The days of the work week are distributed among the up days as they are among all trading days.

H_A: The trading days model does not fit the up days distribution.

Model Think about the assumptions and check the conditions.

- ✓ **Counted Data Condition** We have counts of the days of the week for all trading days and for the "up" days.
- ✓ **Independence Assumption** We have no reason to expect that one day's performance will affect another's, but to be safe we've taken a random sample of days. The randomization should make them far enough apart to alleviate any concerns about dependence.
- ✓ **Randomization Condition** We have a random sample of 1000 days from the time period.
- ✓ **Expected Cell Frequency Condition** All the expected cell frequencies are much larger than 5.

Specify the sampling distribution model.

Name the test you will use.

The conditions are satisfied, so we'll use a χ^2 model with $5 - 1 = 4$ degrees of freedom and do a **chi-square goodness-of-fit test.**

DO

Mechanics To find the expected number of days, we take the fraction of each weekday from all days and multiply by the number of "up" days.

For example, there were 3820 Mondays out of 19,755 trading days.

So, we'd expect there would be $1000 \times 3820/19{,}755$ or 193.369 Mondays among the 1000 "up" days.

Each cell contributes a value equal to $\frac{(Obs - Exp)^2}{Exp}$ to the chi-square sum.

Add up these components. If you do it by hand, it can be helpful to arrange the calculation in a table or spreadsheet.

The P-value is the probability in the upper tail of the χ^2 model. It can be found using software or a table (see Table X in Appendix D).

Large χ^2 statistic values correspond to small P-values, which would lead us to reject the null hypothesis, but the value here is not particularly large.

The expected values are:

Monday: 193.369
Tuesday: 202.582
Wednesday: 203.695
Thursday: 200.607
Friday: 199.747

And we observe:

Monday: 192
Tuesday: 189
Wednesday: 202
Thursday: 199
Friday: 218

$$\chi^2 = \frac{(192 - 193.369)^2}{193.369} + \cdots + \frac{(218 - 199.747)^2}{199.747} = 2.615$$

Using Table X in Appendix D, we find that for a significance level of 5% and 4 degrees of freedom, we'd need a value of 9.488 or more to have a P-value less than .05. Our value of 2.615 is less than that.

Using a computer to generate the P-value, we find:

$$\text{P-value} = P(\chi^2_4 > 2.615) = 0.624$$

(continued)

REPORT

Conclusion Link the P-value to your decision. Be sure to say more than a fact about the distribution of counts. State your conclusion in terms of what the data mean.

MEMO

Re: Stock Market Patterns

Our investigation of whether there are day-of-the-week patterns in the behavior of the DJIA in which one day or another is more likely to be an "up" day found no evidence of such a pattern. Our statistical test indicated that a pattern such as the one found in our sample of trading days would happen by chance about 62% of the time.

We conclude that there is, unfortunately, no evidence of a pattern that could be used to guide investment in the market. We were unable to detect a "weekend" or other day-of-the-week effect in the market.

15.2 Interpreting Chi-Square Values

When we calculated χ^2 for the trading days example, we got 2.615. That value was not large for 4 degrees of freedom, so we were unable to reject the null hypothesis. In general, what *is* big for a χ^2 statistic?

Think about how χ^2 is calculated. In every cell any deviation from the expected count contributes to the sum. Large deviations generally contribute more, but if there are a lot of cells, even small deviations can add up, making the χ^2 value larger. So the more cells there are, the higher the value of χ^2 has to be before it becomes significant. For χ^2, the decision about how big is big depends on the number of degrees of freedom.

Unlike the Normal and t families, χ^2 models are skewed. Curves in the χ^2 family change both shape and center as the number of degrees of freedom grows. For example, Figure 15.1 shows the χ^2 curves for 5 and for 9 degrees of freedom.

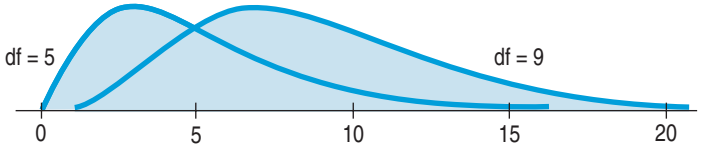


Figure 15.1 The χ^2 curves for 5 and 9 degrees of freedom.

Notice that the value $\chi^2 = 10$ might seem somewhat extreme when there are 5 degrees of freedom, but appears to be rather ordinary for 9 degrees of freedom. Here are two simple facts to help you think about χ^2 models:

- The mode is at $\chi^2 = \text{df} - 2$. (Look at the curves; their peaks are at 3 and 7.)
- The expected value (mean) of a χ^2 model is its number of degrees of freedom. That's a bit to the right of the mode—as we would expect for a skewed distribution.

Goodness-of-fit tests are often performed by people who have a theory of what the proportions *should* be in each category and who believe their theory to be true. In some cases, unlike our market example, there isn't an obvious null hypothesis against which to test the proposed model. So, unfortunately, in those cases, the only null hypothesis available is that the proposed theory is true. And as we know, the hypothesis testing procedure allows us only to reject the null or fail to reject it. We can never confirm that a theory is in fact true; we can never confirm the null hypothesis.

At best, we can point out that the data are consistent with the proposed theory. But this doesn't prove the theory. The data *could* be consistent with the model even if the theory were wrong. In that case, we fail to reject the null hypothesis but can't conclude anything for sure about whether the theory is true.

Why Can't We Prove the Null?

A student claims that it really makes no difference to your starting salary how well you do in your Statistics class. He surveys recent graduates, categorizes them according to whether they earned an A, B, or C in Statistics, and according to whether their starting salary is above or below the median for their class. He calculates the proportion above the median salary for each grade. His null model is that in each grade category, 50% of students are above the median. With 40 respondents, he gets a P-value of .07 and declares that Statistics grades don't matter. But then more questionnaires are returned, and he finds that with a sample size of 70, his P-value is .04. Can he ignore the second batch of data? Of course not. If he could do that, he could claim almost any null model was true just by having too little data to refute it.

15.3 Examining the Residuals

Chi-square tests are always one-sided. The chi-square statistic is always positive, and a large value provides evidence against the null hypothesis (because it shows that the fit to the model is *not* good), while small values provide little evidence that the model doesn't fit. In another sense, however, chi-square tests are really many-sided; a large statistic doesn't tell us *how* the null model doesn't fit. In our market theory example, if we had rejected the uniform model, we wouldn't have known *how* it failed. Was it because there were not enough Mondays represented, or was it that all five days showed some deviation from the uniform?

When we reject a null hypothesis in a goodness-of-fit test, we can examine the residuals in each cell to learn more. In fact, whenever we reject a null hypothesis, it's a good idea to examine the residuals. (We don't need to do that when we fail to reject because when the χ^2 value is small, all of its components must have been small.) Because we want to compare residuals for cells that may have very different counts, we standardize the residuals. We know the mean residual is zero,[1] but we need to know each residual's standard deviation. When we tested proportions, we saw a link between the expected proportion and its standard deviation. For counts, there's a similar link. To standardize a cell's residual, we divide by the square root of its expected value[2]:

$$\frac{(Obs - Exp)}{\sqrt{Exp}}.$$

Notice that these **standardized residuals** are the square roots of the components we calculated for each cell, with the plus (+) or the minus (−) sign indicating whether we observed more or fewer cases than we expected.

The standardized residuals give us a chance to think about the underlying patterns and to consider how the distribution differs from the model. Now that we've divided each residual by its standard deviation, they are z-scores. If the null hypothesis was true, we could even use the 68–95–99.7 Rule to judge how extraordinary the large ones are.

---

[1]Residual = observed − expected. Because the total of the expected values is the same as the observed total, the residuals must sum to zero.

[2]It can be shown mathematically that the square root of the expected value estimates the appropriate standard deviation.

Here are the standardized residuals for the trading days data:

	Standardized Residual $= \frac{(Obs - Exp)}{\sqrt{Exp}}$
Monday	−0.0984
Tuesday	−0.9542
Wednesday	−0.1188
Thursday	−0.1135
Friday	1.292

Table 15.4 Standardized residuals.

None of these values is remarkable. The largest, Friday, at 1.292, is not impressive when viewed as a z-score. The deviations are in the direction suggested by the "weekend effect," but they aren't quite large enough for us to conclude that they are real.

For Example — Examining residuals from a chi-square test

Question: In the call center example (see page 453), examine the residuals to see if any operators stand out as having especially strong or weak performance.

Answer: Because we failed to reject the null hypothesis, we don't expect any of the standardized residuals to be large, but we will examine them nonetheless.

The standardized residuals are the square roots of the components (from the bottom row of the table in the Example on page 453).

Standardized Residuals	−1.03	0.52	−1.55	−0.77	1.03	0.77	−0.52	1.55

As we expected, none of the residuals are large. Even though Atara notices that some of the operators enrolled more than twice the number of new customers as others, the variation is typical (within two standard deviations) of what we would expect if all their performances were, in fact, equal.

15.4 The Chi-Square Test of Homogeneity

Skin care products are big business. According to the American Academy of Dermatology, "the average adult uses at least seven different products each day," including moisturizers, skin cleansers, and hair cosmetics.[3] Growth in the skin care market in China during 2006 was 15%, fueled, in part, by massive economic growth. But not all cultures and markets are the same. Global companies must understand cultural differences in the importance of various skin care products in order to compete effectively.

The GfK Roper Reports® Worldwide Survey, which we first saw in Chapter 3, asked 30,000 consumers in 23 countries about their attitudes on health, beauty, and other personal values. One question participants were asked was how important is "Seeking the utmost attractive appearance" to you? Responses were a scale with 1 = Not at all important and 7 = Extremely important. Is agreement with this

[3] www.aad.org/public/Publications/pamphlets/Cosmetics.htm.

WHO	Respondents in the GfK Roper Reports Worldwide Survey
WHAT	Responses to questions relating to perceptions of food and health
WHEN	Fall 2005; published in 2006
WHERE	Worldwide
HOW	Data collected by GfK Roper Consulting using a multistage design
WHY	To understand cultural differences in the perception of the food and beauty products we buy and how they affect our health

question the same across the five countries for which we have data (China, France, India, U.K., and U.S.)? Here is a contingency table with the counts.

		Country					
		China	France	India	U.K.	U.S.	Total
Appearance	7—Extremely important	197	274	642	210	197	**1520**
	6	257	405	304	252	203	**1421**
	5	315	364	196	348	250	**1473**
	4—Average importance	480	326	263	486	478	**2033**
	3	98	82	41	125	100	**446**
	2	63	46	36	70	58	**273**
	1—Not at all important	92	38	53	62	29	**274**
	Total	**1502**	**1535**	**1535**	**1553**	**1315**	**7440**

Table 15.5 Responses to how important is "Seeking the utmost attractive appearance."

We can compare the countries more easily by examining the column percentages.

		Country					
		China	France	India	U.K.	U.S.	Row %
Appearance	7—Extremely important	13.12%	17.85	41.82	13.52	14.98	**20.43%**
	6	17.11	26.38	19.80	16.23	15.44	**19.10**
	5	20.97	23.71	12.77	22.41	19.01	**19.80**
	4—Average importance	31.96	21.24	17.13	31.29	36.35	**27.33**
	3	6.52	5.34	2.67	8.05	7.60	**5.99**
	2	4.19	3.00	2.35	4.51	4.41	**3.67**
	1—Not at all important	6.13	2.48	3.45	3.99	2.21	**3.68**
	Total	**100%**	**100**	**100**	**100**	**100**	**100%**

Table 15.6 Responses as a percentage of respondents by country.

The stacked bar chart of the responses by country shows the patterns more vividly:

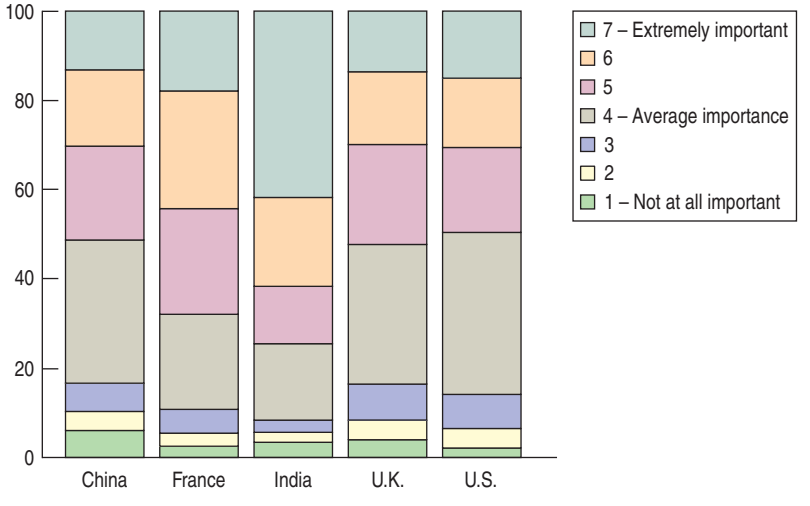


Figure 15.2 Responses to the question how important is "Seeking the utmost attractive appearance" by country. India stands out for the proportion of respondents who said Important or Extremely important.

It seems that India stands out from the other countries. There is a much larger proportion of respondents from India who responded *Extremely Important*. But are the observed differences in the percentages real or just natural sampling variation? Our null hypothesis is that the proportions choosing each alternative are the same for each country. To test that hypothesis, we use a **chi-square test of homogeneity**. This is just another chi-square test. It turns out that the mechanics of the test of this hypothesis are nearly identical to the chi-square goodness-of-fit test we just saw in Section 15.1. The difference is that the goodness-of-fit test compared our observed counts to the expected counts from a *given* model. The test of homogeneity, by contrast, has a null hypothesis that the distributions are the same for all the groups. The test examines the differences between the observed counts and what we'd expect under that assumption of homogeneity.

For example, 20.43% (the row %) of *all* 7440 respondents said that looking good was extremely important to them. If the distributions were homogeneous across the five countries (as the null hypothesis asserts), then that proportion should be the same for all five countries. So 20.43% of the 1315 U.S. respondents, or 268.66, would have said that looking good was extremely important. That's the number we'd *expect* under the null hypothesis.

Working in this way, we (or, more likely, the computer) can fill in expected values for each cell. The following table shows these expected values for each response and each country.

		Country					
		China	France	India	U.K.	U.S.	Total
Appearance	7—Extremely important	306.86	313.60	313.60	317.28	268.66	**1520**
	6	286.87	293.18	293.18	296.61	251.16	**1421**
	5	297.37	303.91	303.91	307.47	260.35	**1473**
	4—Average importance	410.43	419.44	419.44	424.36	359.33	**2033**
	3	90.04	92.02	92.02	93.10	78.83	**446**
	2	55.11	56.32	56.32	56.99	48.25	**273**
	1—Not at all important	55.32	56.53	56.53	57.19	48.43	**274**
	Total	**1502**	**1535**	**1535**	**1553**	**1315**	**7440**

Table 15.7 Expected values for the responses. Because these are theoretical values, they don't have to be integers.

The term *homogeneity* refers to the hypothesis that things are the same. Here, we ask whether the distribution of responses about the importance of looking good is the same across the five countries. The chi-square test looks for differences large enough to step beyond what we might expect from random sample-to-sample variation. It can reveal a large deviation in a single category or small but persistent differences over all the categories—or anything in between.

Assumptions and Conditions

The assumptions and conditions are the same as for the chi-square test for goodness-of-fit. The **Counted Data Condition** says that these data must be counts. You can never perform a chi-square test on a quantitative variable. For example, if Roper had recorded how much respondents spent on skin care products, you wouldn't be able to use a chi-square test to determine whether the mean expenditures in the five countries were the same.[4]

[4]To do that, you'd use a method called Analysis of Variance (see Chapter 21).

Large Samples and Chi-Square Tests

Whenever we test any hypothesis, a very large sample size means that small effects have a greater chance of being statistically significant. This is especially true for chi-square tests. So it's important to look at the effect sizes when the null hypothesis is rejected to see if the differences are practically significant. Don't rely only on the P-value when making a business decision. This applies to many of the examples in this chapter which have large sample sizes typical of those seen in today's business environment.

Independence Assumption. So that we can generalize, we need the counts to be independent of each other. We can check the **Randomization Condition**. Here, we have random samples, so we *can* assume that the observations are independent and draw a conclusion comparing the populations from which the samples were taken.

We must be sure we have enough data for this method to work. The **Sample Size Assumption** can be checked with the **Expected Cell Frequency Condition**, which says that the expected count in each cell must be at least 5. Here, our samples are certainly large enough.

Following the pattern of the goodness-of-fit test, we compute the component for each cell of the table:

$$\textbf{Component} = \frac{(\textit{Obs} - \textit{Exp})^2}{\textit{Exp}}.$$

Summing these components across all cells gives the chi-square value:

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp}.$$

The degrees of freedom are different than they were for the goodness-of-fit test. For a test of homogeneity, there are $(R - 1) \times (C - 1)$ degrees of freedom, where R is the number of rows and C is the number of columns.

In our example, we have $6 \times 4 = 24$ degrees of freedom. We'll need the degrees of freedom to find a P-value for the chi-square statistic.

By Hand

How to find expected values

In a contingency table, to test for homogeneity, we need to find the expected values when the null hypothesis is true. To find the expected value for row i and column j, we take:

$$Exp_{ij} = \frac{Total_{Row\ i} \times Total_{Col\ j}}{Table\ Total}$$

Here's an example:

Suppose we ask 100 people, 40 men and 60 women, to name their magazine preference: *Sports Illustrated*, *Cosmopolitan*, or *The Economist*, with the following result, shown in **Excel**:

	A	B	C	D	E
1	Actual	SI	Cosmo	Economist	Total
2	Men	25	5	10	40
3	Women	10	45	5	60
4	Total	35	50	15	100

Then, for example, the expected value under homogeneity for *Men* who prefer *The Economist* would be:

$$Exp_{13} = \frac{40 \times 15}{100} = 6$$

Performing similar calculations for all cells gives the expected values:

6	Expected	SI	Cosmo	Economist	Total
7	Men	14	20	6	40
8	Women	21	30	9	60
9	Total	35	50	15	100

Guided Example Attitudes on Appearance

How we think about our appearance, in part, depends on our culture. To help providers of beauty products with global markets, we want to examine whether the responses to the question "How important is seeking the utmost attractive appearance to you?" varied in the five markets of China, France, India, the U.K., and the U.S. We will use the data from the GfK Roper Reports Worldwide Survey.

PLAN

Setup State what you want to know.

Identify the variables and context.

We want to know whether the distribution of responses to how important is "Seeking the utmost attractive appearance" is the same for the five countries for which we have data: China, France, India, U.K., and U.S.

Hypotheses State the null and alternative hypotheses.

H_O: The responses are homogeneous (have the same distribution for all five countries).

H_A: The responses are not homogeneous.

Model Think about the assumptions and check the conditions.

We have counts of the number of respondents in each country who choose each response.

- ✓ **Counted Data Condition** *The data are counts of the number of people choosing each possible response.*
- ✓ **Randomization Condition** *The data were obtained from a random sample by a professional global marketing company.*
- ✓ **Expected Cell Frequency Condition** *The expected values in each cell are all at least 5.*

State the sampling distribution model.

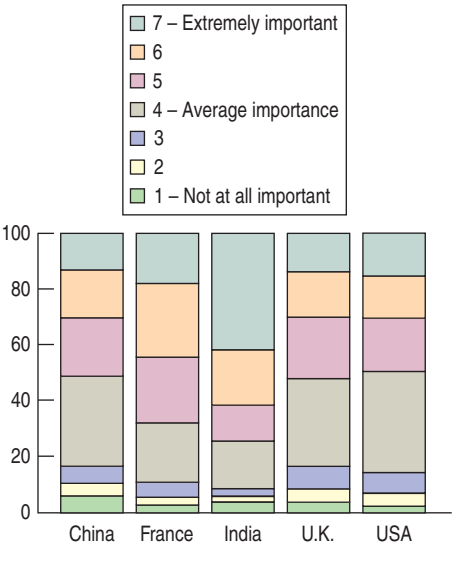
Name the test you will use.

The conditions seem to be met, so we can use a χ^2 model with $(7 - 1) \times (5 - 1) = 24$ degrees of freedom and use a **chi-square test of homogeneity**.

DO

Mechanics Show the expected counts for each cell of the data table. You could make separate tables for the observed and expected counts or put both counts in each cell. A segmented bar chart is often a good way to display the data.

The observed and expected counts are in Tables 15.5 and 15.7. The bar graph shows the column percentages:



	Use software to calculate χ^2 and the associated P-value.	$\chi^2 = 802.64$
	Here, the calculated value of the χ^2 statistic is extremely high, so the P-value is quite small.	P-value $= P(\chi^2_{24} > 802.64) < 0.001$, so we reject the null hypothesis.
REPORT	**Conclusion** State your conclusion in the context of the data. Discuss whether the distributions for the groups appear to be different. For a small table, examine the residuals.	MEMO **Re: Importance of Appearance** Our analysis of the Roper data shows large differences across countries in the distribution of how important respondents say it is for them to look attractive. Marketers of cosmetics are advised to take notice of these differences, especially when selling products to India.

If you find that simply rejecting the hypothesis of homogeneity is a bit unsatisfying, you're in good company. It's hardly a shock that responses to this question differ from country to country especially with samples sizes this large. What we'd really like to know is where the differences were and how big they were. The test for homogeneity doesn't answer these interesting questions, but it does provide some evidence that can help us. A look at the standardized residuals can help identify cells that don't match the homogeneity pattern.

For Example — Testing homogeneity

Question: Although annual inflation in the United States has been low for several years, many Americans fear that inflation may return. In May 2010, a Gallup poll asked 1020 adults nationwide, "Are you very concerned, somewhat concerned, or not at all concerned that inflation will climb?" Does the distribution of responses appear to be the same for Conservatives as Liberals?

Ideology	Very Concerned	Somewhat Concerned	Not at all Concerned	Total
Conservative	232	83	25	**340**
Liberal	143	126	71	**340**
Total	**375 (55.15%)**	**209 (30.74%)**	**96 (14.12%)**	**680**

Answer: This is a test of homogeneity, testing whether the distribution of responses is the same for the two ideological groups. The data are counts, the Gallup poll selected adults randomly (stratified by ideology), and all expected cell frequencies are much greater than 5 (see table below).

There are $(3 - 1) \times (2 - 1)$ or 2 degrees of freedom.

If the distributions were the same, we would expect each cell to have expected values that are 55.15%, 30.74% and 14.12% of the row totals for Very Concerned, Somewhat Concerned and Not at all Concerned respectively. These values can be computed explicitly from:

$$Exp_{ij} = \frac{TotalRow_i \times TotalCol_j}{Table\ Total}$$

So, in the first cell (Conservative, Very Concerned):

$$Exp_{11} = \frac{TotalRow_1 \times TotalCol_1}{Table\ Total} = \frac{340 \times 375}{680} = 187.5$$

(continued)

Expected counts for all cells are:

Expected Numbers	Very Concerned	Somewhat Concerned	Not at all Concerned
Conservative	187.5	104.5	48.0
Liberal	187.5	104.5	48.0

The components $\frac{(Obs - Exp)^2}{Exp}$ are:

Components	Very Concerned	Somewhat Concerned	Not at all Concerned
Conservative	10.56	4.42	11.02
Liberal	10.56	4.42	11.02

Summing these gives $\chi^2 = 10.56 + 4.42 + \cdots + 11.02 = 52.01$, which, with 2 df, has a P-value of < 0.0001.

We, therefore, reject the hypothesis that the distribution of responses is the same for Conservatives and Liberals.

15.5 Comparing Two Proportions

Many employers require a high school diploma. In October 2000, U.S. Department of Commerce researchers contacted more than 25,000 24-year-old Americans to see if they had finished high school and found that 84.9% of the 12,460 men and 88.1% of the 12,678 women reported having high school diplomas. Should we conclude that girls are more likely than boys to complete high school?

The U.S. Department of Commerce gives percentages, but it's easy to find the counts and put them in a table. It looks like this:

	Men	Women	Total
HS diploma	10,579	11,169	**21,748**
No diploma	1,881	1,509	**3,390**
Total	**12,460**	**12,678**	**25,138**

Table 15.8 Numbers of men and women who had earned high school diploma or not, by 2000, in a sample of 25,138 24-year-old Americans.

Overall, $\frac{21{,}748}{25{,}138} = 86.5144\%$ of the sample had received high school diplomas. So, under the homogeneity assumption, we would expect the same percentage of the 12,460 men (or $0.865144 \times 12{,}460 = 10{,}779.7$ men) to have diplomas. Completing the table, the expected counts look like this:

	Men	Women	Total
HS diploma	10,779.7	10,968.3	**21,748**
No diploma	1,680.3	1,709.7	**3,390**
Total	**12,460**	**12,678**	**25,138**

Table 15.9 The expected values.

The chi-square statistic with $(2 - 1) \times (2 - 1) = 1$ df is:

$$\chi^2_1 = \frac{(10579 - 10779.7)^2}{10779.7} + \frac{(11169 - 10968.3)^2}{10968.3} + \frac{(1881 - 1680.3)^2}{1680.3} + \frac{(1509 - 1709.7)^2}{1709.7} = 54.941$$

This has a P-value < 0.001, so we reject the null hypothesis and conclude that the distribution of receiving high school diplomas is different for men and women.

A chi-square test on a 2×2 table, which has only 1 df, is equivalent to testing whether two proportions (in this case, the proportions of men and women with diplomas) are equal. There is an equivalent way of testing the equality of two proportions that uses a z-statistic, and it gives exactly the same P-value. You may encounter the z-test for two proportions, so remember that it's the same as the chi-square test on the equivalent 2×2 table.

Even though the z-test and the chi-square test are equivalent for testing whether two proportions are the same, the z-test can also give a confidence interval. This is crucial here because we rejected the null hypothesis with a large sample size. The confidence interval can tell us how large the difference may be.

Confidence Interval for the Difference of Two Proportions

As we saw, 88.1% of the women and 84.9% of the men surveyed had earned high school diplomas in the United States by the year 2000. That's a difference of 3.2%. If we knew the standard error of that quantity, we could use a z-statistic to construct a confidence interval for the true difference in the population. It's not hard to find the standard error. All we need is the formula[5]:

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

The confidence interval has the same form as the confidence interval for a single proportion, with this new standard error:

$$(\hat{p}_1 - \hat{p}_2) \pm z^*SE(\hat{p}_1 - \hat{p}_2).$$

Confidence interval for the difference of two proportions

When the conditions are met, we can find the confidence interval for the difference of two proportions, $p_1 - p_2$. The confidence interval is

$$(\hat{p}_1 - \hat{p}_2) \pm z^*SE(\hat{p}_1 - \hat{p}_2),$$

where we find the standard error of the difference as

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

from the observed proportions.

The critical value z^* depends on the particular confidence level that you specify.

[5]The standard error of the difference is found from the general fact that the variance of a difference of two independent quantities is the *sum* of their variances. See Chapter 8 for details.

For high school graduation, a 95% confidence interval for the true difference between women's and men's rates is:

$$(0.881 - 0.849) \pm 1.96 \times \sqrt{\frac{(0.881)(0.119)}{12678} + \frac{(0.849)(0.151)}{12460}}$$
$$= (0.0236, 0.0404), \text{ or } 2.36\% \text{ to } 4.04\%.$$

We can be 95% confident that women's rates of having a HS diploma by 2000 were 2.36 to 4.04% higher than men's. With a sample size this large, we can be quite confident that the difference isn't zero. But is it a difference that matters? That, of course, depends on the *reason* we are asking the question. The confidence interval shows us the effect size—or at least, the interval of plausible values for the effect size. If we are considering changing hiring or recruitment policies, this difference may be too small to warrant much of an adjustment even though the difference is statistically "significant." Be sure to consider the effect size if you plan to make a business decision based on rejecting a null hypothesis using chi-square methods.

For Example — A confidence interval for the difference of proportions

Question: In the Gallup poll on inflation (see page 463), 68.2% (232 of 340) of those identifying themselves as Conservative were very concerned about the rise of inflation, but only 42.1% (143 of 340) of Liberals responded the same way. That's a difference of 26.1% in this sample of 680 adults. Find a 95% confidence interval for the true difference.

Answer: The confidence interval can be found from:

$$(\hat{p}_C - \hat{p}_L) \pm z^* SE(\hat{p}_C - \hat{p}_L) \text{ where } SE(\hat{p}_C - \hat{p}_L) = \sqrt{\frac{\hat{p}_C\hat{q}_C}{n_C} + \frac{\hat{p}_L\hat{q}_L}{n_L}} = \sqrt{\frac{(0.682)(0.338)}{340} + \frac{(0.421)(0.579)}{340}} = 0.037.$$

Since we know the 95% confidence critical value for z is 1.96, we have:

$$0.261 \pm 1.96\,(0.037) = (0.188, 0.334).$$

In other words, we are 95% confident that the proportion of Conservatives who are very concerned by inflation is between 18.8% and 33.4% higher than the same proportion of Liberals.

15.6 Chi-Square Test of Independence

We saw that the importance people place on their personal appearance varies a great deal from one country to another, a fact that might be crucial for the marketing department of a global cosmetics company. Suppose the marketing department wants to know whether the age of the person matters as well. That might affect the kind of media channels they use to advertise their products. Do older people feel as strongly as younger people that personal appearance is important?

		Age						
		13–19	20–29	30–39	40–49	50–59	60+	Total
Appearance	7—Extremely important	396	337	300	252	142	93	**1520**
	6	325	326	307	254	123	86	**1421**
	5	318	312	317	270	150	106	**1473**
	4—Average importance	397	376	403	423	224	210	**2033**
	3	83	83	88	93	54	45	**446**
	2	37	43	53	58	37	45	**273**
	1—Not at all important	40	37	53	56	36	52	**274**
	Total	**1596**	**1514**	**1521**	**1406**	**766**	**637**	**7440**

| Table 15.10 Responses to the question about personal appearance by age group.

When we examined the five countries, we thought of the countries as five different groups, rather than as levels of a variable. But here, we can (and probably should) think of *Age* as a second variable whose value has been measured for each respondent along with his or her response to the appearance question. Asking whether the distribution of responses changes with *Age* now raises the question of whether the variables personal *Appearance* and *Age* are independent.

Whenever we have two variables in a contingency table like this, the natural test is a **chi-square test of independence**. Mechanically, this chi-square test is identical to a test of homogeneity. The difference between the two tests is in how we think of the data and, thus, what conclusion we draw.

Here we ask whether the response to the personal appearance question is independent of age. Remember, that for any two events, **A** and **B**, to be independent, the probability of event **A** given that event **B** occurred must be the same as the probability of event **A**. Here, this means the probability that a randomly selected respondent thinks personal appearance is extremely important is the same for all age groups. That would show that the response to the personal *Appearance* question is independent of that respondent's *Age*. Of course, from a table based on data, the probabilities will never be exactly the same. But to tell whether they are different enough, we use a chi-square test of independence.

Homogeneity or Independence?
The only difference between the test for homogeneity and the test for independence is in the decision you need to make.

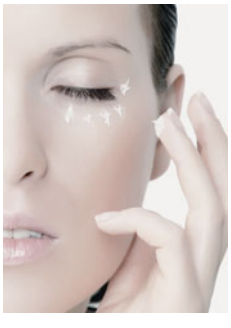
Now we have two categorical variables measured on a single population. For the homogeneity test, we had a single categorical variable measured independently on two or more populations. Now we ask a different question: "Are the variables independent?" rather than "Are the groups homogeneous?" These are subtle differences, but they are important when we draw conclusions.

Assumptions and Conditions

Of course, we still need counts and enough data so that the expected counts are at least five in each cell.

If we're interested in the independence of variables, we usually want to generalize from the data to some population. In that case, we'll need to check that the data are a representative random sample from that population.

Guided Example Personal Appearance and Age



We previously looked at whether responses to the question "How important is seeking the utmost attractive appearance to you?" varied in the five markets of China, France, India, the U.K., and the U.S., and we reported on the cultural differences that we saw. Now we want to help marketers discover whether a person's age influences how they respond to the same question. We have the values of *Age* in six age categories. Rather than six different groups, we can view *Age* as a variable, and ask whether the variables *Age* and *Appearance* are independent.

PLAN

Setup State what you want to know.

Identify the variables and context.

We want to know whether the categorical variables personal Appearance and Age are statistically independent. We have a contingency table of 7440 respondents from a sample of five countries.

(continued)

Hypotheses State the null and alternative hypotheses.

We perform a test of independence when we suspect the variables may not be independent. We are making the claim that knowing the respondents' A age will change the distribution of their response to the question about personal *Appearance*, and testing the null hypothesis that it is *not* true.

H_0: Personal *Appearance* and *Age* are independent.[6]

H_A: Personal *Appearance* and *Age* are not independent.

Model Check the conditions.

- ✓ **Counted Data Condition** We have counts of individuals categorized on two categorical variables.
- ✓ **Randomization Condition** These data are from a randomized survey conducted in 30 countries. We have data from five of them. Although they are not an SRS, the samples within each country were selected to avoid biases.
- ✓ **Expected Cell Frequency Condition** The expected values are all much larger than 5.

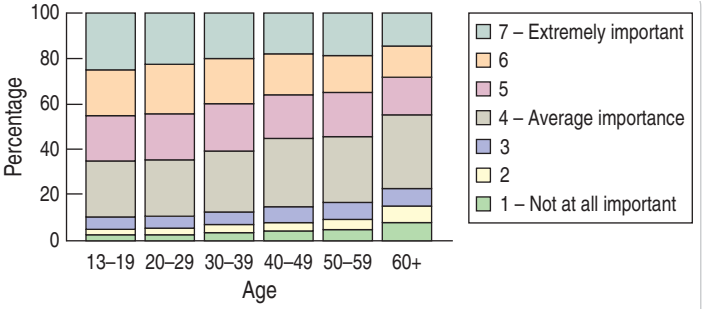
This table shows the expected counts below for each cell. The expected counts are calculated exactly as they were for a test of homogeneity; in the first cell, for example, we expect $\frac{1520}{7440} = 20.43\%$ of 1596 which is 326.065.

Expected Values

Appearance	Age 13–19	20–29	30–39	40–49	50–59	60 +
7—Extremely important	326.065	309.312	310.742	287.247	156.495	130.140
6	304.827	289.166	290.503	268.538	146.302	121.664
5	315.982	299.748	301.133	278.365	151.656	126.116
4—Average importance	436.111	413.705	415.617	384.193	209.312	174.062
3	95.674	90.759	91.178	84.284	45.919	38.186
2	58.563	55.554	55.811	51.591	28.107	23.374
1—Not at all important	58.777	55.758	56.015	51.780	28.210	23.459

The stacked bar graph shows that the response seems to be dependent on *Age*. Older people tend to think personal appearance is less important than younger people.

[6] As in other chi-square tests, the hypotheses are usually expressed in words, without parameters. The hypothesis of independence itself tells us how to find expected values for each cell of the contingency table. That's all we need.



Specify the model.

Name the test you will use.

(The counts are shown in Table 15.10.)

We'll use a χ^2 model with $(7 - 1) \times (6 - 1) = 30$ df and do a **chi-square test of independence**.

DO

Mechanics Calculate χ^2 and find the P-value using software.

The shape of a chi-square model depends on its degrees of freedom. Even with 30 df, this chi-square statistic is extremely large, so the resulting P-value is small.

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp} = 170.7762$$

$$\text{P-value} = P(\chi^2_{30} > 170.7762) < 0.001$$

REPORT

Conclusion Link the P-value to your decision. State your conclusion.

MEMO

Re: Investigation of the relationship between age of consumer and attitudes about personal appearance.

It appears from our analysis of the Roper survey that attitudes on personal *Appearance* are not independent of *Age*. It seems that older people find personal appearance less important than younger people do (on average in the five countries selected).

We rejected the null hypothesis of independence between *Age* and attitudes about personal *Appearance*. With a sample size this large, we can detect very small deviations from independence, so it's almost guaranteed that the chi-square test will reject the null hypothesis. Examining the residuals can help you see the cells that deviate farthest from independence. To make a meaningful business decision, you'll have to look at effect sizes as well as the P-value. We should also look at each country's data individually since country to country differences could affect marketing decisions.

Suppose the company was specifically interested in deciding how to split advertising resources between the teen market and the 30–39-year-old market. How much of a difference is there between the proportions of those in each age group that rated personal *Appearance* as very important (responding either 6 or 7)?

For that we'll need to construct a confidence interval on the difference. From Table 15.10, we find that the percentages of those answering 6 and 7 are 45.17% and 39.91% for the teen and 30–39-year-old groups, respectively. The 95% confidence interval is:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* SE(\hat{p}_1 - \hat{p}_2)$$

$$= (0.4517 - 0.3991) \pm 1.96 \times \sqrt{\frac{(0.4517)(0.5483)}{1596} + \frac{(0.3991)(0.6009)}{1521}}$$

$$= (0.018, 0.087), \text{ or } (1.8\% \text{ to } 8.7\%)$$

This is a statistically significant difference, but now we can see that the difference may be as small as 1.8%. When deciding how to allocate advertising expenditures, it is important to keep these estimates of the effect size in mind.

For Example — A chi-square test of independence

Question: In May 2010, the Gallup poll asked U.S. adults their opinion on whether they are in favor of or opposed to using profiling to identify potential terrorists at airports, a practice used routinely in Israel, but not in the United States. Does opinion depend on age? Or are opinion and age independent? Here are numbers similar to the ones Gallup found (the percentages are the same, but the totals have been changed to make the calculations easier).

	Age				
	18–29	30–49	50–64	65+	Total
Favor	57	66	77	87	**287**
Oppose	43	34	23	13	**113**
Total	**100**	**100**	**100**	**100**	**400**

Answer: The null hypothesis is that *Opinion* and *Age* are independent. We can view this as a test of independence as opposed to a test of homogeneity if we view *Age* and *Opinion* are variables whose relationship we want to understand. This was a random sample and there are at least 5 expected responses in every cell. The expected values are calculated using the formula:

$$Exp_{ij} = \frac{TotalRow_i \times TotalCol_j}{Table\ Total} \Rightarrow$$

$$Exp_{11} = \frac{TotalRow_1 \times TotalCol_1}{Table\ Total} = \frac{287 \times 100}{400} = 71.75$$

Expected Values	Age				
	18–29	30–49	50–64	65+	Total
Favor	71.75	71.75	71.75	71.75	**287**
Oppose	28.25	28.25	28.25	28.25	**113**
Total	**100**	**100**	**100**	**100**	**400**

The components are:

Components	Age			
	18–29	30–49	50–64	65+
Favor	3.03	0.46	0.38	3.24
Oppose	7.70	1.17	0.98	8.23

There are $(r - 1) \times (c - 1) = 1 \times 3 = 3$ degrees of freedom. Summing all the components gives:

$$\chi^2_3 = 3.03 + 0.46 + \cdots + 8.23 = 25.20,$$

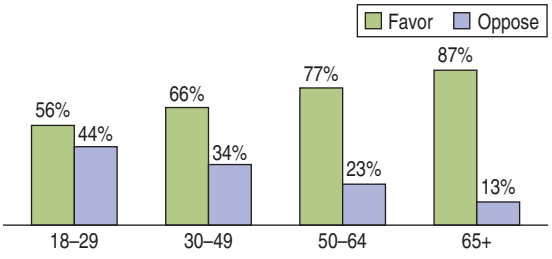
which has a P-value < 0.0001.

Thus, we reject the null hypothesis and conclude that *Age* and *Opinion* about Profiling are not independent. Looking at the residuals,

Residuals	Age			
	18–29	30–49	50–64	65+
Favor	−1.74	−0.68	0.62	1.80
Oppose	2.78	1.08	−0.99	−2.87

we see a pattern. These two variable fail to be independent because increasing age is associated with more favorable attitudes toward profiling.

Bar charts arranged in *Age* order make the pattern clear:



Just Checking

Which of the three chi-square tests would you use in each of the following situations—goodness-of-fit, homogeneity, or independence?

1. A restaurant manager wonders whether customers who dine on Friday nights have the same preferences among the chef's four special entrées as those who dine on Saturday nights. One weekend he has the wait staff record which entrées were ordered each night. Assuming these customers to be typical of all weekend diners, he'll compare the distributions of meals chosen Friday and Saturday.
2. Company policy calls for parking spaces to be assigned to everyone at random, but you suspect that may not be so. There are three lots of equal size: lot A, next to the building; lot B, a bit farther away; and lot C on the other side of the highway. You gather data about employees at middle management level and above to see how many were assigned parking in each lot.
3. Is a student's social life affected by where the student lives? A campus survey asked a random sample of students whether they lived in a dormitory, in off-campus housing, or at home and whether they had been out on a date 0, 1–2, 3–4, or 5 or more times in the past two weeks.

Chi-square tests and causation

Chi-square tests are common. Tests for independence are especially widespread. Unfortunately, many people interpret a small P-value as proof of causation. We know better. Just as correlation between quantitative variables does not demonstrate causation, a failure of independence between two categorical variables does not show a cause-and-effect relationship between them, nor should we say that one variable *depends* on the other.

The chi-square test for independence treats the two variables symmetrically. There is no way to differentiate the direction of any possible causation from one variable to the other. While we can see that attitudes on personal *Appearance* and *Age* are related, we can't say that getting older *causes* you to change attitudes. And certainly it's not correct to say that changing attitudes on personal appearance makes you older.

Of course, there's never any way to eliminate the possibility that a lurking variable is responsible for the observed lack of independence. In some sense, a failure of independence between two categorical variables is less impressive than a strong, consistent association between quantitative variables. Two categorical variables can fail the test of independence in many ways, including ways that show no consistent pattern of failure. Examination of the chi-square standardized residuals can help you think about the underlying patterns.

What Can Go Wrong?

- **Don't use chi-square methods unless you have counts.** All three of the chi-square tests apply only to counts. Other kinds of data can be arrayed in two-way tables. Just because numbers are in a two-way table doesn't make them suitable for chi-square analysis. Data reported as proportions or percentages can be suitable for chi-square procedures, *but only after they are converted to counts.* If you try to do the calculations without first finding the counts, your results will be wrong.
- **Beware large samples.** Beware *large* samples? That's not the advice you're used to hearing. The chi-square tests, however, are unusual. You should be wary of chi-square tests performed on very large samples. No hypothesized distribution fits perfectly, no two groups are exactly homogeneous, and two variables are rarely perfectly independent. The degrees of freedom for chi-square tests don't grow with the sample size. With a sufficiently large sample size, a chi-square test can always reject the null hypothesis. But we have no measure of how far the data are from the null model. There are no confidence intervals to help us judge the effect size except in the case of two proportions.
- **Don't say that one variable "depends" on the other just because they're not independent.** "Depend" can suggest a model or a pattern, but variables can fail to be independent in many different ways. When variables fail the test for independence, it may be better to say they are "associated."

Ethics in Action

Deliberately Different specializes in unique accessories for the home such as hand-painted switch plates and hand-embroidered linens, offered through a catalog and a website. Its customers tend to be women, generally older, with relatively high household incomes. Although the number of customer visits to the site has remained the same, management noticed that the proportion of customers visiting the site who make a purchase has been declining. Megan Cally, the product manager for Deliberately Different, was in charge of working with the market research firm hired to examine this problem. In her first meeting with Jason Esgro, the firm's consultant, she directed the conversation toward website design. Jason mentioned several reasons for consumers abandoning online purchases, the two most common being concerns about transaction security and unanticipated shipping/handling charges. Because Deliberately Different's shipping charges are reasonable, Megan asked him to look further into the issue of security concerns. They developed a survey that randomly sampled customers who had visited the website. They contacted these customers by e-mail and asked them to respond to a brief survey, offering the chance of winning a prize, which would be awarded at random among the respondents. A total of 2450 responses were received. The analysis of the responses included chi-square tests for independence, checking to see if responses on the security question were independent of gender and income category. Both tests were significant, rejecting the null hypothesis of independence. Megan reported to management that concerns about online transaction security were dependent on gender, and income, so Deliberately Different began to explore ways in which they could assure their older female customers that transactions on the website are indeed secure. As product manager, Megan was relieved that the decline in purchases was not related to product offerings.

ETHICAL ISSUE *The chance of rejecting the null hypothesis in a chi-square test for independence increases with sample size. Here the sample size is very large. In addition, it is misleading to state that concerns about security depend on gender, age, and income. Furthermore, patterns of association were not examined (for instance, with varying age categories). Finally, as product manager, Megan intentionally steered attention away from examining the product offerings, which could be a factor in declining purchases. Instead she reported to management that they have pinpointed the problem without noting that they had not explored other potential factors (related to Items A and H, ASA Ethical Guidelines).*

ETHICAL SOLUTION *Interpret results correctly, cautioning about the large sample size and looking for any patterns of association, realizing that there is no way to estimate the effect size.*

What Have We Learned?

Learning Objectives

- Recognize when a chi-square test of goodness of fit, homogeneity, or independence is appropriate.
- For each test, find the expected cell frequencies.
- For each test, check the assumptions and corresponding conditions and know how to complete the test.
  - Counted data condition.
  - Independence assumption; randomization makes independence more plausible.
  - Sample size assumption with the expected cell frequency condition; expect at least 5 observations in each cell.
- Interpret a chi-square test.
  - Even though we might believe the model, we cannot prove that the data fit the model with a chi-square test because that would mean confirming the null hypothesis.
- Examine the standardized residuals to understand what cells were responsible for rejecting a null hypothesis.
- Compare two proportions.
- State the null hypothesis for a test of independence and understand how that is different from the null hypothesis for a test of homogeneity.
  - Both are computed the same way. You may not find both offered by your technology. You can use either one as long as you interpret your result correctly.

Terms

Chi-square models

Chi-square models are skewed to the right. They are parameterized by their degrees of freedom and become less skewed with increasing degrees of freedom.

Chi-square (or chi-squared) statistic

The chi-square statistic is found by summing the chi-square components. Chi-square tests can be used to test goodness-of-fit, homogeneity, or independence.

Chi-square goodness-of-fit test

A test of whether the distribution of counts in one categorical variable matches the distribution predicted by a model. A chi-square test of goodness-of-fit finds

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp},$$

where the expected counts come from the predicting model. It finds a P-value from a chi-square model with $n - 1$ degrees of freedom, where n is the number of categories in the categorical variable.

Chi-square test of homogeneity

A test comparing the distribution of counts for two or more groups on the same categorical variable. A chi-square test of homogeneity finds

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp},$$

where the expected counts are based on the overall frequencies, adjusted for the totals in each group. We find a P-value from a chi-square distribution with $(R - 1) \times (C - 1)$ degrees of freedom, where R gives the number of categories (rows) and C gives the number of independent groups (columns).

Chi-square test of independence

A test of whether two categorical variables are independent. It examines the distribution of counts for one group of individuals classified according to both variables. A chi-square test of *independence* uses the same calculation as a test of homogeneity. We find a P-value from a chi-square distribution with $(R - 1) \times (C - 1)$ degrees of freedom, where R gives the number of categories in one variable and C gives the number of categories in the other.

(continued)

Standardized residual

In each cell of a two-way table, a standardized residual is the square root of the chi-square component for that cell with the sign of the *Observed* – *Expected* difference:

$$\frac{(Obs - Exp)}{\sqrt{Exp}}$$

When we reject a chi-square test, an examination of the standardized residuals can sometimes reveal more about how the data deviate from the null model.

Technology Help: Chi-Square

Most statistics packages associate chi-square tests with contingency tables. Often chi-square is available as an option only when you make a contingency table. This organization can make it hard to locate the chi-square test and may confuse the three different roles that the chi-square test can take. In particular, chi-square tests for goodness-of-fit may be hard to find or missing entirely. Chi-square tests for homogeneity are computationally the same as chi-square tests for independence, so you may have to perform the mechanics as if they were tests of independence and interpret them afterward as tests of homogeneity.

Most statistics packages work with data on individuals rather than with the summary counts. If the only information you have is the table of counts, you may find it more difficult to get a statistics package to compute chi-square. Some packages offer a way to reconstruct the data from the summary counts so that they can then be passed back through the chi-square calculation, finding the cell counts again. Many packages offer chi-square standardized residuals (although they may be called something else).

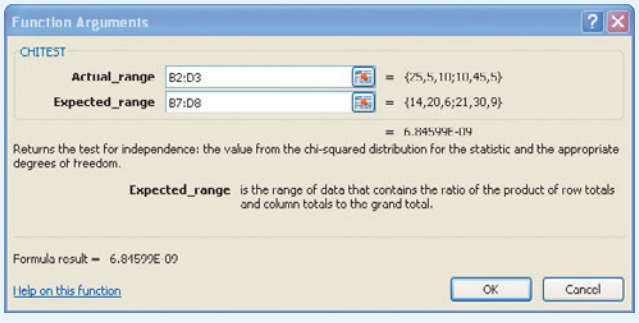
EXCEL

XLSTAT

Excel offers the function **CHITEST (actual_range, expected_range),** which computes a chi-square P-value for independence. Both ranges are of the form UpperLeftCell: LowerRightCell, specifying two rectangular tables. The two tables must be of the same size and shape. The function is called CHISQ.TEST in Excel 2010.

Comments

In order to use this function, you will have to compute the expected values table, typically using the column total divided by the sample size method.



JMP

From the **Analyze** menu,

- Select **Fit Y by X**.
- Choose one variable as the Y, response variable, and the other as the X, factor variable. Both selected variables must be Nominal or Ordinal.
- JMP will make a plot and a contingency table. Below the contingency table, **JMP** offers a **Tests** panel. In that panel, the Chi Square for independence is called a **Pearson ChiSquare.** The table also offers the P-value.
- Click on the contingency Table title bar to drop down a menu that offers to include a **Deviation** and Cell **Chi square** in each cell of the table.

Comments

JMP will choose a chi-square analysis for a **Fit Y by X** if both variables are nominal or ordinal (marked with an N or O), but not otherwise. Be sure the variables have the right type. Deviations are the observed—expected differences in counts. Cell chi-squares are

the squares of the standardized residuals. Refer to the deviations for the sign of the difference. Look under **Distributions** in the **Analyze** menu to find a chi-square test for goodness-of-fit.

MINITAB

From the **Start** menu,

- Choose the **Tables** submenu.
- From that menu, choose **Chi Square Test**. . . .
- In the dialog, identify the columns that make up the table. Minitab will display the table and print the chi-square value and its P-value.

Comments

Alternatively, select the **Cross Tabulation** . . . command to see more options for the table, including expected counts and standardized residuals.

SPSS

From the **Analyze** menu,

- Choose the **Descriptive Statistics** submenu.
- From that submenu, choose **Crosstabs**. . . .
- In the Crosstabs dialog, assign the row and column variables from the variable list. Both variables must be categorical.
- Click the **Cells** button to specify that standardized residuals should be displayed.
- Click the **Statistics** button to specify a chi-square test.

Comments

SPSS offers only variables that it knows to be categorical in the variable list for the Crosstabs dialog. If the variables you want are missing, check that they have the right type.

Health Insurance

In 2010 the U.S. Congress passed the historic health care reform bill that will provide some type of coverage for the 32 million Americans currently without health care insurance. Just how widespread was the lack of medical coverage? The media claims that the segments of the population most at risk are women, children, the elderly and the poor. The tables give the number of uninsured (in thousands) by sex, by age and by household income in 2008.[7] Using the appropriate summary statistics, graphical displays, statistical tests, and confidence intervals, investigate the accuracy of the media's statement using these data. Be sure to discuss your assumptions, methods, results, and conclusions. (Note: some rows and colums may not add exactly to totals due to rounding.)

	Sex		
	Male	Female	Total
Uninsured	25,208	21,131	**46,340**
Insured	122,886	132,257	**255,143**
Total	**148,094**	**153,388**	**301,483**

	Age			
	0-17	18-64	65+	Total
Uninsured	7,348	38,345	646	**46,340**
Insured	67,161	150,841	37,142	**255,143**
Total	**74,510**	**189,185**	**37,788**	**301,483**

(continued)

[7]Source: U.S. Census Bureau, Current Population Survey, Annual Social and Economic Supplement, 2009. http://www.census.gov/hhes/www/cpstables/032009/health/h01_001.htm.

	Household Income				
	<$25,000	$25,000–$49,999	$50,000–$74,999	$75,000+	Total
Uninsured	13,673	14,908	8,034	9,725	**46,340**
Insured	42,142	54,712	49,491	108,798	**255,143**
Total	**55,814**	**69,621**	**57,525**	**118,523**	**301,483**

Loyalty Program

A marketing executive tested two incentives to see what percentage of customers would enroll in a new web-based loyalty program. The customers were asked to log on to their accounts on the Web and provide some demographic and spending information. As an incentive, they were offered either Nothing (No Offer), Free flight insurance on their next flight (Free Insurance), or a free companion Airline ticket (Free Flight). The customers were segmented according to their past year's spending patterns as spending primarily in one of five areas: *Travel*, *Entertainment*, *Dining*, *Household*, or *Balanced*. The executive wanted to know whether the incentives resulted in different enrollment rates (*Response*). Specifically, she wanted to know how much higher the enrollment rate for the free flight was compared to the free insurance. She also wanted to see whether *Spending Pattern* was associated with *Response*. Using the data **Loyalty_Program**, write up a report for the marketing executive using appropriate graphics, summary statistics, statistical tests, and confidence intervals.

Exercises

SECTION 15.1

1. If there is no seasonal effect on human births, we would expect equal numbers of children to be born in each season (winter, spring, summer and fall). A student takes a census of her statistics class and finds that of the 120 students in the class, 25 were born in winter, 35 in spring, 32 in summer and 28 in fall. She wonders if the excess in the spring is an indication that births are not uniform throughout the year.

a) What is the expected number of births in each season if there is no "seasonal effect" on births?
b) Compute the χ^2 statistic.
c) How many degrees of freedom does the χ^2 statistic have?

2. At a major credit card bank, the percentages of people who historically apply for the Silver, Gold and Platinum cards are 60%, 30% and 10% respectively. In a recent sample of customers responding to a promotion, of 200 customers, 110 applied for Silver, 55 for Gold and 35 for Platinum. Is there evidence to suggest that the percentages for this promotion may be different from the historical proportions?

a) What is the expected number of customers applying for each type of card in this sample if the historical proportions are still true?
b) Compute the χ^2 statistic.
c) How many degrees of freedom does the χ^2 statistic have?

SECTION 15.2

3. For the births in Exercise 1,

a) If there is no seasonal effect, about how big, on average, would you expect the χ^2 statistic to be (what is the mean of the χ^2 distribution)?
b) Does the statistic you computed in Exercise 1 seem large in comparison to this mean? Explain briefly.
c) What does that say about the null hypothesis?
d) Find the $\alpha = 0.05$ critical value for the χ^2 distribution with the appropriate number of df.
e) Using the critical value, what do you conclude about the null hypothesis at $\alpha = 0.05$?

4. For the customers in Exercise 2,

a) If the customers apply for the three cards according to the historical proportions, about how big, on average,

would you expect the χ^2 statistic to be (what is the mean of the χ^2 distribution)?
b) Does the statistic you computed in Exercise 2 seem large in comparison to this mean? Explain briefly.
c) What does that say about the null hypothesis?
d) Find the $\alpha = 0.05$ critical value for the χ^2 distribution with the appropriate number of df.
e) Using the critical value, what do you conclude about the null hypothesis at $\alpha = 0.05$?

SECTION 15.3

5. For the data in Exercise 1,

a) Compute the standardized residual for each season.
b) Are any of these particularly large? (Compared to what?)
c) Why should you have anticipated the answer to part b?

6. For the data in Exercise 2,

a) Compute the standardized residual for each type of card.
b) Are any of these particularly large? (Compared to what?)
c) What does the answer to part b say about this new group of customers?

SECTION 15.4

7. An analyst at a local bank wonders if the age distribution of customers coming for service at his branch in town is the same as at the branch located near the mall. He selects 100 transactions at random from each branch and researches the age information for the associated customer. Here are the data:

	Age			
	Less than 30	30–55	56 or older	Total
In-town branch	20	40	40	100
Mall branch	30	50	20	100
Total	100	100	200	200

a) What is the null hypothesis?
b) What type of test is this?
c) What are the expected numbers for each cell if the null hypothesis is true?
d) Find the χ^2 statistic.
e) How many degrees of freedom does it have?
f) Find the critical value at $\alpha = 0.05$.
g) What do you conclude?

8. A market researcher working for the bank in Exercise 2 wants to know if the distribution of applications by card is the same for the past three mailings. She takes a random sample of 200 from each mailing and counts the number applying for Silver Gold and Platinum. The data follow:

	Type of Card			
	Silver	Gold	Platinum	Total
Mailing 1	120	50	30	200
Mailing 2	115	50	35	200
Mailing 3	105	55	40	200
Total	340	155	105	600

a) What is the null hypothesis?
b) What type of test is this?
c) What are the expected numbers for each cell if the null hypothesis is true?
d) Find the χ^2 statistic.
e) How many degrees of freedom does it have?
f) Find the critical value at $\alpha = 0.05$.
g) What do you conclude?

SECTION 15.5

9. Marketers want to know about the differences between men's and women's use of the Internet. A Pew Research poll in April 2009 from a random sample of US adults found that 2393 of 3037 men use the Internet, at least occasionally, while 2378 of 3166 women did.

a) Find the proportions of men and women who said they use the Internet at least occasionally.
b) What is the difference in proportions?
c) What is the standard error of the difference?
d) Find a 95% confidence interval for the difference between percentages of usage by men and women nationwide.

10. From the same poll as in Exercise 9, marketers want to know about the differences in use of the Internet by income. Of 2741 who reported earning less than $50,000 a year, 1813 said they used the Internet. Of the 2353 people who reported earnings of $50,000 a year or more, 2201 said they used the Internet.

a) Find the proportions of those earning at least $50,000 a year and those earning less than $50,000 a year who said they use the Internet.
b) What is the difference in proportions?
c) What is the standard error of the difference?
d) Find a 95% confidence interval for the difference in the proportion of people who use the Internet between the two income groups.

SECTION 15.6

11. The same poll as in Exercise 9 also asked the questions "Did you use the Internet yesterday?" and "Are you White, Black, or Hispanic/Other?" Is the response to the question about the Internet independent of race? The data follow:

	Did You Use the Internet Yesterday?	
	Yes	No
White	2546	856
Black	314	146
Hispanic/Other	431	174

a) Under the null hypothesis, what are the expected values?
b) Compute the χ^2 statistic.
c) How many degrees of freedom does it have?
d) Find the critical value for $\alpha = 0.05$.
e) What do you conclude?

12. The same poll as in Exercise 9 also asked the questions "Did you use the Internet yesterday?" and "What is your educational level?" Is the response to the question about the internet independent of educational level? The data follow:

	Did You Use the Internet Yesterday?	
	Yes	No
Less Than High School	209	131
High School	932	550
Some College	958	346
College Grad	1447	247

a) Under the null hypothesis, what are the expected values?
b) Compute the χ^2 statistic.
c) How many degrees of freedom does it have?
d) Find the critical value for $\alpha = 0.01$.
e) What do you conclude?

CHAPTER EXERCISES

13. Concepts. For each of the following situations, state whether you'd use a chi-square goodness-of-fit test, chi-square test of homogeneity, chi-square test of independence, or some other statistical test.

a) A brokerage firm wants to see whether the type of account a customer has (Silver, Gold, or Platinum) affects the type of trades that customer makes (in person, by phone, or on the Internet). It collects a random sample of trades made for its customers over the past year and performs a test.
b) That brokerage firm also wants to know if the type of account affects the size of the account (in dollars). It performs a test to see if the mean size of the account is the same for the three account types.
c) The academic research office at a large community college wants to see whether the distribution of courses chosen (Humanities, Social Science, or Science) is different for its residential and nonresidential students. It assembles last semester's data and performs a test.

14. Concepts, part 2. For each of the following situations, state whether you'd use a chi-square goodness-of-fit test, a chi-square test of homogeneity, a chi-square test of independence, or some other statistical test.

a) Is the quality of a car affected by what day it was built? A car manufacturer examines a random sample of the warranty claims filed over the past two years to test whether defects are randomly distributed across days of the work week.
b) A researcher for the American Booksellers Association wants to know if retail sales/sq. ft. is related to serving coffee or snacks on the premises. She examines a database of 10,000 independently owned bookstores testing whether retail sales (dollars/sq. ft.) is related to whether or not the store has a coffee bar.
c) A researcher wants to find out whether education level (some high school, high school graduate, college graduate, advanced degree) is related to the type of transaction most likely to be conducted using the Internet (shopping, banking, travel reservations, auctions). He surveys 500 randomly chosen adults and performs a test.

15. Dice. After getting trounced by your little brother in a children's game, you suspect that the die he gave you is unfair. To check, you roll it 60 times, recording the number of times each face appears. Do these results cast doubt on the die's fairness?

Face	Count
1	11
2	7
3	9
4	15
5	12
6	6

a) If the die is fair, how many times would you expect each face to show?
b) To see if these results are unusual, will you test goodness-of-fit, homogeneity, or independence?
c) State your hypotheses.
d) Check the conditions.
e) How many degrees of freedom are there?
f) Find χ^2 and the P-value.
g) State your conclusion.

16. Quality control. Mars, Inc. says that the colors of its M&M's® candies are 14% yellow, 13% red, 20% orange,

24% blue, 16% green and 13% brown. (www.mms.com/us/about/products/milkchocolate). On his way home from work the day he was writing these exercises, one of the authors bought a bag of plain M&M's. He got 29 yellow, 23 red, 12 orange, 14 blue, 8 green, and 20 brown. Is this sample consistent with the company's advertised proportions? Test an appropriate hypothesis and state your conclusion.

a) If the M&M's are packaged in the advertised proportions, how many of each color should the author have expected in his bag of M&M's?
b) To see if his bag was unusual, should he test goodness-of-fit, homogeneity, or independence?
c) State the hypotheses.
d) Check the conditions.
e) How many degrees of freedom are there?
f) Find χ^2 and the P-value.
g) State a conclusion.

17. **Quality control, part 2.** A company advertises that its premium mixture of nuts contains 10% Brazil nuts, 20% cashews, 20% almonds, 10% hazelnuts, and that the rest are peanuts. You buy a large can and separate the various kinds of nuts. Upon weighing them, you find there are 112 grams of Brazil nuts, 183 grams of cashews, 207 grams of almonds, 71 grams of hazelnuts, and 446 grams of peanuts. You wonder whether your mix is significantly different from what the company advertises.

a) Explain why the chi-square goodness-of-fit test is not an appropriate way to find out.
b) What might you do instead of weighing the nuts in order to use a χ^2 test?

18. **Sales rep travel.** A sales representative who is on the road visiting clients thinks that, on average, he drives the same distance each day of the week. He keeps track of his mileage for several weeks and discovers that he averages 122 miles on Mondays, 203 miles on Tuesdays, 176 miles on Wednesdays, 181 miles on Thursdays, and 108 miles on Fridays. He wonders if this evidence contradicts his belief in a uniform distribution of miles across the days of the week. Is it appropriate to test his hypothesis using the chi-square goodness-of-fit test? Explain.

19. **Maryland lottery.** For a lottery to be successful, the public must have confidence in its fairness. One of the lotteries in Maryland is Pick-3 Lottery, where 3 random digits are drawn each day.[8] A fair game depends on every value (0 to 9) being equally likely at each of the three positions. If not, then someone detecting a pattern could take advantage of that and beat the lottery. To investigate the randomness, we'll look at data collected over a recent 32-week period. Although the winning numbers look like three-digit numbers, in fact, each digit is a randomly drawn numeral. We have 654 random digits in all. Are each of the digits from 0 to 9 equally likely? Here is a table of the frequencies.

Group	Count	%
0	62	9.480
1	55	8.410
2	66	10.092
3	64	9.786
4	75	11.468
5	57	8.716
6	71	10.856
7	74	11.315
8	69	10.550
9	61	9.327

a) Select the appropriate procedure.
b) Check the assumptions.
c) State the hypotheses.
d) Test an appropriate hypothesis and state your results.
e) Interpret the meaning of the results and state a conclusion.

20. **Employment discrimination?** Census data for New York City indicate that 29.2% of the under-18 population is white, 28.2% black, 31.5% Latino, 9.1% Asian, and 2% are of other ethnicities. The New York Civil Liberties Union points out that of 26,181 police officers, 64.8% are white, 14.5% black, 19.1% Hispanic, and 1.4% Asian. Do the police officers reflect the ethnic composition of the city's youth?

a) Select the appropriate procedure.
b) Check the assumptions.
c) State the hypotheses.
d) Test an appropriate hypothesis and state your results.
e) Interpret the meaning of the results and state a conclusion.

T **21.** **Titanic.** Here is a table showing who survived the sinking of the *Titanic* based on whether they were crew members or passengers booked in first-, second-, or third-class staterooms.

	Crew	First	Second	Third	Total
Alive	212	202	118	178	**710**
Dead	673	123	167	528	**1491**
Total	**885**	**325**	**285**	**706**	**2201**

[8]Source: Maryland State Lottery Agency, www.mdlottery.com.

a) If we draw an individual at random from this table, what's the probability that we will draw a member of the crew?
b) What's the probability of randomly selecting a third-class passenger who survived?
c) What's the probability of a randomly selected passenger surviving, given that the passenger was in a first-class stateroom?
d) If someone's chances of surviving were the same regardless of their status on the ship, how many members of the crew would you expect to have lived?
e) State the null and alternative hypotheses we would test here (and the name of the test).
f) Give the degrees of freedom for the test.
g) The chi-square value for the table is 187.8, and the corresponding P-value is barely greater than 0. State your conclusions about the hypotheses.

T **22. Promotion discrimination?** The table shows the rank attained by male and female officers in the New York City Police Department (NYPD). Do these data indicate that men and women are equitably represented at all levels of the department? (All possible ranks in the NYPD are shown.)

Rank	Male	Female
Officer	21,900	4281
Detective	4058	806
Sergeant	3898	415
Lieutenant	1333	89
Captain	359	12
Higher ranks	218	10

a) What's the probability that a person selected at random from the NYPD is a female?
b) What's the probability that a person selected at random from the NYPD is a detective?
c) Assuming no bias in promotions, how many female detectives would you expect the NYPD to have?
d) To see if there is evidence of differences in ranks attained by males and females, will you test goodness-of-fit, homogeneity, or independence?
e) State the hypotheses.
f) Test the conditions.
g) How many degrees of freedom are there?
h) Find the chi-square value and the associated P-value.
i) State your conclusion.
j) If you concluded that the distributions are not the same, analyze the differences using the standardized residuals of your calculations.

23. Birth order and college choice. Students in an Introductory Statistics class at a large university were classified by birth order and by the college they attend.

College	Birth Order (1 = oldest or only child) 1	2	3	4 or more	Total
Arts and Sciences	34	14	6	3	57
Agriculture	52	27	5	9	93
Social Science	15	17	8	3	43
Professional	13	11	1	6	31
Total	114	69	20	21	224

Expected Values

College	Birth Order (1 = oldest or only child) 1	2	3	4 or more
Arts and Sciences	29.0089	17.5580	5.0893	5.3438
Agriculture	47.3304	28.6473	8.3036	8.7188
Social Science	21.8839	13.2455	3.8393	4.0313
Professional	15.7768	9.5491	2.7679	2.9063

a) What kind of chi-square test is appropriate—goodness-of-fit, homogeneity, or independence?
b) State your hypotheses.
c) State and check the conditions.
d) How many degrees of freedom are there?
e) The calculation yields $\chi^2 = 17.78$, with $P = 0.0378$. State your conclusion.
f) Examine and comment on the standardized residuals. Do they challenge your conclusion? Explain.

Standardized Residuals

College	Birth Order (1 = oldest or only child) 1	2	3	4 or more
Arts and Sciences	0.92667	−0.84913	0.40370	−1.01388
Agriculture	0.67876	−0.30778	−1.14640	0.09525
Social Science	−1.47155	1.03160	2.12350	−0.51362
Professional	−0.69909	0.46952	−1.06261	1.81476

24. Automobile manufacturers. *Consumer Reports* uses surveys given to subscribers of its magazine and website (www.ConsumerReports.org) to measure reliability in automobiles. This annual survey asks about problems that consumers have had with their cars, vans, SUVs, or trucks during the previous 12 months. Each analysis is based on the number of problems per 100 vehicles.

	Origin of Manufacturer			
	Asia	Europe	U.S.	Total
No Problems	88	79	83	250
Problems	12	21	17	50
Total	100	100	100	300

	Expected Values		
	Asia	Europe	U.S.
No Problems	83.33	83.33	83.33
Problems	16.67	16.67	16.67

a) State your hypotheses.
b) State and check the conditions.
c) How many degrees of freedom are there?
d) The calculation yields $\chi^2 = 2.928$, with P = 0.231. State your conclusion.
e) Would you expect that a larger sample might find statistical significance? Explain.

25. Cranberry juice. It's common folk wisdom that cranberries can help prevent urinary tract infections in women. A leading producer of cranberry juice would like to use this information in their next ad campaign, so they need evidence of this claim. In 2001, the *British Medical Journal* reported the results of a Finnish study in which three groups of 50 women were monitored for these infections over 6 months. One group drank cranberry juice daily, another group drank a lactobacillus drink, and the third group drank neither of those beverages, serving as a control group. In the control group, 18 women developed at least one infection compared with 20 of those who consumed the lactobacillus drink and only 8 of those who drank cranberry juice. Does this study provide supporting evidence for the value of cranberry juice in warding off urinary tract infections in women?

a) Select the appropriate procedure.
b) Check the assumptions.
c) State the hypotheses.
d) Test an appropriate hypothesis and state your results.
e) Interpret the meaning of the results and state a conclusion.
f) If you concluded that the groups are not the same, analyze the differences using the standardized residuals of your calculations.

26. Car company. A European manufacturer of automobiles claims that their cars are preferred by the younger generation and would like to target university students in their next ad campaign. Suppose we test their claim with our own survey. A random survey of autos parked in the student lot and the staff lot at a large university classified the brands by country of origin, as seen in the following table. Are there differences in the national origins of cars driven by students and staff?

		Driver	
		Student	Staff
Origin	American	107	105
	European	33	12
	Asian	55	47

a) Is this a test of independence or homogeneity?
b) Write appropriate hypotheses.
c) Check the necessary assumptions and conditions.
d) Find the P-value of your test.
e) State your conclusion and analysis.

27. Market segmentation. The Chicago Female Fashion Study[9] surveyed customers to determine characteristics of the "frequent" shoppers at different department stores in the Chicago area. Suppose you are a marketing manager at one of the department stores. You would like to know if a customer's shopping frequency and her age are related. Here are the data:

		Age			
		18–24	25–44	45–54	55 or over
Shopping Frequency	Never/Hardly Ever	32	171	45	24
	1–2 times/yr	18	134	40	37
	3–4 times/yr	21	109	48	27
	≥ 5 times/yr	39	134	71	50

		Standardized Residuals Age			
		18–24	25–44	45–54	55 or over
Shopping Frequency	Never/Hardly Ever	0.3803	1.7974	−1.4080	−2.2094
	1–2 times/yr	−1.4326	0.7595	−0.9826	0.9602
	3–4 times/yr	−0.3264	−0.3151	0.9556	−0.2425
	≥ 5 times/yr	1.1711	−2.1360	1.4235	1.4802

[9]Original *Market Segmentation Exercise* prepared by K. Matsuno, D. Kopcso, and D.Tigert, Babson College in 1997 (Babson Case Serics #133-C97A-U).

a) Is this a test of homogeneity or independence?
b) Write an appropriate hypothesis.
c) Are the conditions for inference satisfied?
d) The calculation yields $\chi^2 = 26.084$, P-value = 0.002. State your conclusion.
e) Given the standardized residuals in the table, state a complete conclusion.

28. Seafood company. A large company in the northeastern United States that buys fish from local fishermen and distributes them to major companies and restaurants is considering launching a new ad campaign on the health benefits of fish. As evidence, they would like to cite the following study. Medical researchers followed 6272 Swedish men for 30 years to see if there was any association between the amount of fish in their diet and prostate cancer ("Fatty Fish Consumption and Risk of Prostate Cancer," *Lancet*, June 2001).

Fish Consumption	Prostate Cancer	
	No	Yes
Never/seldom	110	14
Small part of diet	2420	201
Moderate part	2769	209
Large part	507	42

a) Is this a survey, a retrospective study, a prospective study, or an experiment? Explain.
b) Is this a test of homogeneity or independence?
c) Do you see evidence of an association between the amount of fish in a man's diet and his risk of developing prostate cancer?
d) Does this study prove that eating fish does not prevent prostate cancer? Explain.

29. Shopping. A survey of 430 randomly chosen adults finds that 47 of 222 men and 37 of 208 women had purchased books online.

a) Is there evidence that the sex of the person and whether they buy books online are associated?
b) If your conclusion in fact proves to be wrong, did you make a Type I or Type II error?
c) Give a 95% confidence interval for the difference in proportions of buying online for men and women.

30. Information technology. A recent report suggests that Chief Information Officers (CIO's) who report directly to Chief Financial Officers (CFO's) rather than Chief Executive Officers (CEO's) are more likely to have IT agendas that deal with cost cutting and compliance (SearchCIO.com, March 14, 2006). In a random sample of 535 companies, it was found that CIO's reported directly to CFO's in 173 out of 335 service firms and in 95 out of 200 manufacturing companies.

a) Is there evidence that type of business (service versus manufacturing) and whether or not the CIO reports directly to the CFO are associated?
b) If your conclusion proves to be wrong, did you make a Type I or Type II error?
c) Give a 95% confidence interval for the difference in proportions of companies in which the CIO reports directly to the CFO between service and manufacturing firms.

31. Fast food. GfK Roper Consulting gathers information on consumer preferences around the world to help companies monitor attitudes about health, food, and health care products. They asked people in many different cultures how they felt about the following statement: *I try to avoid eating fast foods.*

In a random sample of 800 respondents, 411 people were 35 years old or younger, and, of those, 197 agreed (completely or somewhat) with the statement. Of the 389 people over 35 years old, 246 people agreed with the statement.

a) Is there evidence that the percentage of people avoiding fast food is different in the two age groups?
b) Give a 90% confidence interval for the difference in proportions.

32. Computer gaming. In order to effectively market electronic games, a manager wanted to know what age group of boys played more. A survey in 2006 found that 154 of 223 boys aged 12–14 said they "played computer or console games like Xbox or PlayStation . . . or games online." Of 248 boys aged 15–17, 154 also said they played these games.

a) Is there evidence that the percentage of boys who play these types of games is different in the two age groups?
b) Give a 90% confidence interval for the difference in proportions.

33. Foreclosure rates. The two states with the highest home foreclosure rates in March 2008 were Nevada and Colorado (realestate.msn.com, April 2008). In the second quarter of 2008, there were 8 foreclosures in a random sample of 1098 homes in Nevada, and 6 in a sample of 1460 homes in Colorado.

a) Is there evidence that the percentage of foreclosures is different in the two states?
b) Give a 90% confidence interval for the difference in proportions.

34. Labor force. Immigration reform has focused on dividing illegal immigrants into two groups: long-term and short-term. According to a recent report, short-term unauthorized workers make up nearly 6% of the U.S. labor force in construction (Pew Hispanic Center Fact Sheet, April 13, 2006). The regions of the country with the lowest percentage of unauthorized short-term immigrant construction workers are the Northeast and the Midwest. In a

random sample of 958 construction workers from the Northeast, 66 are illegal short-term immigrants. In the Midwest, 42 out of a sample of 1070 are illegal short-term immigrants.

a) Is there evidence that the percentage of construction workers who are illegal short-term immigrants differs in the two regions?

b) Give a 90% confidence interval for the difference in proportions.

35. Market segmentation, part 2. The survey described in Exercise 27 also investigated the customers' marital status. Using the same definitions for *Shopping Frequency* as in Exercise 27, the calculations yielded the following table. Test an appropriate hypothesis for the relationship between marital status and the frequency of shopping at the same department store as in Exercise 27, and state your conclusions.

Counts				
	Single	Widowed	Married	Total
Never/Hardly Ever	105	5	162	272
1–2 times/yr	53	15	161	229
3–4 times/yr	57	8	140	205
≥ 5 times/yr	72	15	207	294
Total	287	43	670	1000

36. Investment options. The economic slowdown in early 2008 and the possibility of future inflation prompted a full service brokerage firm to gauge the level of interest in inflation-beating investment options among its clients. It surveyed a random sample of 1200 clients asking them to indicate the likelihood that they would add inflation-linked annuities and bonds to their portfolios within the next year. The table below shows the distribution of responses by the investors' tolerance for risk. Test an appropriate hypothesis for the relationship between risk tolerance and the likelihood of investing in inflation linked options.

		Risk Tolerance			
		Averse	Neutral	Seeking	Total
Likelihood of Investing in Inflation-Linked Options	Certain Will Invest	191	93	40	324
	Likely to Invest	82	106	123	311
	Not Likely to Invest	64	110	101	275
	Certain Will not Invest	63	91	136	290
	Total	400	400	400	1200

37. Accounting. The Sarbanes Oxley (SOX) Act was passed in 2002 as a result of corporate scandals and in an attempt to regain public trust in accounting and reporting practices. Two random samples of 1015 executives were surveyed and asked their opinion about accounting practices in both 2000 and in 2006. The table below summarizes all 2030 responses to the question, "Which of the following do you consider most critical to establishing ethical and legal accounting and reporting practices?" Did the distribution of responses change from 2000 to 2006?

		2000	2006
Responses	Training	142	131
	IT Security	274	244
	Audit Trails	152	173
	IT Policies	396	416
	No Opinion	51	51

a) Select the appropriate procedure.
b) Check the assumptions.
c) State the hypotheses.
d) Test an appropriate hypothesis and state your results.
e) Interpret the meaning of the results and state a conclusion.

38. Entrepreneurial executives. A leading CEO mentoring organization offers a program for chief executives, presidents, and business owners with a focus on developing entrepreneurial skills. Women and men executives that recently completed the program rated its value. Are perceptions of the program's value the same for men and women?

		Men	Women
Perceived Value	Excellent	3	9
	Good	11	12
	Average	14	8
	Marginal	9	2
	Poor	3	1

a) Will you test goodness-of-fit, homogeneity, or independence?
b) Write appropriate hypotheses.
c) Find the expected counts for each cell, and explain why the chi-square procedures are not appropriate for this table.

39. Market segmentation, part 3. The survey described in Exercise 27 also investigated the customers' emphasis on *Quality* by asking them the question: "For the same amount

of money, I will generally buy one good item rather than several of lower price and quality." Using the same definitions for *Shopping Frequency* as in Exercise 27, the calculations yielded the following table. Test an appropriate hypothesis for the relationship between a customer's emphasis on *Quality* and the *Shopping Frequency* at this department store.

a) Select the appropriate procedure.
b) Check the assumptions.
c) State the hypotheses.
d) Test an appropriate hypothesis and state your results.
e) Interpret the meaning of the results and state a conclusion.

Counts

	Moderately Disagree	Moderately Disagree/Agree	Moderately Agree	Total
Never/Hardly Ever	15	97	160	**272**
1–2 times/yr	28	107	94	**229**
3–4 times/yr	30	90	85	**205**
≥ 5 times/yr	35	140	119	**294**
Total	**108**	**434**	**458**	**1000**

40. Online shopping. A recent report concludes that while Internet users like the convenience of online shopping, they do have concerns about privacy and security (*Online Shopping*, Washington, DC, Pew Internet & American Life Project, February 2008). Respondents were asked to indicate their level of agreement with the statement "I don't like giving my credit card number or personal information online." The table gives a subset of responses. Test an appropriate hypothesis for the relationship between age and level of concern about privacy and security online.

Age Category	Strongly Agree	Agree	Disagree	Strongly Disagree	Total
Ages 18–29	127	147	138	10	**422**
Ages 30–49	141	129	78	55	**403**
Ages 50–64	178	102	64	51	**395**
Ages 65 +	180	132	54	14	**380**
Total	**626**	**510**	**334**	**130**	**1600**

a) Select the appropriate procedure.
b) Check the assumptions.
c) State the hypotheses.
d) Test an appropriate hypothesis and state your results.
e) Interpret the meaning of the results and state a conclusion.

41. Entrepreneurial executives again. In some situations where the expected counts are too small, as in Exercise 38, we can complete an analysis anyway. We can often proceed after combining cells in some way that makes sense and also produces a table in which the conditions are satisfied. Here is a new table displaying the same data, but combining "Marginal" and "Poor" into a new category called "Below Average."

Perceived Value	Men	Women
Excellent	3	9
Good	11	12
Average	14	8
Below Average	12	3

a) Find the expected counts for each cell in this new table, and explain why a chi-square procedure is now appropriate.
b) With this change in the table, what has happened to the number of degrees of freedom?
c) Test your hypothesis about the two groups and state an appropriate conclusion.

42. Small business. The director of a small business development center located in a mid-sized city is reviewing data about its clients. In particular, she is interested in examining if the distribution of business owners across the various stages of the business life cycle is the same for white-owned and Hispanic-owned businesses. The data are shown below.

Stage in Business	White-Owned	Hispanic-Owned
Planning	11	9
Starting	14	11
Managing	20	2
Getting Out	15	1

a) Will you test goodness-of-fit, homogeneity, or independence?
b) Write the appropriate hypotheses.
c) Find the expected counts for each cell and explain why chi-square procedures are not appropriate for this table.
d) Create a new table by combining categories so that a chi-square procedure can be used.
e) With this change in the table, what has happened to the number of degrees of freedom?
f) Test your hypothesis about the two groups and state an appropriate conclusion.

43. Racial steering. A subtle form of racial discrimination in housing is "racial steering." Racial steering occurs when real estate agents show prospective buyers only homes in neighborhoods already dominated by that family's race. This violates the Fair Housing Act of 1968. According to an article in *Chance* magazine (Vol. 14, no. 2, 2001), tenants at a large apartment complex recently filed a lawsuit alleging racial steering. The complex is divided into two parts: Section A and Section B. The plaintiffs claimed that white potential renters were steered to Section A, while African-Americans were steered to Section B. The following table displays the data that were presented in court to show the locations of recently rented apartments. Do you think there is evidence of racial steering?

	New Renters		
	White	Black	Total
Section A	87	8	95
Section B	83	34	117
Total	170	42	212

44. Titanic, again. Newspaper headlines at the time and traditional wisdom in the succeeding decades have held that women and children escaped the *Titanic* in greater proportion than men. Here's a table with the relevant data. Do you think that survival was independent of whether the person was male or female? Defend your conclusion.

	Female	Male	Total
Alive	343	367	710
Dead	127	1364	1491
Total	470	1731	2201

45. Racial steering, revisited. Find a 95% confidence interval for the difference in the proportions of Black renters in the two sections for the data in Exercise 43.

46. Titanic, one more time. Find a 95% confidence interval for the difference in the proportion of women who survived and the proportion of men who survived for the data in Exercise 44.

47. Industry sector and outsourcing. Many companies have chosen to outsource segments of their business to external providers in order to cut costs and improve quality and/or efficiencies. Common business segments that are outsourced include Information Technology (IT) and Human Resources (HR). The data below show the types of outsourcing decisions made (no outsourcing, IT only, HR only, both IT and HR) by a sample of companies from various industry sectors.

Industry Sector	No Outsourcing	IT Only	HR Only	Both IT and HR
Healthcare	810	6429	4725	1127
Financial	263	1598	549	117
Industrial Goods	1031	1269	412	99
Consumer Goods	66	341	305	197

Do these data highlight significant differences in outsourcing by industry sector?

a) Select the appropriate procedure.
b) Check the assumptions.
c) State the hypotheses.
d) Test an appropriate hypothesis and state your results.
e) Interpret the meaning of the results and state a conclusion.

48. Industry sector and outsourcing, part 2. Consider only the companies that have outsourced their IT and HR business segments. Do these data suggest significant differences between companies in the financial and industrial goods sectors with regard to their outsourcing decisions?

Industry Sector	IT Only	HR Only	Both IT and HR
Financial	1598	549	117
Industrial Goods	1269	412	99

a) Select the appropriate procedure.
b) Check the assumptions.
c) State the hypotheses.
d) Test an appropriate hypothesis and state your results.
e) Interpret the meaning of the results and state the conclusion.

		Employee Job Satisfaction			
		Very Satisfied	Satisfied	Somewhat Satisfied	Not Satisfied
Management Styles	Exploitative Authoritarian	27	82	43	48
	Benevolent Authoritarian	50	19	56	75
	Laissez Faire	52	88	26	34
	Consultative	71	83	20	26
	Participative	101	59	20	20

49. **Management styles.** Use the survey results in the table at the top of the page to investigate differences in employee job satisfaction among organizations in the United States with different management styles.

a) Select the appropriate procedure.
b) Check the assumptions.
c) State the hypotheses.
d) Test an appropriate hypothesis and state your results.
e) Interpret the meaning of the results and state a conclusion.

50. **Ranking companies.** Every year Fortune Magazine lists the 100 best companies to work for, based on criteria such as pay, benefits, turnover rate, and diversity. In 2008, the top three were Google, Quicken Loans, and Wegmans Food Markets (*Fortune*, February 4, 2008). Of the best 100 companies to work for, 33 experienced double digit job growth (10%–68%), 49 experienced single digit job growth (1%–9%), and 18 experienced no growth or a decline. A closer examination of the top 30 showed that 15 had job growth in the double digits, 11 in the single digits, and only 4 had no growth or a decline. Is there anything unusual about job growth among the 30 top companies?

a) Select the appropriate procedure.
b) Check the assumptions.
c) State the hypotheses.
d) Test an appropriate hypothesis and state your results.
e) Interpret the meaning of the results and state a conclusion.

51. **Businesses and blogs.** The Pew Internet & American Life Project routinely conducts surveys to gauge the impact of the Internet and technology on daily life. A recent survey asked respondents if they read online journals or blogs, an Internet activity of potential interest to many businesses. A subset of the data from this survey *(February–March 2007 Tracking Data Set)* shows responses to this question. Test whether reading online journals or blogs is independent of generation.

		Read online journal or blog			
		Yes, Yesterday	Yes, but not Yesterday	No	Total
Generation	Gen-Y (18–30)	29	35	62	**126**
	Gen X (31–42)	12	34	137	**183**
	Trailing Boomers (43–52)	15	34	132	**181**
	Leading Boomers (53–61)	7	22	83	**112**
	Matures (62+)	6	21	111	**138**
	Total	**69**	**146**	**525**	**740**

52. **Businesses and blogs again.** The Pew Internet & American Life Project survey described in Exercise 51 also asked respondents if they ever created or worked on their own online journal or blog. Again, a subset of the data from this survey *(February–March 2007 Tracking Data Set)* shows responses to this question. Test whether creating online journals or blogs is independent of generation.

		Create online journal or blog . . .			
		Yes/ Yesterday	Yes/Not Yesterday	No	Total
Generation	Gen Y (18–30)	18	24	85	**127**
	Gen X (31–42)	6	15	162	**183**
	Boomers (43–61)	5	15	273	**293**
	Matures (62+)	3	3	132	**138**
	Total	**32**	**57**	**652**	**741**

53. **Information systems.** In a recent study of enterprise resource planning (ERP) system effectiveness, researchers asked companies about how they assessed the success of their ERP systems. Out of 335 manufacturing companies surveyed, they found that 201 used return on investment (ROI), 100 used reductions in inventory levels, 28 used improved data quality, and 6 used on-time delivery. In a survey of 200 service firms, 40 used ROI, 40 used inventory levels, 100 used improved data quality, and 20 used on-time delivery. Is there evidence that the measures used to assess ERP system effectiveness differ between service and manufacturing firms? Perform the appropriate test and state your conclusion.

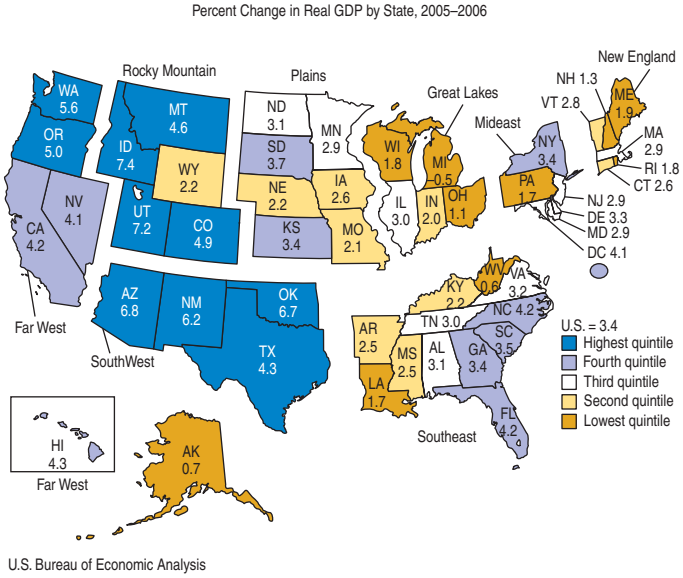
	GDP		
	Top 40%	Bottom 60%	Total
West (Far West, Southwest, and Rocky Mtn.)	5	10	**15**
Midwest (Great Lakes and Plains States)	5	7	**12**
Southeast	5	7	**12**
Northeast (Mideast and New England States)	5	7	**12**
Total	**20**	**31**	**51**

	GDP % Change		
	Top 40%	Bottom 60%	Total
West (Far West, Southwest, and Rocky Mtn.)	13	2	**15**
Midwest (Great Lakes and Plains States)	2	10	**12**
Southeast	4	8	**12**
Northeast (Mideast and New England States)	2	10	**12**
Total	**21**	**30**	**51**

54. **U.S. Gross Domestic Product.** The U.S. Bureau of Economic Analysis provides information on the Gross Domestic Product (GDP) in the United States by state (www.bea.gov). The Bureau recently released figures that showed the real GDP by state for 2007. Using the data in the table at the top of the page, examine if GDP and *Region* of the country are independent. (Alaska and Hawaii are part of the West Region. D.C is included in the Mideast Region.)

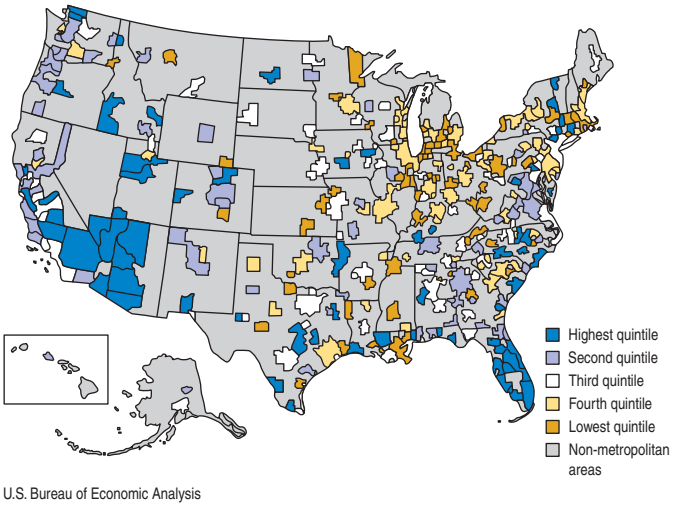
55. **Economic growth.** The U.S. Bureau of Economic Analysis also provides information on the growth of the U.S. economy (www.bea.gov). The Bureau recently released figures that they claimed showed a growth spurt in the western region of the United States. Using the table and map below, determine if the percent change in real GDP by state for 2005–2006 was independent of region of the country. (Alaska and Hawaii are part of the West Region. D.C is included in the Mideast Region.)

56. **Economic growth, revisited.** The U.S. Bureau of Economic Analysis provides information on the GDP in the United States by metropolitan area (www.bea.gov). The Bureau recently released figures that showed the percent change in real GDP by metropolitan area for 2004–2005. Using the data in the following table, examine if there is independence of the growth in metropolitan GDP and region of the country. (Alaska and Hawaii are part of the West Region. Some of the metropolitan areas may have been combined for this analysis.)



	GDP Growth		
	Top two quintiles (top 40%)	Bottom three quintiles (bottom 60%)	Total
West (Far West, Southwest, and Rocky Mtn.)	62	46	**108**
Midwest (Great Lakes and Plains States)	9	87	**96**
Southeast	38	58	**96**
Northeast (Mideast and New England States)	12	36	**48**
Total	**121**	**227**	**348**

Percent Change in Real GDP by Metropolitan Area, 2004–2005



U.S. Bureau of Economic Analysis

Just Checking Answers

1. This is a test of homogeneity. The clue is that the question asks whether the distributions are alike.
2. This is a test of goodness-of-fit. We want to test the model of equal assignment to all lots against what actually happened.
3. This is a test of independence. We have responses on two variables for the same individuals.

CASE Study

Investment Strategy Segmentation

In the aftermath of the financial crisis of 2008, brokerage firms struggled to get individual investors back into the stock market. To gain competitive advantage, market analysts in nearly every industry segment their customers in order to place advertisements where they can have the most impact. The brokerage business is no different. Because different groups of people invest differently, customizing advertising to the needs of these groups leads to more efficient advertising placement and response. Brokerage firms get their information about the investment practices of individuals from a variety of sources, among which the U.S. Census Bureau figures prominently.

The U.S. Census Bureau, with the help of the Bureau of Labor Statistics (BLS) and the Internal Revenue Service (IRS), monitors the incomes and expenditures of Americans. A random sample of Americans are surveyed periodically about their investment practices. In the file CSIII.txt you'll find a random sample of 1000 people from the 48,842 records found in the file Census Income Data set on the University of California at Irvine machine learning repository. This subset was sampled from those that showed some level of investment in the stock market as evidenced by claiming *Capital Gains ($)*, *Capital Losses ($)*, or *Dividends ($)*. Included as well are the demographic variables for these people: *Age (years)*, *Sex (male/female)*, *Union Member (Yes/No)*, *Citizenship (several categories)*, *College (No College/Some)*, *Married (Married/Single)*, *Filer Status (Joint/Single)*.

In order to support her segmentation efforts, a market analyst at an online brokerage firm wants to study differences in investment behaviors. If she can find meaningful differences in the types and amounts of investing that various groups engage in, she can use that information to inform the advertising and marketing departments in their strategies to attract new investors. Using the techniques of Part III, including confidence intervals and hypothesis tests, what differences in investment behaviors can you find among the various demographic groups?

Some specific questions to consider:

1. Do men and women invest similarly? Construct confidence intervals for the differences in mean *Capital Gains*, *Capital Losses* and *Dividends*.

   Be sure to make a suitable display to check assumptions and conditions. If you find outlier(s), consider the analysis with and without the outlier(s).
2. Make a suitable display to compare investment results for the various levels of *Citizenship*.
3. Do those who file singly have the same investment results as those who file jointly? Select, perform, and interpret an appropriate test.
4. Compare differences in investment results for the other demographic variables, being careful to check assumptions and conditions.

   Once again, make suitable displays. Are there outliers to be concerned with? Discuss.

Summarize your findings and conclusions about the investment practices of various groups. Write a short report in order to help the market analyst.

This page intentionally left blank