Business Statistics A First Course 3rd Edition Sharpe Solutions Manual

Chapter 2
Displaying and Describing Categorical Data

What's It About?

We introduce students to distributions of categorical variables. The mathematics is easy (summaries are just percentages) and the graphs are straightforward (pie charts and bar graphs). We challenge students to uncover the story the data tell and to write about it in complete sentences in context.

Then we up the ante, asking them to compare distributions in two-way tables. Constructing comparative graphs, discussing conditional distributions, and considering (informally) the idea of independence give students a look at issues that require deeper thought, careful analysis, and clear writing.

Comments

Most texts do not deal with conditional distributions, independence, and confounding (Simpson's paradox) so soon. Our experience is that students can get lulled into a false sense of security in the early part of this course, if all they see is things like means and histograms that they have dealt with since middle school. They think the course is going to be easy, and they may not recognize the level of sophistication that is required until it's too late. The ideas in Chapter 4 are not hard and are introduced only informally, but they do require some thought. Students will find it difficult to make clear explanations. We want these ideas to be interesting, to engage imaginations, and to challenge students. We hope the level of thought required will get their attention and arouse their interest.

It is probably beginning to dawn on your students that this isn't a math class. At the very least, they are going to be expected to write often and clearly. In business, the numeric solution alone is never sufficient. There is always a real-world understanding or conclusion.

Looking Ahead

There are many important skills and ideas here that prepare students for later topics. They need to think about the type of data, checking a condition before plunging ahead. They need to think about what comparisons will answer the questions posed, and write clear explanations in context. They begin to think about independence, one of the most important issues in Statistics. And, in Simpson's paradox, they see the need to think more deeply to avoid being misled by lurking or confounding variables.

Class Do's

Continue to emphasize precision of vocabulary (and notation). These are an important part of clear communication and critical to success.

Emphasize *Plan-Do-Report* right from the start. The key to doing well in Statistics is to plan carefully by understanding the business questions on the table and ask what statistical techniques can address those issues before starting to write an answer. And then, after doing some calculations or other work, to write clear and concise report of what it all means. Your students may rebel at first at having to write sentences, much less paragraphs, in a course they may have thought was a math class. They are used to just doing the *Do*. *Report* is at least 50% of each solution.

Copyright © 2017 Pearson Education, Inc.

Visit TestBankDeal.com to get complete for all chapters

If you make that point consistently right from the start of the course it becomes second nature soon, and puts each student in the right mindset for writing solid answers. Continually remind them: ***Answers are sentences, not numbers.***

Weave the key step of checking the assumptions and conditions into the fabric of doing Statistics. It's easy: have students check that the data are being treated as categorical before they proceed with pie charts, conditional distributions, and the like. As the course goes on, thinking about assumptions and conditions will help students Plan appropriate statistical procedures. Start now.

Discuss categorical data and appropriate summaries: numerical (counts/percentages) and graphical (pie charts, bar graphs). Discuss *distribution, frequency, relative frequency*.

It gets more interesting when we make comparisons (using bivariate data): for example, social networking by country. Discuss two-way tables, *marginal* and *conditional* distributions. Use of social networking would likely be interesting to many businesses, but looking at the differences in social networking use by across different countries adds much more to the discussion. You can emphasize the vocabulary by asking things like "What is the marginal frequency distribution of social networking?" vs. "What is the conditional relative frequency distribution of social networking in Britain?"

Make sure students can correctly sort out answers to similar sounding questions:

1. What percent of respondents are British internet users who use social networking?
2. What percent of British respondents use social networking?
3. What percent of those who use social networking are in Germany?

Raise the issue of independence. It's not formal independence yet, just the general idea that if social networking and country were independent, the percentages for either gender would mirror the class as a whole, or the percentages of Liberal, Moderate, and Conservative would be the same for both genders. If they are not, we encounter what the politicians refer to as the "gender gap." Statisticians would say this indicates that voting preference is not independent of gender.

Pay attention in each chapter to the What Can Go Wrong? (WCGW) sections. Helping students avoid common pitfalls is one of the keys to success in this course.

Simpson's Paradox is fun, but don't overemphasize it. It's not a critical issue, but it's a good discussion point about making valid comparisons, and not overlooking lurking or confounding variables.

The Importance of What You Don't Say

Probability. You can see that we are patrolling the perimeter of probability. Concepts like relative frequency, conditional relative frequency, and independence cry out for a formal discussion in probabilistic terms. Don't heed the cry. You and we know that we are setting up the habits of thought that students will need for learning about probability. But this isn't the time to discuss the formalities. Or even to say the word "probability" out loud. (Notice that the book doesn't use the term in this chapter at all.) Talk about "relative frequency" instead. In this class probability is a relative frequency, so we are encouraging students to think about the concepts correctly.

Copyright © 2017 Pearson Education, Inc.

Class Examples

1. If you have collected class data about gender and political view, you can use it here. Help students develop their Plan-Do-Report skills with questions like:
   - What percent of the class are women with liberal political views?
   - What percent of the liberals are women?
   - What percent of the women are liberals?
   - What is the marginal frequency distribution of political views?
   - What is the conditional relative frequency distribution of gender among conservatives?
   - Are gender and political view independent?
2. Is the color distribution of M&Ms independent of the type of candy? Break open bags of plain and peanut M&Ms and count the colors. (Then eat the data…)
3. Simpson's paradox example:
   It's the last inning of important game. Your team is a run down with the bases loaded and two outs. The pitcher is due up, so you'll be sending in a pinch-hitter. There are 2 batters available on the bench. Whom should you send in to bat? First show the students the overall success history of the two players.

Player	Overall	vs LHP	vs RHP
A	33 for 103	28 for 81	5 for 22
B	45 for 151	12 for 32	33 for 119

A's batting average is higher than B's (.320 vs. .298), so he looks like the better choice. Someone, though, will raise the issue that it matters whether the pitcher throws right- or left-handed. Now add the rest of the table. It turns out that B has a higher batting average against both right- and left-handed pitching, even though his overall average is lower. Students are stunned.

Here's an explanation. B hits better against both right- and left-handed pitchers. So no matter the pitcher, B is a better choice. So why is his batting "average" lower? Because B sees a lot more right-handed pitchers than A, and (at least for these guys) right-handed pitchers are harder to hit. For some reason, A is used mostly against left-handed pitchers, so A has a higher average.

Suppose all you know is that A bats .227 against righties and .346 against lefties. Ask the students to guess his overall batting average. It could be anywhere between .227 and .346, depending on how many righties and lefties he sees. And B's batting average may slide between .277 and .375. These intervals overlap, so it's quite possible that A's batting average is either higher or lower than B's, depending on the mix of pitchers they see.

Pooling the data loses important information and leads to the wrong conclusion. We always should take into account any factor that might matter.

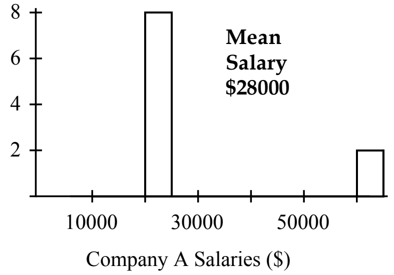
Copyright © 2017 Pearson Education, Inc.

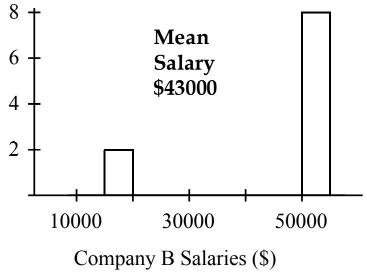
4. Refer to Simpson, again. Here's a nice thought problem to pose to the class; give them a few minutes to work it out. Two companies have labor and management classifications of employees. Company A's laborers have a higher average salary than company B's, as do Company A's managers. But overall company B pays a higher average salary. How can that be? And which is the better way to compare earning potential at the two companies?

Solution:

Make sure you first point out that this example deals with quantitative variables, not categorical. The paradox can be explained when you realize that Company A must employ a greater percentage of laborers than Company B. Also, Company A must employ a smaller percentage of managers than Company B. If laborers earn salaries that are considerably lower than managers, the salaries of Company A's laborers will pull the company average down, and the salaries of Company B's managers will pull the company average up.

The proper way to compare the companies is to use the salaries that are broken down by job type. Using the overall average salary leads to a misleading conclusion.





Basic Exercises

1. The following data show responses to the question "What is your primary source for news?" from a sample of college students.

Internet	Newspaper	Internet	TV	Internet
Newspaper	TV	Internet	Internet	TV
Newspaper	TV	TV	Newspaper	TV
Internet	Internet	Internet	Internet	Internet
TV	Internet	Internet	TV	TV

a. Prepare a frequency table for these data.
b. Prepare a relative frequency table for these data.
c. Based on the frequencies, construct a bar chart.
d. Based on relative frequencies, construct a pie chart.

2. A cable company surveyed its customers and asked how likely they were to bundle other services, such as phone and Internet, with their cable TV. The following data show the responses.

Very Likely	Unlikely	Unlikely	Very Likely
Likely	Unlikely	Likely	Likely
Unlikely	Unlikely	Likely	Likely
Very Likely	Unlikely	Unlikely	Very Likely
Unlikely	Unlikely	Unlikely	Likely

Copyright © 2017 Pearson Education, Inc.

a. Prepare a frequency table for these data.
b. Prepare a relative frequency table for these data.
c. Based on frequencies, construct a bar chart.
d. Based on relative frequencies, construct a pie chart.

3. A membership survey at a local gym asked whether weight loss or fitness was the primary goal for joining. Of 200 men surveyed, 150 responded fitness and the rest responded weight loss. Of 250 women surveyed, 175 responded weight loss and the rest responded fitness.

a. Construct a contingency table.
b. How many members have fitness as their primary goal for joining the gym?
c. How many members have weight loss as their primary goal for joining the gym?
d. Based on the results, should the owner of the gym emphasize one goal over the other? Explain.

4. The following contingency table shows students by major and home state for a small private school in the northeast U.S.

Major Program of Study

Home State	*Biology*	*Accounting*	*History*	*Education*
PA	80	65	55	100
NJ	50	40	65	95
NY	75	50	45	80
MD	65	55	40	40

a. Give the marginal frequency distribution for home state.
b. Give the marginal frequency distribution for major program of study.
c. What percentage of students major in accounting and come from PA?
d. What percentage of students major in education and come from NY?

5. The following contingency table shows students by major and home state for a small private school in the northeast U.S.

Major Program of Study

Home State	*Biology*	*Accounting*	*History*	*Education*
PA	80	65	55	100
NJ	50	40	65	95
NY	75	50	45	80
MD	65	55	40	40

a. Find the conditional distribution (in percentages) of major distribution for the home state of NJ.
b. Find the conditional distribution (in percentages) of major distribution for the home state of MD.
c. Construct segmented bar charts for these two conditional distributions.
d. What can you say about these two conditional distributions?

Copyright © 2017 Pearson Education, Inc.

6. The following contingency table shows students by major and home state for a small private school in the northeast U.S.

Major Program of Study

Home State	*Biology*	*Accounting*	*History*	*Education*
PA	80	65	55	100
NJ	50	40	65	95
NY	75	50	45	80
MD	65	55	40	40

a. Find the conditional distribution (in percentages) of home state distribution for the biology major.
b. Find the conditional distribution (in percentages) of home state distribution for the education major.
c. Construct segmented bar charts for these two conditional distributions.
d. What can you say about these two conditional distributions?

Copyright © 2017 Pearson Education, Inc.

ANSWERS

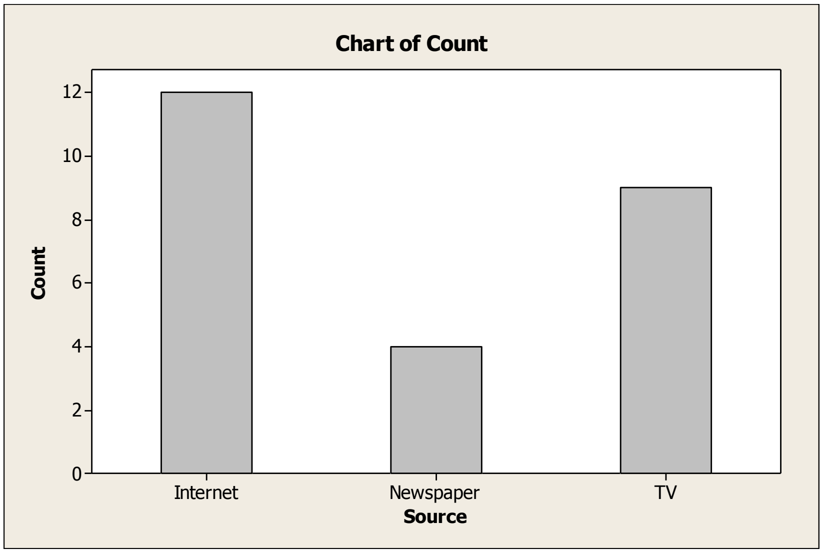
1. a.

News Source	***Number of Students***
Internet	12
Newspaper	4
TV	9

b.

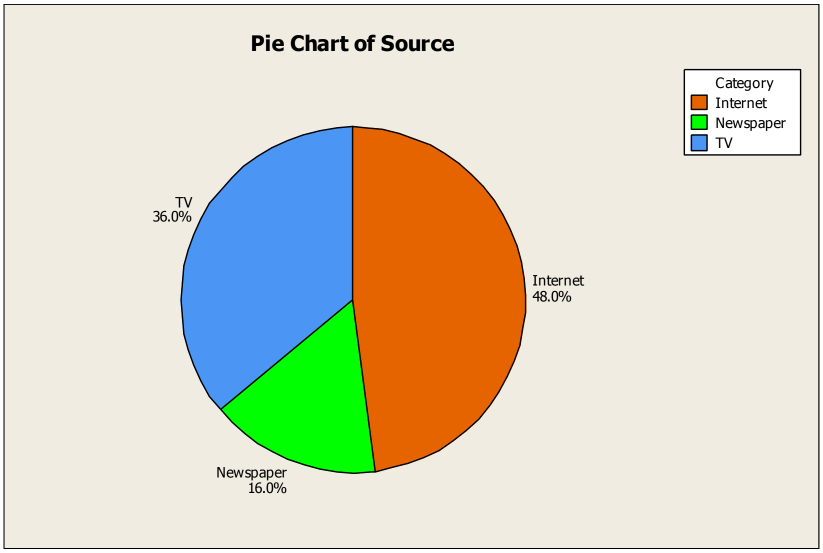
News Source	***% of Students***
Internet	48 %
Newspaper	16 %
TV	36 %

c.



Copyright © 2017 Pearson Education, Inc.

d.



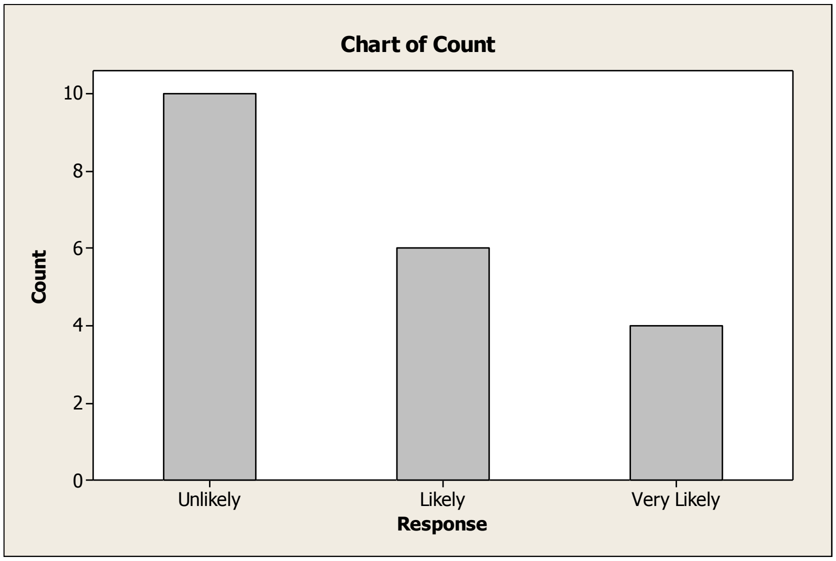
2. a.

Response	***Number of Consumers***
Unlikely	10
Likely	6
Very Likely	4

b.

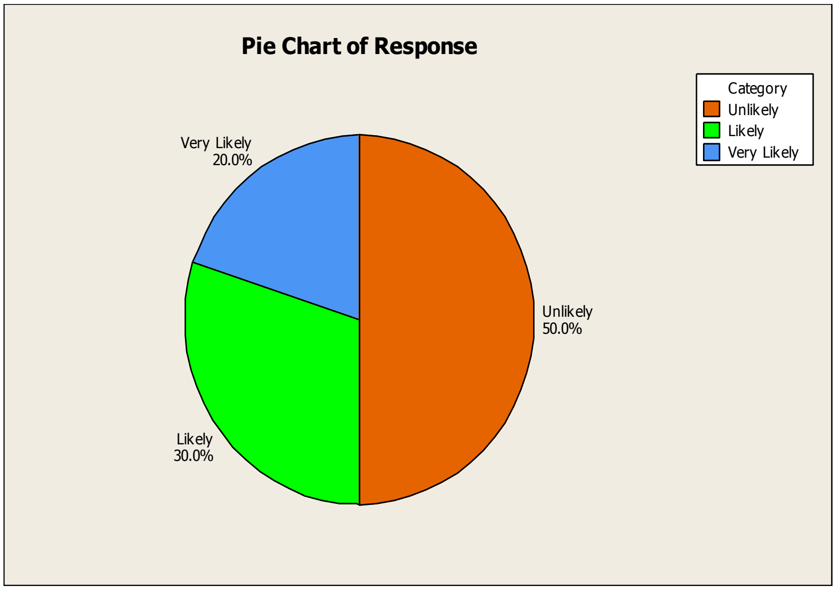
Response	***% of Consumers***
Unlikely	50 %
Likely	30 %
Very Likely	20 %

c.



Copyright © 2017 Pearson Education, Inc.

d.



3. a.

Goal for Gym Membership

Gender	*Fitness*	*Weight Loss*	*Total*
Men	150	50	200
Women	75	175	250
Total	225	225	450

b. 225

c. 225

d. No. 50% of the membership is pursuing each goal.

4. a.

Home State	***Number of Students***
PA	300
NJ	250
NY	250
MD	200

b.

Major	***Number of Students***
Biology	270
Accounting	210
History	205
Education	315

c. 6.5 %

d. 8 %

Copyright © 2017 Pearson Education, Inc.

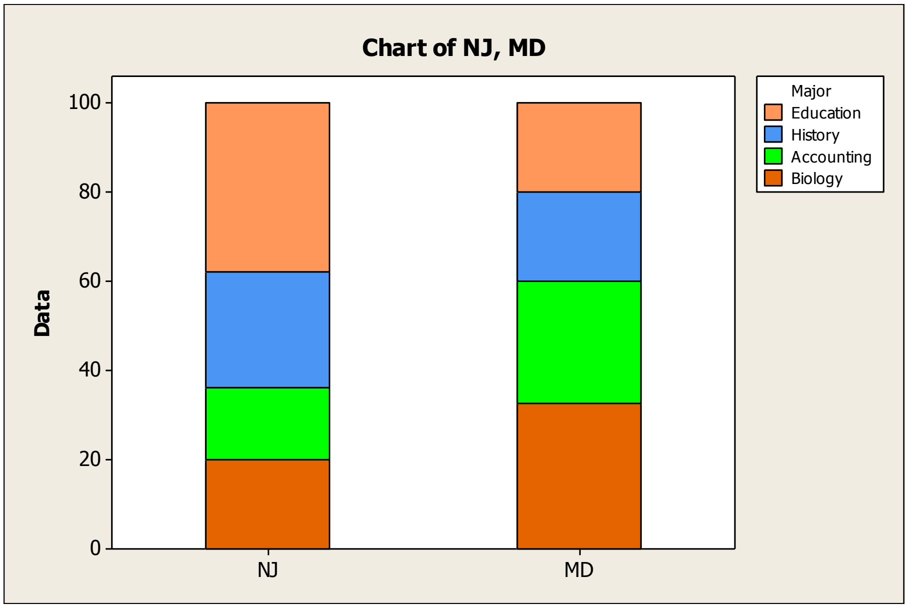
5. a.

Major	***Conditional for NJ***
Biology	20 %
Accounting	16 %
History	26 %
Education	38 %

b.

Major	***Conditional for MD***
Biology	32.5 %
Accounting	27.5 %
History	20 %
Education	20 %

c.



d. More biology and accounting majors come from MD compared to NJ.

6. a.

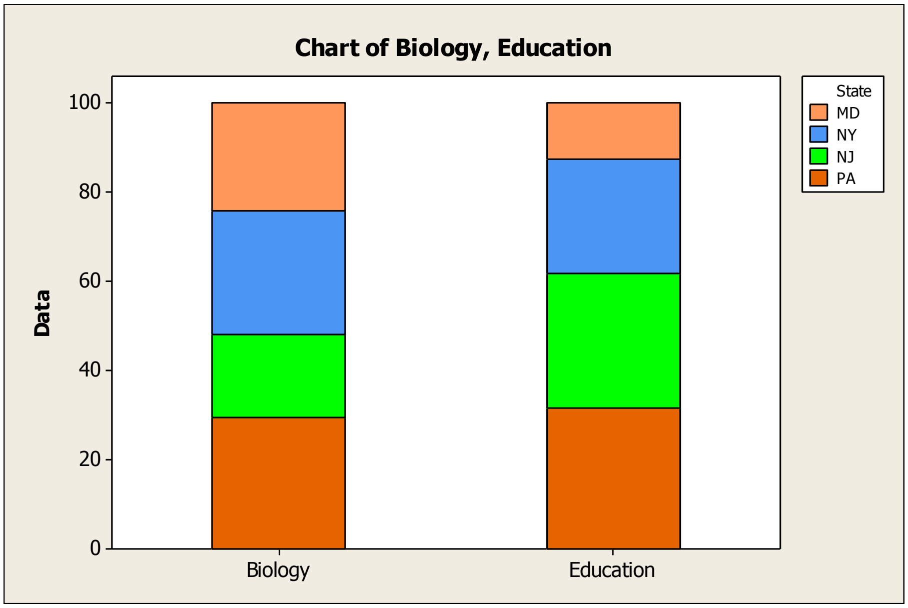
Home State	***Conditional for Biology***
PA	29.6 %
NJ	18.5 %
NY	27.8 %
MD	24.1 %

b.

Home State	***Conditional for Education***
PA	31.7 %
NJ	30.2 %
NY	25.4 %
MD	12.7

Copyright © 2017 Pearson Education, Inc.

c.



d. Fewer education majors are from MD and more are from NJ compared with biology majors.

Copyright © 2017 Pearson Education, Inc.

Chapter 15 – Multiple Regression

SECTION EXERCISES

SECTION 15.1

1.

a) $\hat{y} = 20,986.09 - 7483.10(2) + 93.84(1000) = \$99,859.89.$

b) The residual for the house that just sold for $135,000 is $135,000 - $99,859.89 = $35,140.11.

c) The house sold for more than the model predicted. The model underestimated the selling price.

2.

a) $\hat{y} = 28.4 + 11.37(15) + 2.91(20) = 257.15$ calories

b) The residual for this candy that has 227 calories per serving is 227 – 257.15 = –30.15 calories.

c) The candy has fewer calories that the model predicted. The model overestimated the calories in her candy.

SECTION 15.2

3.

a) $\widehat{USGross} = -22.9898 + 1.13442 Budget + 24.9724 Stars - 0.403296 RunTime$

b) After allowing for the effects of *Stars* and *RunTime*, each additional million dollars in the budget for making the film yields about 1.13 million dollars in gross revenue.

4.

The manager is incorrectly interpreting the coefficient causally. The model says that longer films had smaller gross incomes (after allowing for budget and Stars), but it doesn't say that making a movie shorter will increase its gross. In fact, cutting arbitrarily would, for example, probably reduce the Star rating. Also, the coefficient is negative in the presence of the two other variables.

SECTION 15.3

5.

a) **Linearity**: The scatterplot shows that the relationship is reasonably linear; no curvature is evident in the scatterplot. The linearity condition is satisfied.

b) **Equal Spread:** The scattering of points seems to increase as the budget increases. The points are much more scattered (spread apart) to the right than to the left. Consequently, it appears that the equal spread condition is not satisfied.

c) **Normality:** A scatterplot of two variables gives us no information about the distribution of the residuals. Therefore we cannot determine if the normality assumption is satisfied.

6.

a) **Linearity**: A histogram gives us no information about the form of the relationship between two variables. Therefore, we cannot determine if the linearity condition is satisfied.

b) **Nearly Normal Condition:** The histogram shows that the distribution is unimodal and slightly skewed to the right. However, the presence of an outlier to the right is apparent in the histogram. This would violate the nearly normal condition.

c) **Equal Spread condition:** While a histogram shows the spread of a distribution, a histogram does not show if the spread is consistent for various values of *x*. We would need to see either a scatterplot of the data or a residual plot to determine if this were the case. Therefore, we cannot determine if the equal spread condition is satisfied.

Copyright © 2017 Pearson Education, Inc.

SECTION 15.4

7.

a) The null hypothesis for testing the coefficient associated with *Stars* is: H$_0$: $\beta_{Stars} = 0$.

b) The *t*-statistic for this test is given by:

$$t_{n-k-1} = \frac{b_j - 0}{SE(b_j)}$$

$$t_{120-3-1} = \frac{24.9724}{5.884} = 4.24$$

c) The associated P-value is ≤ 0.0001.

d) We can reject the null hypothesis. There is sufficient evidence to suggest that the coefficient of *Stars* is not equal to zero. *Stars* is a significant predictor of a film's gross income.

8.

a) The null hypothesis for testing the coefficient associated with *Run Time* is: H$_0$: $\beta_{RunTime} = 0$.

b) The *t*-statistic for this test is given by:

$$t_{n-k-1} = \frac{b_j - 0}{SE(b_j)}$$

$$t_{120-3-1} = \frac{-0.403296}{0.2513} = -1.60$$

c) The *t*-statistic is negative because the regression coefficient for *Run Time* is negative.

d) The associated P-value is 0.1113.

e) We cannot reject the null hypothesis (at $\alpha = 0.05$). There is not sufficient evidence to suggest that the coefficient of *Run Time* is not equal to zero. *Run Time* is not a significant predictor of a film's gross income.

SECTION 15.5

9.

a) For this regression equation, $R^2 = 0.474$ or 47.4%. This tells us that 47.4% of the variation in the dependent variable *USGross* is explained by the regression model that includes the predictors *Budget, Run Time* and *Stars.*

b) The *Adjusted* R^2 value is slightly lower than the value of R^2. This is because the *Adjusted* R^2 is adjusted downward (imposing a "penalty") for each new predictor variable added to the model. Consequently, it allows the comparison of models with different numbers of predictor variables.

10.

a) To compute R^2 from the values in the table, we have

$$R^2 = \frac{SSR}{SST} = \frac{224995}{474794} = 0.474 \text{ or } 47.4\%$$

b) From the table, we see that $F = 34.8$.

c) The null hypothesis tested with the *F*-statistic is

$$H_0 : \beta_{Budget} = \beta_{Stars} = \beta_{RunTime} = 0$$

d) The P-value (see table) is very small. Therefore we can reject the null hypothesis and conclude that at least one slope coefficient is significantly different from zero (or that at least one predictor is significant in explaining *USGross*.

Copyright © 2017 Pearson Education, Inc.

CHAPTER EXERCISES

11. Police salaries 2013.

a) **Linearity condition**: The scatterplots appear at least somewhat linear but there is a lot of scatter.
Randomization condition (Independence Assumption): States may not be a random sample but may be independent of each other.
Equal spread condition (Equal Variance Assumption): The scatterplot of Violent Crime vs Police Officer Wage looks less spread to the right but may just have fewer data points. Residual plots are not provided to analyze.
Nearly Normal condition (Normality Assumption): To check this condition, we will need to look at the residuals which are not provided for this example.

b) The R^2 of that regression would be $(0.051)^2 = 0.0026 = 0.26\%$.

12. Ticket prices.

a) **Linearity condition**: The first two scatterplots appear linear. The last one has a lot of scatter.
Randomization condition (Independence Assumption): These data are collected over time and may not be mutually independent. We should check for time-related patterns.
Equal spread condition (Equal Variance Assumption): The scatterplots show no tendency to thicken or spread.
Nearly Normal condition (Normality Assumption): To check this condition, we will need to look at the residuals which are not provided for this example.

b) The R^2 of that regression would be $(0.961)^2 = 0.9235 = 92.35\%$.

13. Police salaries, part 2.

a) The regression model:

$$\widehat{\mathit{Violent\ Crime}} = 1370.22 + 0.795\mathit{Police\,Officer\,Wage} - 12.641\mathit{Graduation\ Rate}$$

b) After allowing for the effects of *Graduation Rate*, states with higher *Police Officer Wages* have more *Violent Crime* at the rate of 0.7953 crimes per 100,000 for each dollar per hour of average wage.

c) $\widehat{\mathit{Violent\ Crime}} = 1370.22 + 0.795*20 - 12.641*70 = 501.25$ crimes per 100,000.

d) The prediction is not very good. The R^2 of that regression is only 22.4%.

14. Ticket prices, part 2.

a) The regression model:

$$\widehat{\mathit{Receipts}} = -18.32 + 0.076\mathit{Paid\ Attendance} + 0.007\,\#\mathit{Shows} + 0.24\mathit{Average\,Ticket\ Price}$$

b) After allowing for the effects of the *# Shows* and *Average Ticket Price*, each thousand customers account for about \$76,000 in receipts. That's about \$76 per customer, which is very close to the average ticket price.

c) $\widehat{\mathit{Receipts}} = -18.32 + 0.076\times 200 + 0.007\times 30 + 0.24\times 70 = \$13.89$ million

d) The prediction is very good. The R^2 of that regression is 99.9%.

15. Police salaries, part 3.

a) $0.221 = \dfrac{0.7947}{3.598}$

b) There are 50 states used in this model. The degrees of freedom are shown to be 47, which is equal to $n - k - 1 = 50 - 2 - 1 = 47$. There are two predictors.

c) The *t*-ratio is negative because the coefficient is negative meaning that *Graduation Rate* contributes negatively to the regression.

Copyright © 2017 Pearson Education, Inc.

16. Ticket prices, part 3.

a) $126.7 = \dfrac{0.076}{0.0006}$

b) There are 78 weeks used in this model. The degrees of freedom are shown to be 74, which is equal to $n - k - 1 = 78 - 3 - 1 = 74$. There are three predictors.

c) The *t*-ratio is negative because the coefficient is negative, meaning that the intercept is negative.

17. Police salaries, part 4.

a) The hypotheses are: $\text{H}_0 : \beta_{Officer} = 0$; $\text{H}_A : \beta_{Officer} \neq 0$.

b) P = 0.8262 which is not small enough to reject the null hypothesis at $\alpha = 0.05$ and conclude that the coefficient is different from zero.

18. Ticket prices, part 4.

a) The hypotheses are: $\text{H}_0 : \beta_{\#Shows} = 0$; $\text{H}_A : \beta_{\#Shows} \neq 0$.

b) P = 0.116 which is too large to reject the null hypothesis at $\alpha = 0.05$. The coefficient may be zero. Practically speaking, the # *Shows* doesn't contribute to this regression model.

c) The coefficient of # *Shows* reports the relationship after allowing for the effects of *Paid Attendance* and *Average Ticket Price*. The scatterplot and correlation were only concerned with the relationship between *Receipts* and # *Shows* (two variables).

19. Police salaries, part 5. This is a causal interpretation, which is not supported by regression. For example, among states with high graduation rates, it may be that those with higher violent crime rates spend more to hire police officers, or states with higher costs of living must pay more to attract qualified police officers but also have higher crime rates.

20. Ticket prices, part 5. This is a causal interpretation, which is not supported by regression. Also, it attempts to interpret a coefficient without taking account of the other variables in the model. For example, the number of paid attendees is surely related to the number of shows. If there were fewer shows, there would be fewer attendees.

21. Police salaries, part 6.
Constant Variance Condition (Equal Spread): met by the residuals vs. predicted plot.
Nearly Normal Condition: met by the Normal probability plot.

22. Ticket prices, part 6.
Constant Variance Condition (Equal Spread): met by the residuals vs. predicted plot.
Nearly Normal Condition: met by the Normal probability plot
Independence Assumption: There doesn't appear to be very much autocorrelation when the residuals are plotted against time.

23. Real estate prices.

a) Incorrect: Doesn't mention other predictors; suggests direct relationship between only two variables: *Age* and *Price*.

b) Correct

c) Incorrect: Can't predict *x* from *y*

d) Incorrect interpretation of R^2 (this model accounts for 92% of the of the variability in *Price*)

24. Wine prices.

a) Incorrect: Doesn't mention other predictors; suggests direct relationship between only two variables: *Age* and *Price*.

b) Incorrect interpretation of R^2 (this model accounts for 92% of the of the variability in *Price*)

c) Incorrect: Doesn't mention other predictors; suggests direct relationship between only two variables: *Tasting Score* and *Price*.

d) Correct

Copyright © 2017 Pearson Education, Inc.

25. Appliance sales.

- **a)** Incorrect: This is likely to be extrapolation since it is unlikely that they observed any data points with no advertising of any kind.
- **b)** Incorrect: Suggests a perfect relationship
- **c)** Incorrect: Can't predict one explanatory variable (x) from another
- **d)** Correct

26. Wine prices, part 2.

- **a)** Incorrect: Doesn't mention other predictors
- **b)** Correct
- **c)** Incorrect: Can't predict one explanatory variable (x) from another
- **d)** Incorrect: Can't predict x from y

27. Cost of pollution.

- **a)** The negative sign of the coefficient for ln(*number of employees*) means that for businesses that have the same amount of sales, those with more employees spend less per employee on pollution abatement on average. The sign of the coefficient for ln(*sales*) is positive. This means that for businesses with the same number of employees, those with larger sales spend more on pollution abatement on average.
- **b)** The logarithms mean that the effects become less severe (in dollar terms) as companies get larger either in *Sales* or in *Number of Employees*.

28. OECD economic regulations.

- **a)** No, it says that *after allowing for the effects of all the other predictors in the model,* the effect of more regulation on GDP is negative.
- **b)** The F is clearly significant. We can be confident that the regression coefficients aren't all zero.
- **c)** It makes sense that 1988 GDP is a good predictor of 1998 GDP. All the other predictors are only helpful after taking into consideration this one variable.

29. Home prices.

- **a)** $\widehat{Price} = -152,037 + 9530 Baths + 139.87 Area$
- **b)** R^2 =71.1%
- **c)** For houses with the same number of bathrooms, each square foot of area is associated with an increase of $139.87 in the price of the house, on average.
- **d)** The regression model says that for houses of the same size, there is no evidence that those with more bathrooms are priced higher. It says nothing about what would actually happen if a bathroom were added to a house.

30. Home prices, part 2. The residuals are right skewed, and the residuals versus fitted values plot shows a possible outlier, which may be the cause of the skewness in the other residuals. The outlier should be examined and either corrected or set aside and the regression recomputed. In addition, there is a clear pattern (negative linear before the outlier) in the residual vs. fitted plot.

31. Secretary performance.

- **a)** The regression equation:

$$\widehat{Salary} = 9.788 + 0.110 Service + 0.053 Education + 0.071 Test\ Score + 0.004 Typing\ wpm + 0.065 Dictation\ wpm$$

$$\widehat{Salary} = 9.788 + 0.110 \times 120 + 0.053 \times 9 + 0.071 * 50 + 0.004 \times 60 + 0.065 \times 30 = 29.205$$

- **b)** $29,205
- **c)** The t-value is 0.013 with 24 df and a P-value = 0.9897 (two-tailed), which is not significant at $\alpha = 0.05$.
- **d)** You could take out the explanatory variable *typing speed* since it is not significant.
- **e)** *Age* is likely to be collinear with several of the other predictors already in the model. For example, secretaries with longer terms of *Service* will naturally also be older.

Copyright © 2017 Pearson Education, Inc.

32. Wal-Mart revenue.

a) The regression equation:

$$\widehat{Revenue} = 87.0089 + 0.0001 Retail\ Sales + 0.000011 Personal\ Consumption - 0.345 CPI$$

b) After allowing for the effects of the other predictors in the model, a change of 1 point in the CPI is associated with a decrease of 0.345 billion dollars on average in Wal-Mart revenue. Possibly higher prices (increased CPI) lead customers to shop less.

c) The Normal probability plot looks reasonably straight, so the Nearly Normal condition is met. With a P-value of 0.007, it is very unlikely that the true coefficient is zero.

33. Gross domestic product.

a) This model explains less than 4% of the variation in *GDP per Capita*. The P-value is not particularly low.

b) Because more education is general associated with a higher standard of living, it is not surprising that the simple association between *Primary Completion Rate* and GDP is positive.

c) The coefficient now is measuring the association between *GDP/Capita* and *Primary Completion Rate* after account for the two other predictors.

34. Lobster industry 2012, revisited.

a) $\widehat{LogValue} = 0.856 + 0.563 Traps - 0.000044 Fishers + 0.00381 Pounds / Trap$

b) Residuals show no pattern and have equal spread. Normal probability plot is straight. There is a question whether values from year to year are mutually independent.

c) After allowing for the number of *Traps* and *pounds/trap*, the *LogValue* of the lobster catch decreases by 0.000044 per *Fisher*. This doesn't mean that a smaller number of fishers would lead to a more valuable harvest. It is likely that *Traps* and *Fishers* are correlated, affecting the value and meaning of their coefficients.

d) The hypotheses are: $\mathrm{H}_0 : \beta_{lbs/trap} = 0$; $\mathrm{H}_A : \beta_{lbs/trap} \neq 0$; P-value = 0.0114 is below the common alpha level of 0.05, so we can reject the null hypothesis. However, this is not strong evidence that pounds/trap is an important predictor of the harvest value.

35. Lobster industry 2012, part 2.

a) $\widehat{EstimatedPrice / lb} = 1.094 + 1.236 Traps(M) - 0.000149 Fishers - 0.0180 Pounds\ /\ Trap$

b) Residuals show greater spread on the right and a possible outlier on the high end. We might wonder if the values from year to year are mutually independent. We should interpret the model with caution.

c) The hypotheses are: $\mathrm{H}_0 : \beta_{pounds/trap} = 0$; $\mathrm{H}_A : \beta_{pounds/trap} \neq 0$; P-value = 0.0011. It appears that *Pounds/Trap* does contribute to the model.

d) No, we can't draw causal conclusions from a regression. A change in *pounds/trap* would likely affect other variables in the model.

e) The adjusted R^2 accounts for the number of predictors and says to prefer the more complex model.

36. HDI.

a)
$$\widehat{HDI} = 0.09 + 0.01376 ExpectedYearsofSchooling + 0.00333 LifeExpectancy - 0.00012 MaternalMortality$$
$$+0.01686 MeanYrsSchool + 0.000745 PopUrban + 0.000000802 GDP / Capita + 0.00046 CellPhones / 100$$

b) No. There appear to be two outliers with HDI higher than predicted. They make the residuals non-Normal.

c) The hypotheses are: $\mathrm{H}_0 : \beta_{YrsSchool} = 0$; $\mathrm{H}_A : \beta_{YrsSchool} \neq 0$; *t*-statistic 8 is quite large, so we can reject the null hypothesis.

d) The Normality assumption is violated. Most likely the standard deviation of the residuals is inflated. That would tend to make the t-ratios smaller. This one is so large that we probably can feel safe in rejecting the null hypothesis anyway.

Copyright © 2017 Pearson Education, Inc.

37. Wal-Mart revenue, part 2.

a) $\widehat{PBE} = 87.0 - 0.345\,CPI + 0.000011\,Personal\ Consumption + 0.0001\,Retail\ Sales$

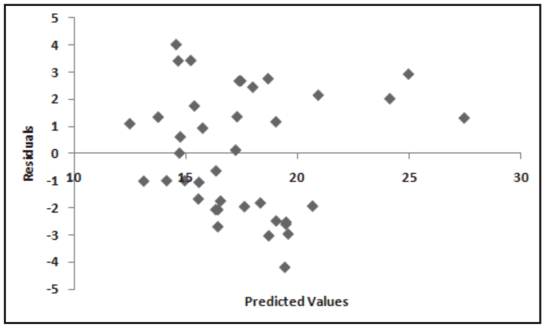
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	87.00892605	33.59897163	2.589631	0.013908
CPI	-0.344795233	0.120335014	-2.86529	0.007002
Personal Consumption	1.10842E-05	4.40271E-06	2.51759	0.016546
Retail Sales Index	0.000103152	1.54563E-05	6.67378	1.01E-07

Regression Statistics	
Multiple R	0.816425064
R Square	0.666549886
Adjusted R Square	0.637968448
Standard Error	2.326701861
Observations	39

b) $R^2 = 66.7\%$ and all *t*-ratios are significant. It looks like these variables can account for much of the variation in Wal-Mart revenue.

38. Wal-Mart revenue, part 3.

a) The plot does not show any pattern or spread. There are a few high values to the right that could be considered outliers.



b) The December results correspond to the high values. Performing the regression analysis again without the four December values:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-22.6410169	35.98903632	-0.62911	0.533887
CPI	0.04756868	0.129728836	0.366678	0.71635
Personal Consumption	7.8443E-07	4.27905E-06	0.183319	0.855741
Retail Sales Index	1.3382E-05	2.26191E-05	0.591614	0.558399

Regression Statistics	
Multiple R	0.64980747
R Square	0.42224975
Adjusted R Square	0.36633844
Standard Error	1.87418242
Observations	35

Copyright © 2017 Pearson Education, Inc.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	79.58196867	26.52732	7.552134	0.000623298
Residual	31	108.8893521	3.51256		
Total	34	188.4713207			

c) Without the December values, none of these variables is obviously different from zero. None of the *t*-ratios or P-values are significant. The F-statistic of 7.55 is highly significant with a P-value of < 0.001 indicating that the slope coefficients are not zero. The regression analysis as a whole doesn't provide much insight into the Wal-mart revenues. It appears that the model of the previous exercise was only about holiday sales.

39. Clinical trials.

a) $\widehat{Logit(Drop)} = -0.4419 - 0.0379\,Age + 0.0468HDRS$

b) To find the predicted log odds (logit) of the probability that a 30-year-old patient with an HDRS score of 30 will drop out of the study: set *Age* = 30 and *HDRS* = 30 in the estimated regression equation: – 0.4419 – 0.0379 (30) – 0.0468(30) = –0.1749.

c) The predicted dropout probability of that patient is

$$\hat{p} = \frac{1}{1+e^{-(-0.1749)}} = \frac{1}{1+1.19} = 0.4564$$

d) To find the predicted log odds (logit) of the probability that a 60-year-old patient with an HDRS score of 8 will drop out of the study: set *Age* = 60 and *HDRS* = 8 in the estimated regression equation: – 0.4419 – 0.0379 (60) + 0.0468(8) = –2.342.

e) The associated predicted probability is

$$\hat{p} = \frac{1}{1+e^{-(-2.342)}} = \frac{1}{1+10.42} = 0.0877$$

40. Cost of higher education.

a) $\widehat{Logit(Type)} = -13.1461 + 0.08455Top10\% + 0.000259\,\$\,/\,Student$

The Outlier Condition is satisfied because there are not outliers in either predictor, but this is not a sample, but the top 25 colleges and universities in the U.S. It may be used for predicting, but the inference is not clear.

b) Yes; the P-value is < 0.05.

c) Yes; the P-value is < 0.05.

41. Motorcycles.

The scatterplot of *MSRP* versus *Wheelbase* indicates that this relationship is not linear. While there is a positive relationship between *Wheelbase* and *MSRP*, a curved pattern is evident. The scatterplot of *MSRP* versus *Displacement* indicates that this relationship is positive and linear. The scatterplot of *MSRP* versus *Bore* indicates that this relationship is also positive and linear. Based on these plots, it appears that both *Displacement* and *Bore* would be better predictors of *MSRP* than *Wheelbase.*

42. Motorcycles, part 2.

a) The hypotheses are: $\text{H}_0 : \beta_{Bore} = 0$; $\text{H}_A : \beta_{Bore} \neq 0$; P-value = 0.0108 which is greater than 0.05, so we fail to reject the null hypothesis.

b) Although *Bore* might be individually significant in predicting *MSRP*, in the multiple regression, after allowing for the effects of *Displacement*, it doesn't add enough to the model to have a coefficient that is clearly different from zero.

Copyright © 2017 Pearson Education, Inc.

43. Motorcycles, part 3.

a) Yes, with an R^2 = 90.9% says that most of the variability of *MSRP* is accounted for by this model.

b) No, in a regression model, you can't predict an explanatory variable from the response variable.

44. Demographics.

a) The only model that seems to do poorly is the one that omits *murder.* The other three are hard to choose among.

$$\widehat{Life}\exp = 70.1421 - 0.238597(Murder) + 0.039059(HSgrad) + 0.000095(Income)$$

$$R^2 = 66.4\%$$

$$\widehat{Life}\exp = 69.7354 - 0.258132(Murder) + 0.051791(HSgrad) + 0.253982(Illiteracy)$$

$$R^2 = 66.8\%$$

$$\widehat{Life}\exp = 71.1638 - 0.273632(Murder) + 0.000381(Income) + 0.036869(Illiteracy)$$

$$R^2 = 63.7\%$$

b) Each of the models has at least one coefficient with a large P-value. This predictor variable could be omitted to simplify the model without degrading it too much.

c) No. Regression models cannot be interpreted that way. Association is not the same thing as causation.

d) Plots of the residual highlight Hawaii, Alaska, and Utah as possible outliers. These seem to the principal violations of the assumptions and conditions.

45. Burger King nutrition.

a) With an R^2 = 100%, the model should make excellent predictions.

b) The value of *s*, 3.140 calories, is very small compared to the initial standard variation of *calories*. This means that the model fits the data quite well, leaving very little variation unaccounted for.

c) No, the residuals are not all 0. Indeed, we know that their standard deviation is *s* = 3.140 calories. They are very small compared with the original values. The true value of R^2 was likely rounded up to 100%.

46. Health expenditures.

a) $\widehat{Expenditures} = 0.1994 + 0.232 ExpectedYrsofSchooling + 0.051 InternetUsers / 100 people$

b) Residuals show no pattern and have equal spread. There is one fairly high residual and one fairly low residual, but neither seem to be large outliers. Normal probability plot is fairly straight. Assuming that the errors are independent, the conditions are met.

c) The hypotheses are: $\text{H}_0 : \beta_{Years} = 0$; $\text{H}_\text{A} : \beta_{Years} \neq 0$; the *t*-value is 2.81 (93df) and a P-value = 0.0006; we reject the null hypothesis of 0 slope.

d) No. The model says nothing about causality. It says that accounting for *Expected Year of Schooling*, higher numbers of Internet users are generally associated with higher health expenditures.

Copyright © 2017 Pearson Education, Inc.

Ethics in Action

Kenneth’s Ethical Dilemma: With all 5 independent variables included, gender shows no significant effect on sales performance and Kenneth wants to eliminate it from the model. Nicole reminds him that women had a history of being offered lower starting base salaries and when that is removed, gender is significant. When all variables are together, the detailed effects are confounded and gender is not an issue in sales performance.

Undesirable Consequences: Eliminating gender as a predictor of sales performance may mask the effects of gender and actually give an incorrect conclusion.

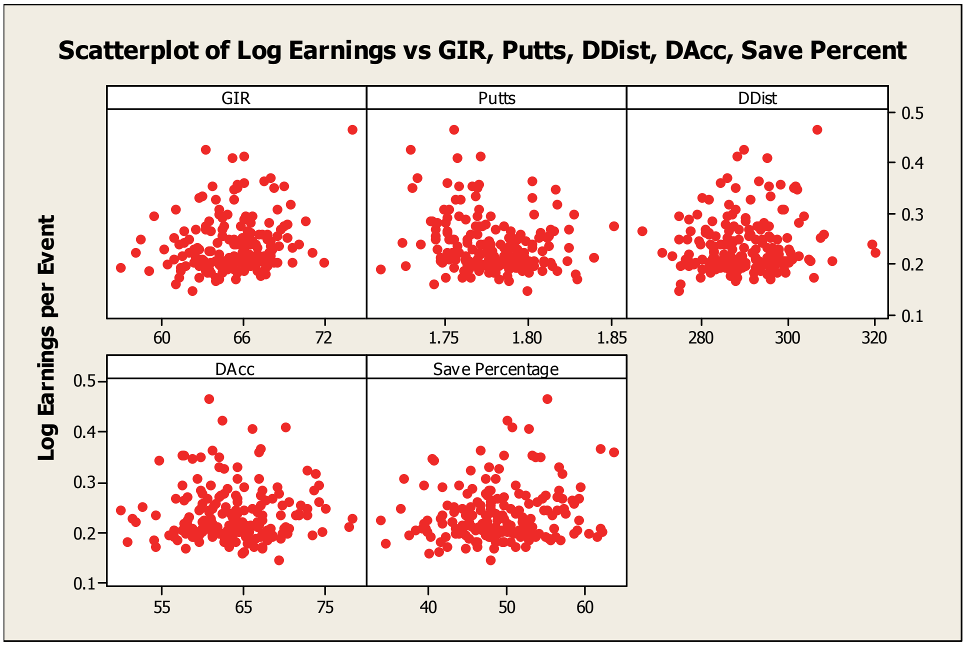
Ethical Solution: Kenneth needs to listen to Nicole’s logic and her recollection of history regarding women and lower starting base salaries. Because the starting base salaries were inequitable, it is confounded with gender and starting base salaries and should be eliminated from the model until the data reflect the court order adjustment.

For further information on the official American Statistical Association’s Ethical Guidelines, visit:
http://www.amstat.org/about/ethicalguidelines.cfm
The Ethical Guidelines address important ethical considerations regarding professionalism and responsibilities.

Brief Case – Golf Success

Report:

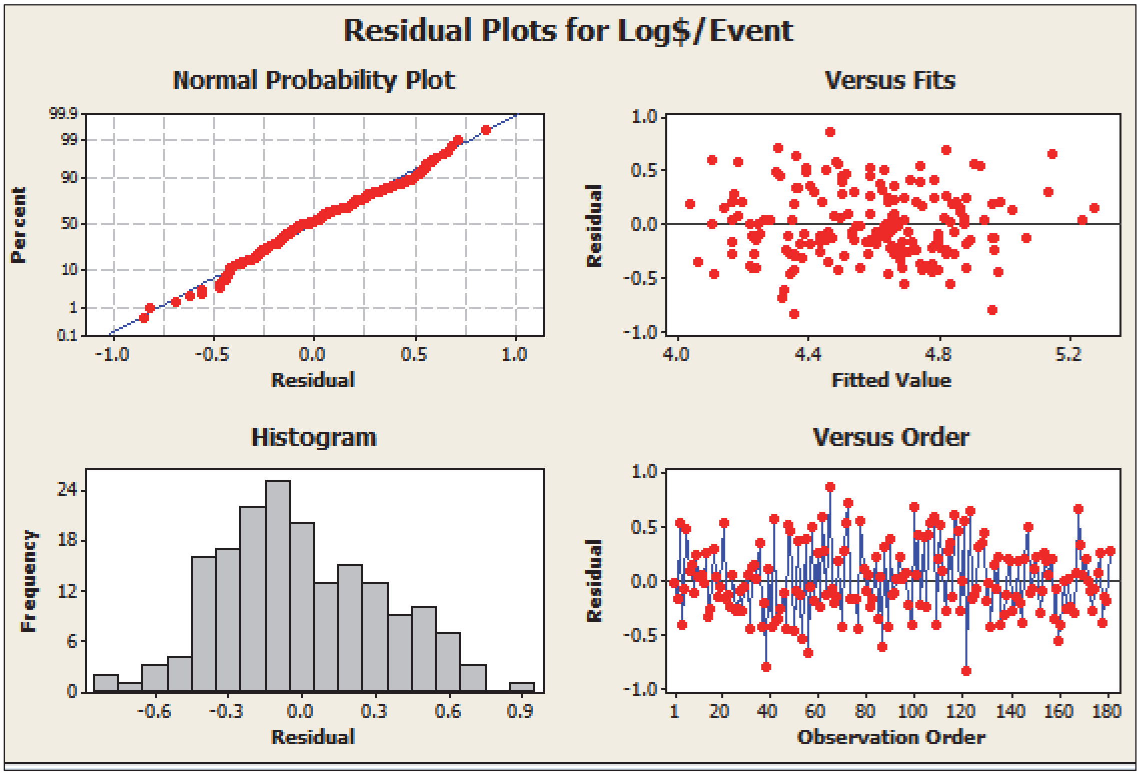
Of the potential variables considered for predicting golfers’ success (measured in log earnings per event), the best model includes two significant independent variables: *GIR* and *Putts*. *GIR* stands for “Greens in Regulation” and is defined as the percentage of holes played in which the ball is on the green with two or more strokes left for par. The variable *Putts* is the average number of putts per hole in which the green was reached in regulation. In the scatterplots of *Log Earnings* versus all potential independent variables (shown below), weak linear relationships are observed (*Log Earnings* has a weak positive linear association with *GIR* and a weak negative linear association with *Putts*). The model is $\widehat{LogEarningsperEvent} = 0.828 + 0.00497\,GIR - 0.515\,Putts$. The model is significant with a moderate explanatory power ($R^2 = 37.2\%$). This model explains less than 40% of the variability in golfers’ success. Examination of the residuals plotted against fitted values indicates that the equal spread condition is reasonably satisfied, but the histogram of residuals is skewed right indicating problems with the nearly normal condition even though a log transformation of earnings is used.

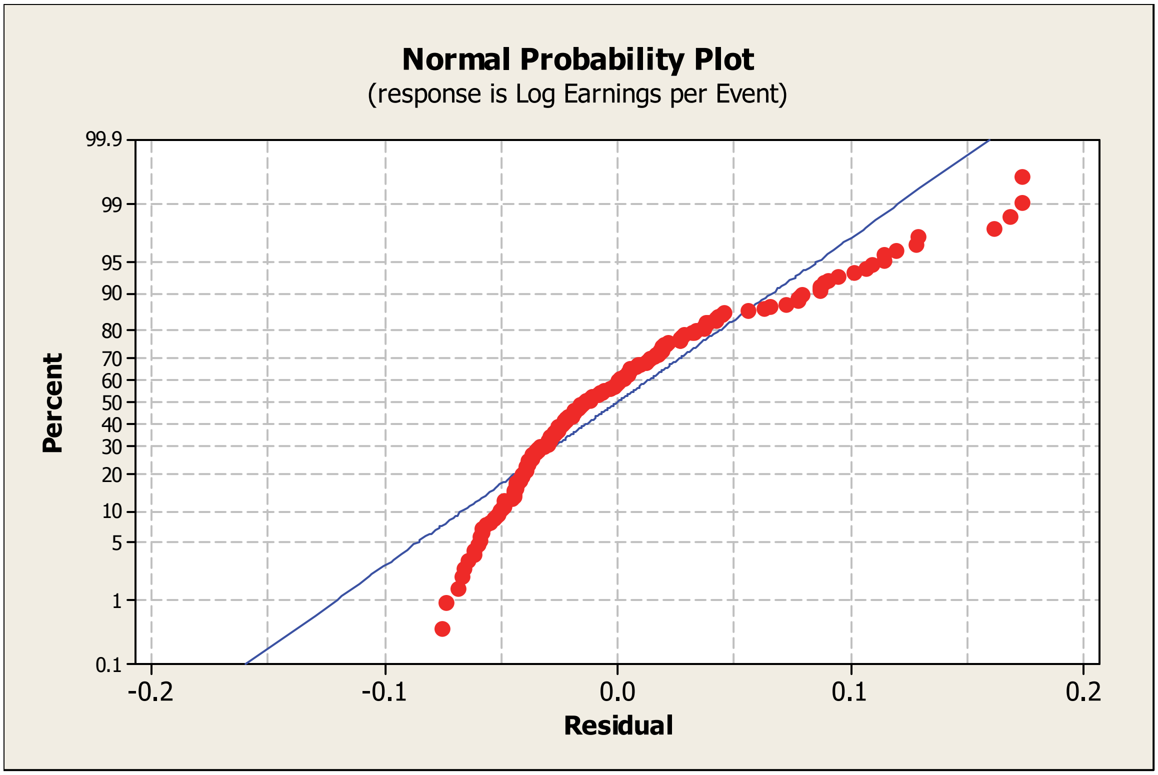


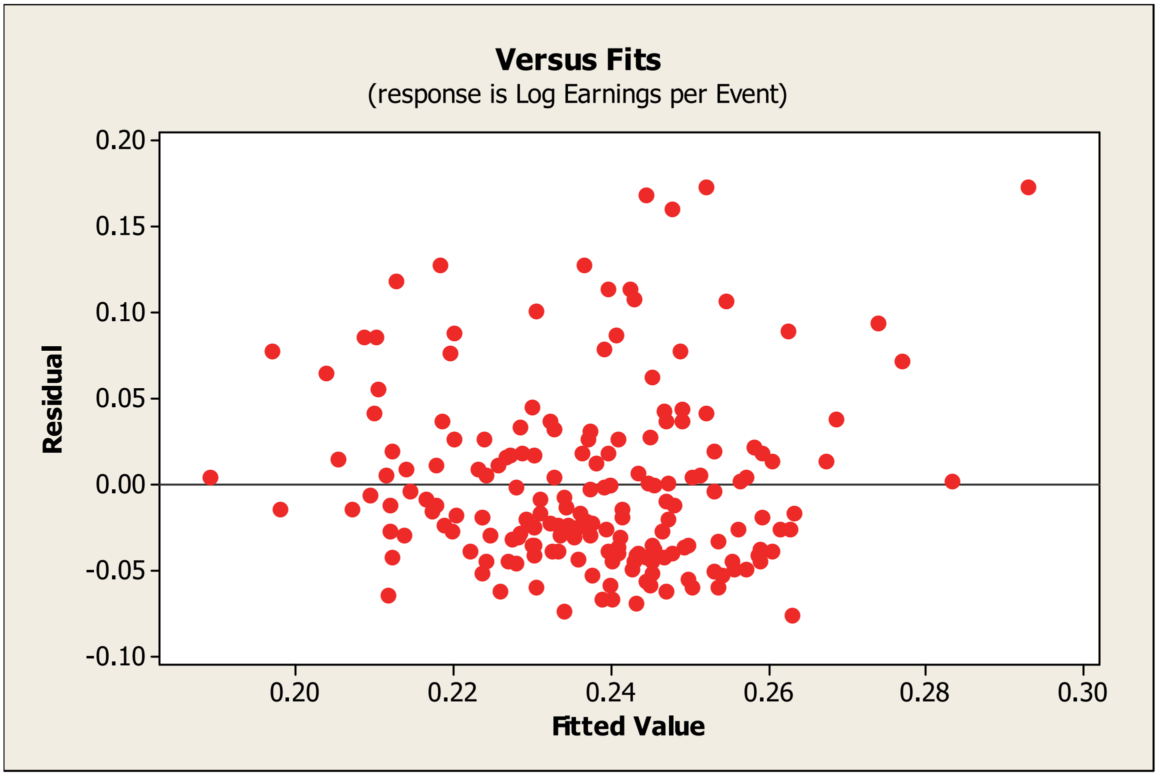
Copyright © 2017 Pearson Education, Inc.

General Regression Analysis: Log$/Event versus GREENS IN REG., PUTT AVG.

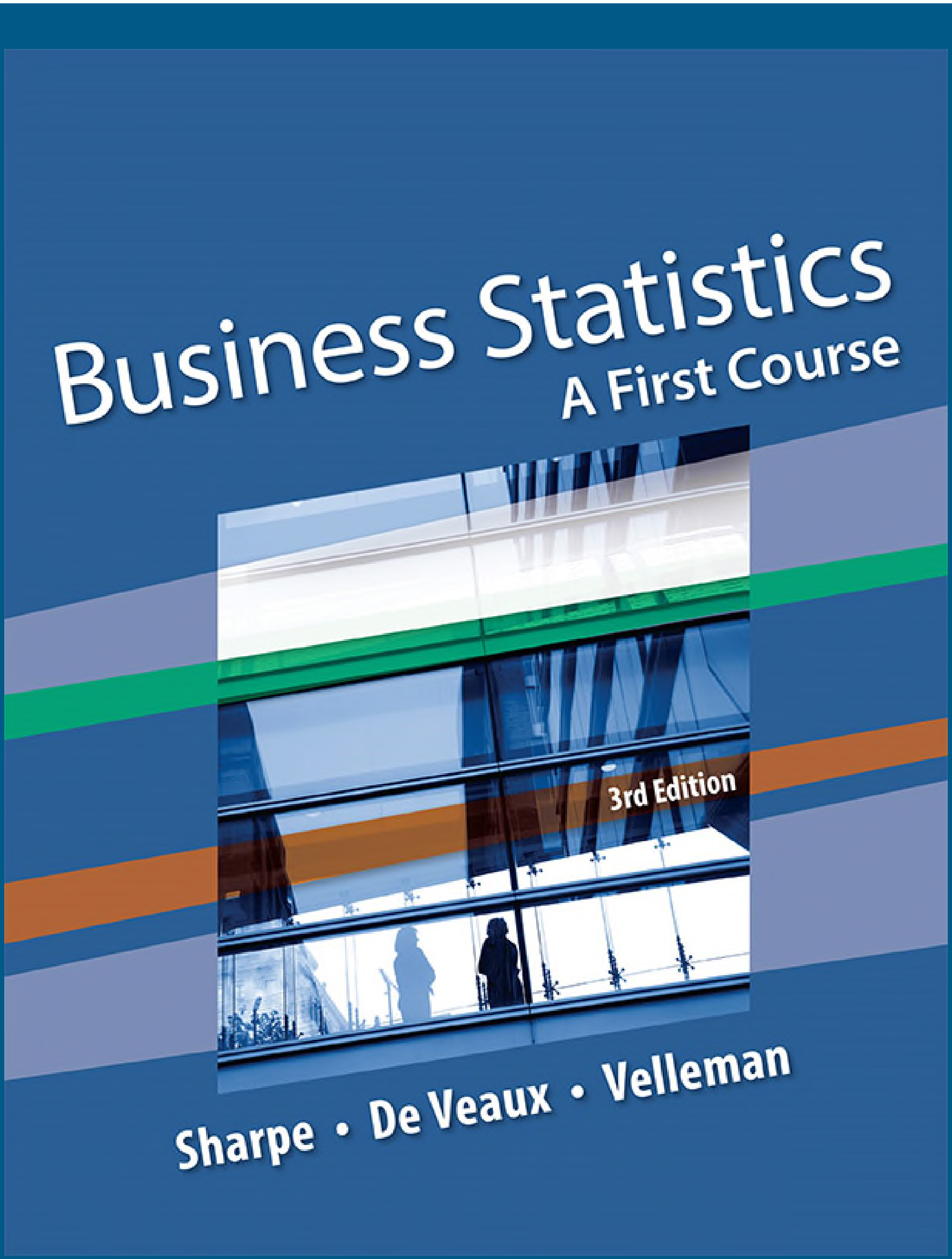
```
Regression Equation
Log$/Event  =  11.6983 + 0.0641226 GREENS IN REG. - 6.31982 PUTT AVG.

Coefficients

Term                 Coef  SE Coef         T      P
Constant          11.6983  1.37020   8.53767  0.000
GREENS IN REG.     0.0641  0.00862   7.44206  0.000
PUTT AVG.         -6.3198  0.76064  -8.30860  0.000

Summary of Model

S = 0.329649     R-Sq = 37.20%    R-Sq(adj) = 36.50%
PRESS = 20.0679  R-Sq(pred) = 34.85%
```







Copyright © 2017 Pearson Education, Inc.



Chapter 2

Displaying and Describing Categorical Data

Copyright © 2017, 2015 Pearson Education. All rights reserved.

PEARSON

2.1 Summarizing a Categorical Variable

A *frequency table* organizes data by recording totals and category names as in the table below.

The names of the categories label each row in the frequency table.

Some tables report counts, others report percentages, and many report both.

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.1 Summarizing a Categorical Variable

Example: Super Bowl

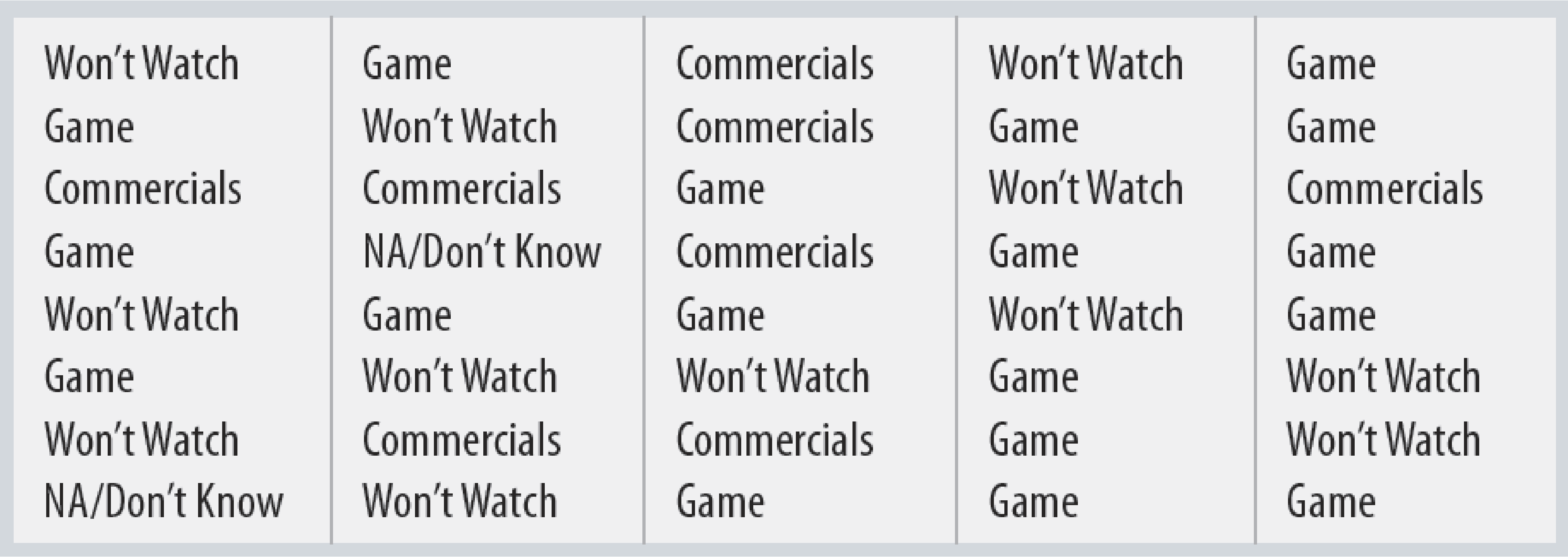
The Super Bowl, the championship game of the National Football League of the United States, is an important annual social event for Americans, with tens of millions of viewers. The ads that air during the game are expensive: a 30-second ad during the 2013 Super Bowl cost about $4M.

Polls often ask whether respondents are more interested in the game or the commercials. Here are 40 responses from one such poll. (next slide)

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.1 Summarizing a Categorical Variable

Example: Super Bowl



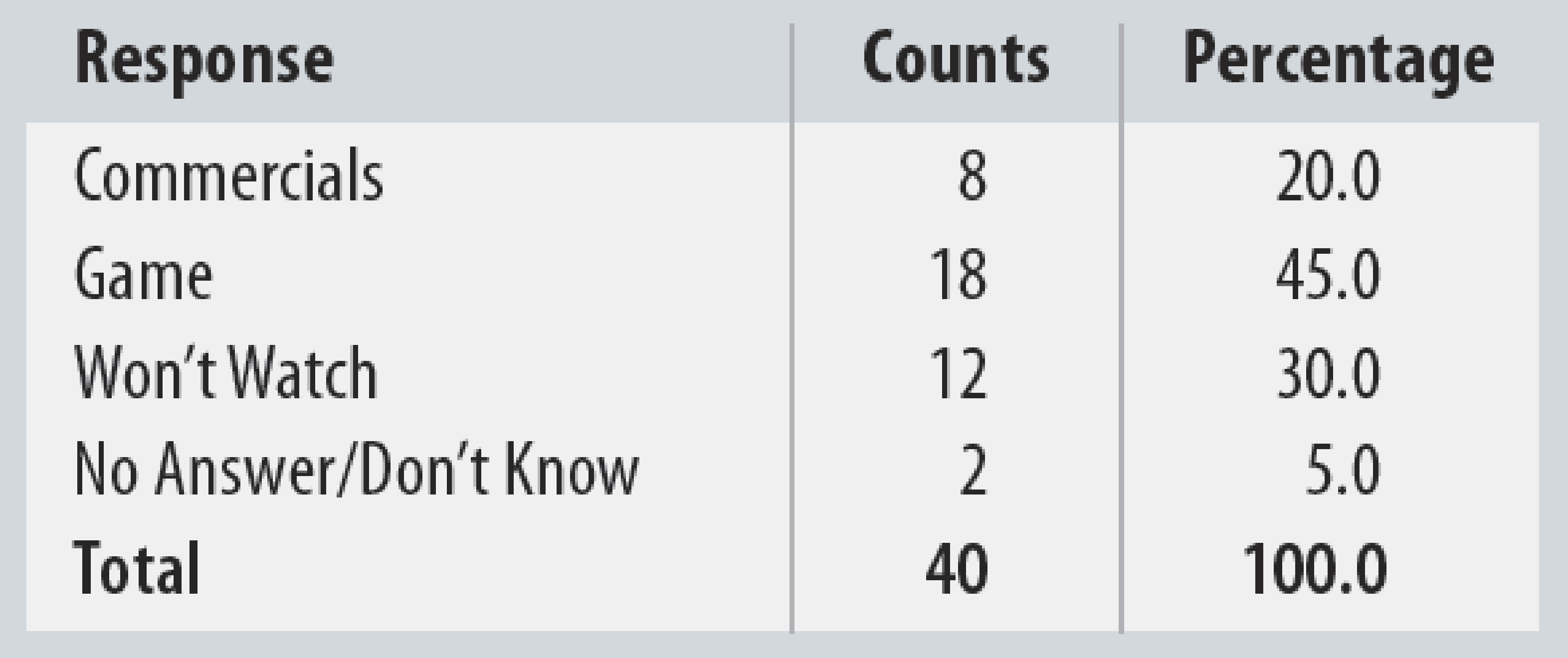
Won't Watch	Game	Commercials	Won't Watch	Game
Game	Won't Watch	Commercials	Game	Game
Commercials	Commercials	Game	Won't Watch	Commercials
Game	NA/Don't Know	Commercials	Game	Game
Won't Watch	Game	Game	Won't Watch	Game
Game	Won't Watch	Won't Watch	Game	Won't Watch
Won't Watch	Commercials	Commercials	Game	Won't Watch
NA/Don't Know	Won't Watch	Game	Game	Game

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.1 Summarizing a Categorical Variable

Example: Super Bowl

Make a frequency table for this variable. Include counts and percentages.



Response	Counts	Percentage
Commercials	8	20.0
Game	18	45.0
Won't Watch	12	30.0
No Answer/Don't Know	2	5.0
Total	**40**	**100.0**

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.2 Displaying a Categorical Variable

The Three Rules of Data Analysis

Make a picture. Make a picture. Make a picture. Pictures …

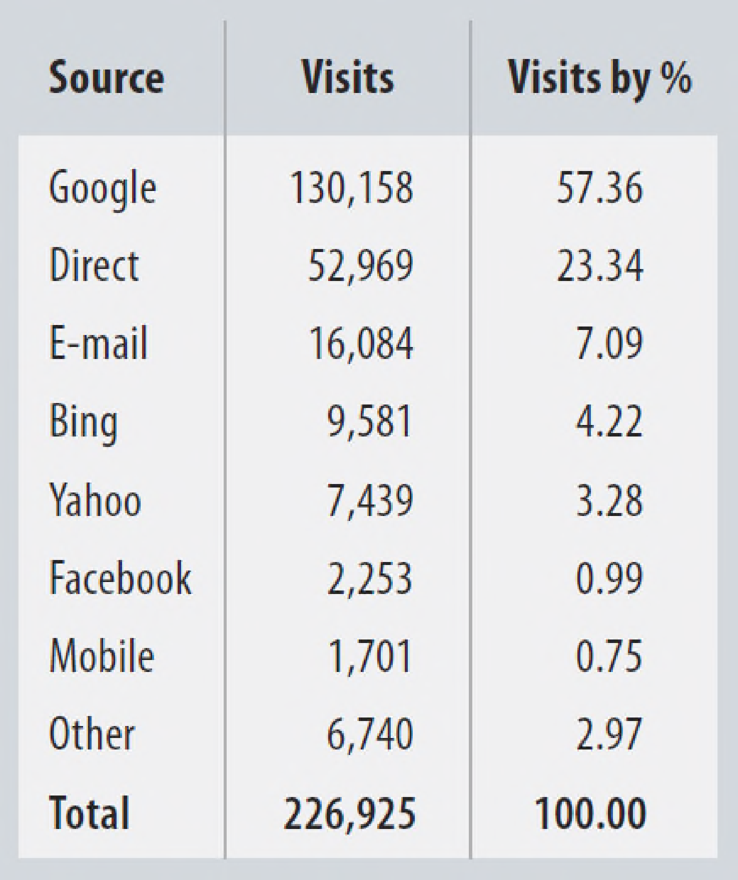
- reveal things that can’t be seen in a table of numbers.
- show important features and patterns in the data.
- provide an excellent means for reporting findings to others.

Copyright © 2017, 2015 Pearson Education. All rights reserved.

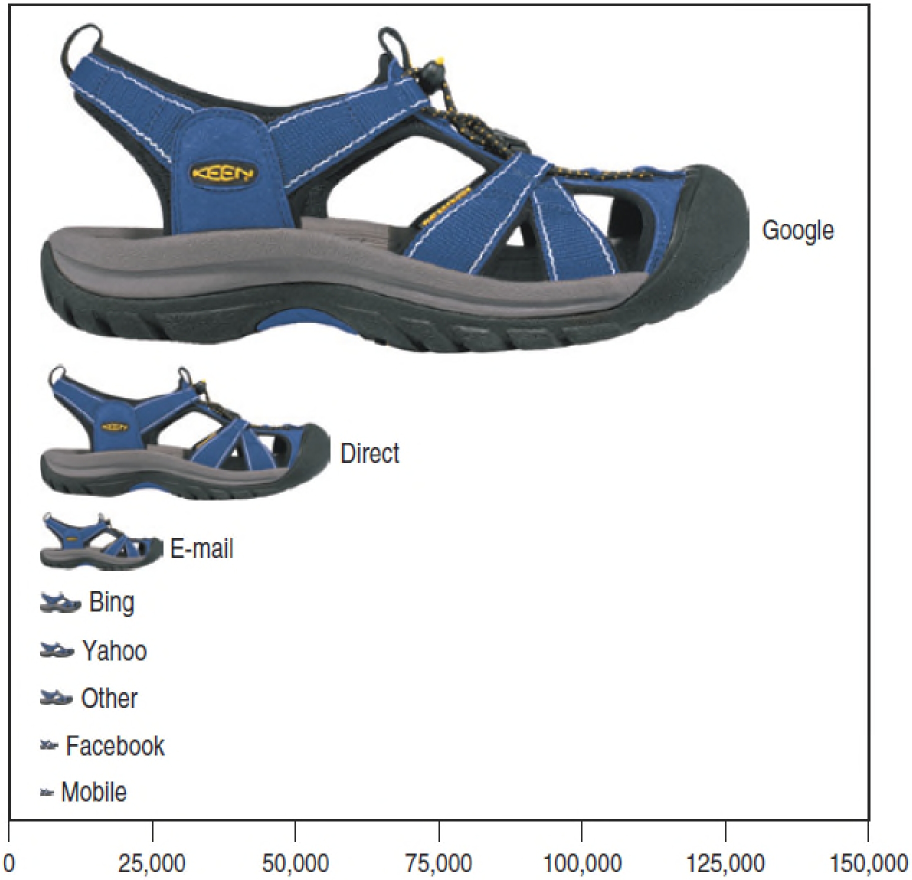
2.2 Displaying a Categorical Variable

The Area Principle

The figure given distorts the data from the frequency table.



Source	Visits	Visits by %
Google	130,158	57.36
Direct	52,969	23.34
E-mail	16,084	7.09
Bing	9,581	4.22
Yahoo	7,439	3.28
Facebook	2,253	0.99
Mobile	1,701	0.75
Other	6,740	2.97
Total	**226,925**	**100.00**



Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.2 Displaying a Categorical Variable

The Area Principle

The best data displays observe the *area principle*: the area occupied by a part of the graph should correspond to the magnitude of the value it represents.

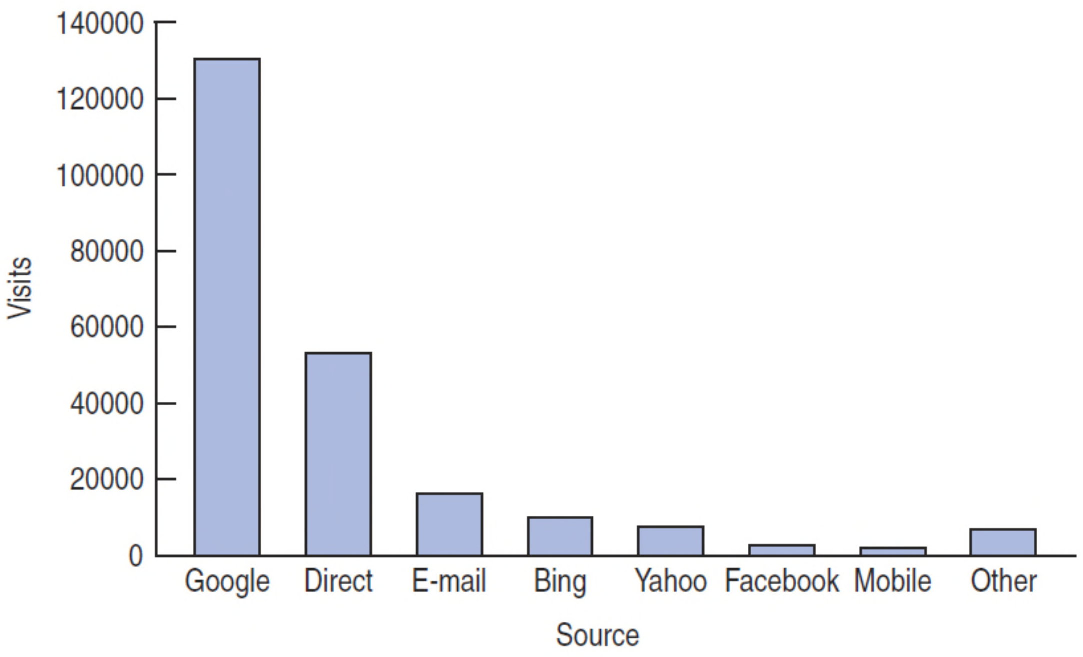
Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.2 Displaying a Categorical Variable

Bar Charts

A *bar chart* displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison.

The bar graph here gives a more *accurate* visual impression of the sandal data, though it may not be as visually entertaining.



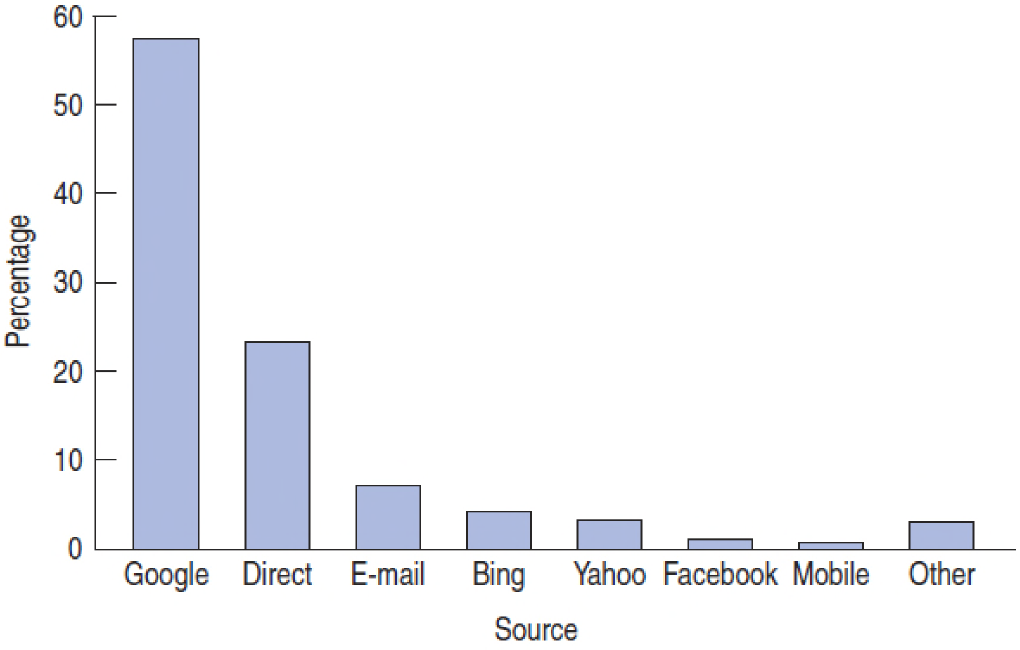
Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.2 Displaying a Categorical Variable

Bar Charts

If the counts are replaced with percentages, the data can be displayed in a *relative frequency bar chart*.

The relative frequency bar chart looks the same as the bar chart, but shows the proportion of visits in each category rather than counts.

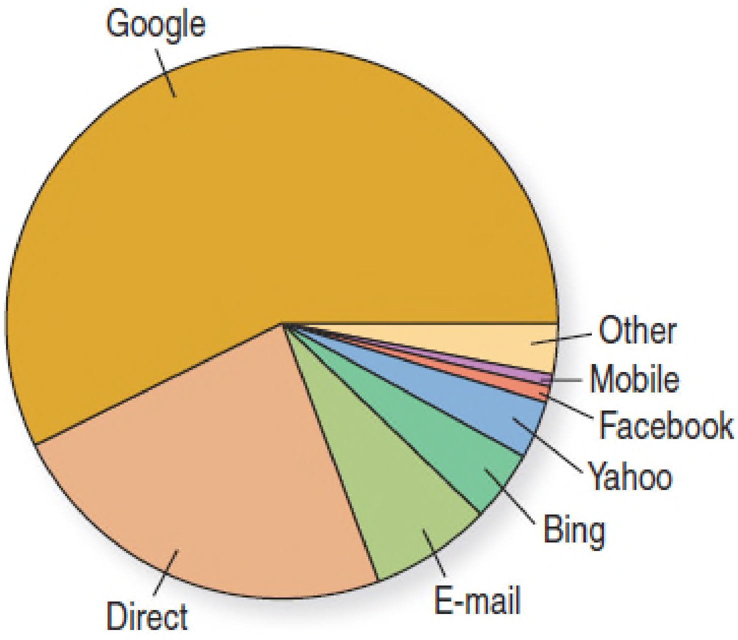


Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.2 Displaying a Categorical Variable

Pie Charts

Pie charts show the whole group of cases as a circle sliced into pieces with sizes proportional to the fraction of the whole in each category. The KEEN Inc. data is displayed below.



Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.2 Displaying a Categorical Variable

Before making a bar chart or pie chart,

- the data must satisfy the *Categorical Data Condition*: the data are counts or percentages of individuals in categories.
- be sure the categories don’t overlap.
- consider what you are attempting to communicate about the data.

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.3 Exploring Two Categorical Variables: Contingency Tables

- To show how two categorical variables are related, we can create a contingency table.

- Contingency tables show how individuals are distributed along each variable depending on the value of the other variable.

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.3 Exploring Two Categorical Variables: Contingency Tables

- The *marginal distribution* of a variable in a contingency table is the total count that occurs when the value of that variable is held constant.

- Each *cell* of a contingency table (any intersection of a row and column of the table) gives the count for a combination of values of the two variables.

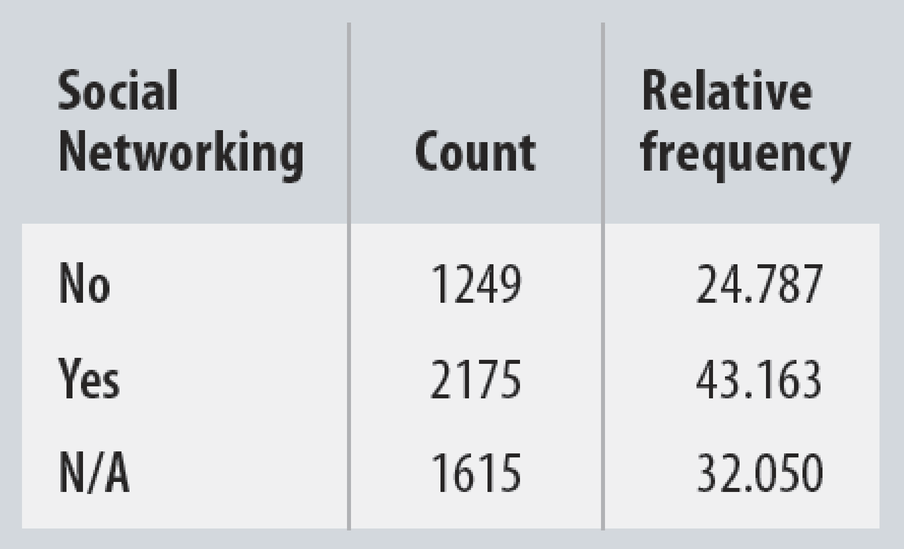
- Rather than displaying the data as counts, a table may display the data as a percentage – as a *total percent*, *row percent*, or *column percent*, which show percentages with respect to the total count, row count, or column count, respectively.

Copyright © 2017, 2015 Pearson Education. All rights reserved.

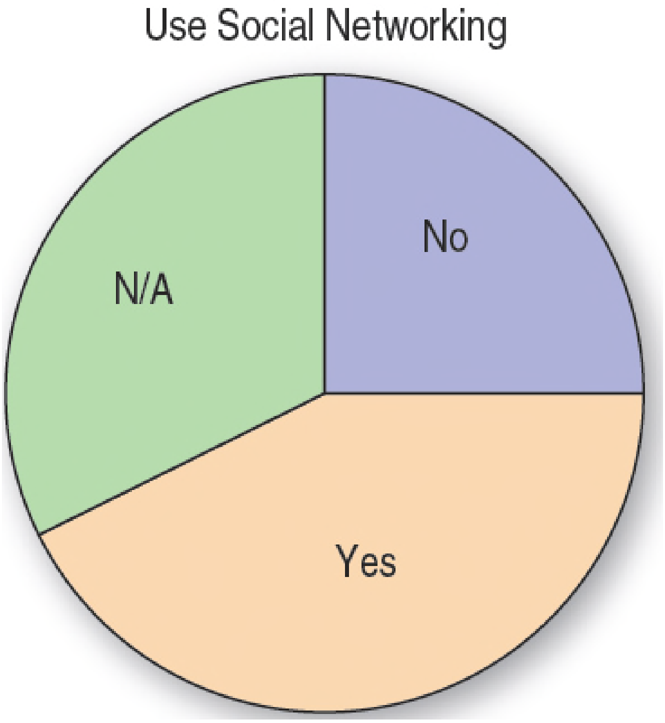
2.3 Exploring Two Categorical Variables: Contingency Tables

Example: Pew Research

One question of interest to business decision makers is how common it is for citizens of different countries to use social networking and whether they have it available to them.



Social Networking	Count	Relative frequency
No	1249	24.787
Yes	2175	43.163
N/A	1615	32.050

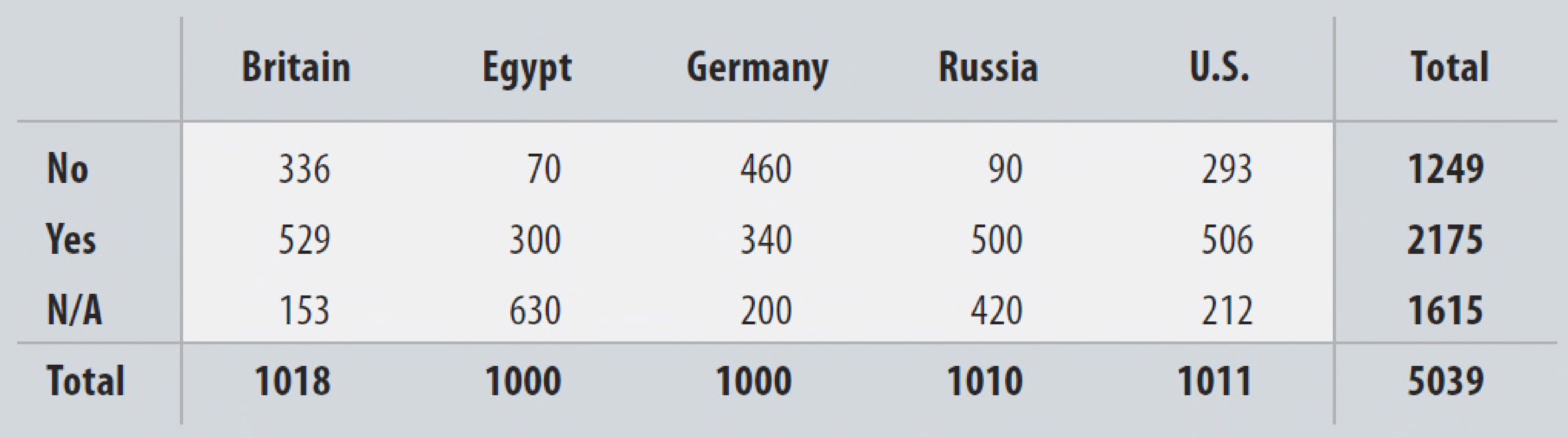


Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.3 Exploring Two Categorical Variables: Contingency Tables

Example: Pew Research

But if we want to target our online customer relations with social networks differently in different countries, wouldn’t it be more interesting to know how social networking use varies from country to country?



	Britain	Egypt	Germany	Russia	U.S.	Total
No	336	70	460	90	293	**1249**
Yes	529	300	340	500	506	**2175**
N/A	153	630	200	420	212	**1615**
Total	**1018**	**1000**	**1000**	**1010**	**1011**	**5039**

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.3 Exploring Two Categorical Variables: Contingency Tables

Conditional Distributions

Variables may be restricted to show the distribution for just those cases that satisfy a specified condition. This is called a *conditional distribution*.

The more interesting questions are contingent on something.

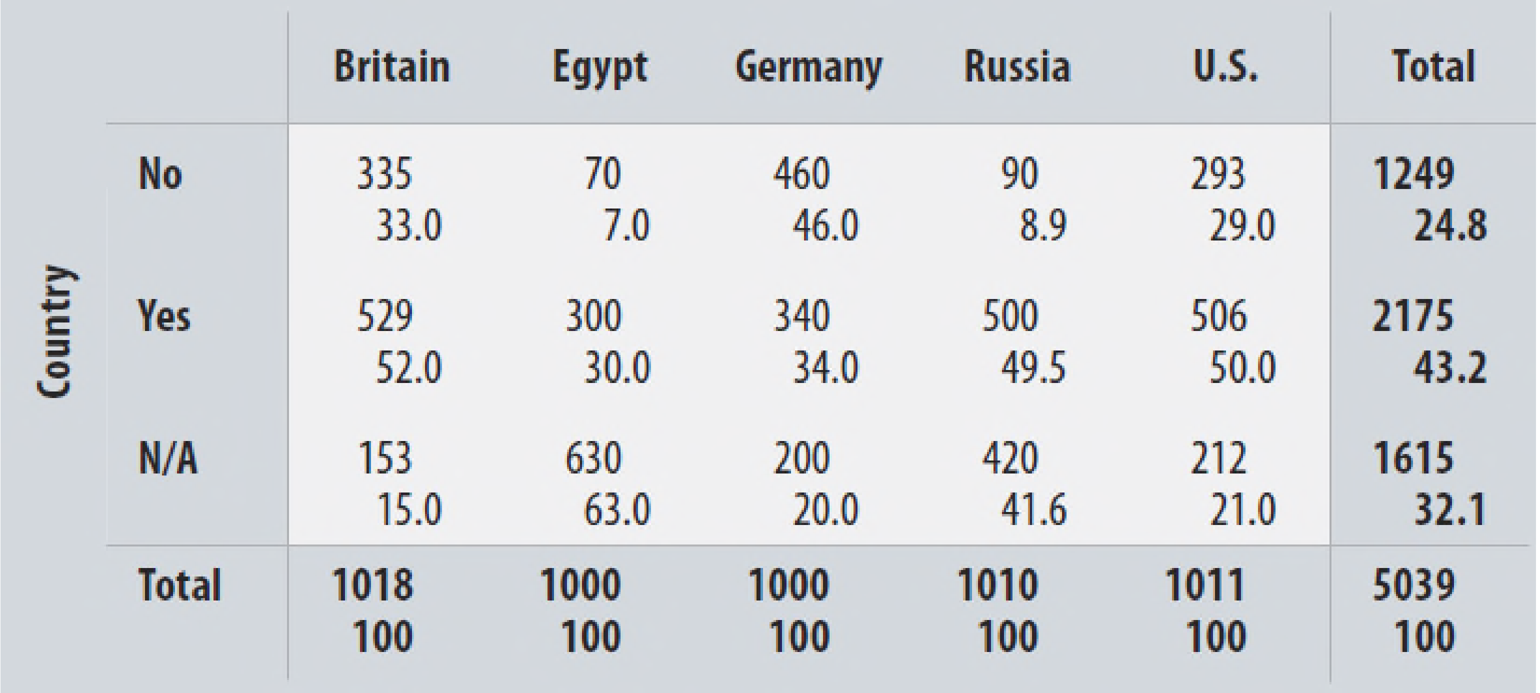
We'd like to know, for example, whether these countries are similar in use and availability of social networking.

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.3 Exploring Two Categorical Variables: Contingency Tables

Conditional Distributions

The conditional distribution of *Social Networking* conditioned on two values of *Country*. This table shows column percentages.



		Britain	Egypt	Germany	Russia	U.S.	Total
Country	No	335 33.0	70 7.0	460 46.0	90 8.9	293 29.0	**1249** **24.8**
	Yes	529 52.0	300 30.0	340 34.0	500 49.5	506 50.0	**2175** **43.2**
	N/A	153 15.0	630 63.0	200 20.0	420 41.6	212 21.0	**1615** **32.1**
	Total	**1018** **100**	**1000** **100**	**1000** **100**	**1010** **100**	**1011** **100**	**5039** **100**

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.3 Exploring Two Categorical Variables: Contingency Tables

Conditional Distributions

Variables can be related in many ways, so it is typically easier to ask if they are *not* related.

In a contingency table, when the distribution of one variable is the same for all categories of another variable, we say that the variables are *independent*.

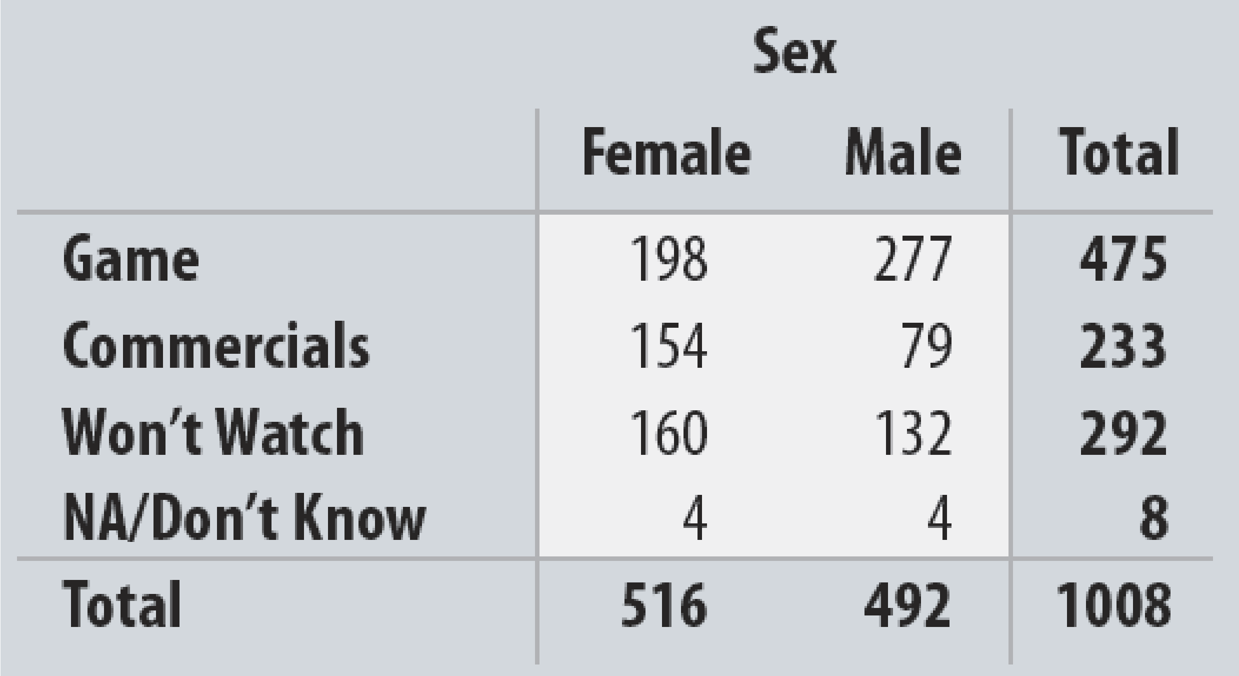
This tells us there is no association between these variables.

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.3 Exploring Two Categorical Variables: Contingency Tables

Example: Super Bowl

Here is a contingency table of the responses for 1008 adult U.S. respondents to the question about watching the Super Bowl discussed previously:



	Sex		
	Female	Male	Total
Game	198	277	**475**
Commercials	154	79	**233**
Won't Watch	160	132	**292**
NA/Don't Know	4	4	**8**
Total	**516**	**492**	**1008**

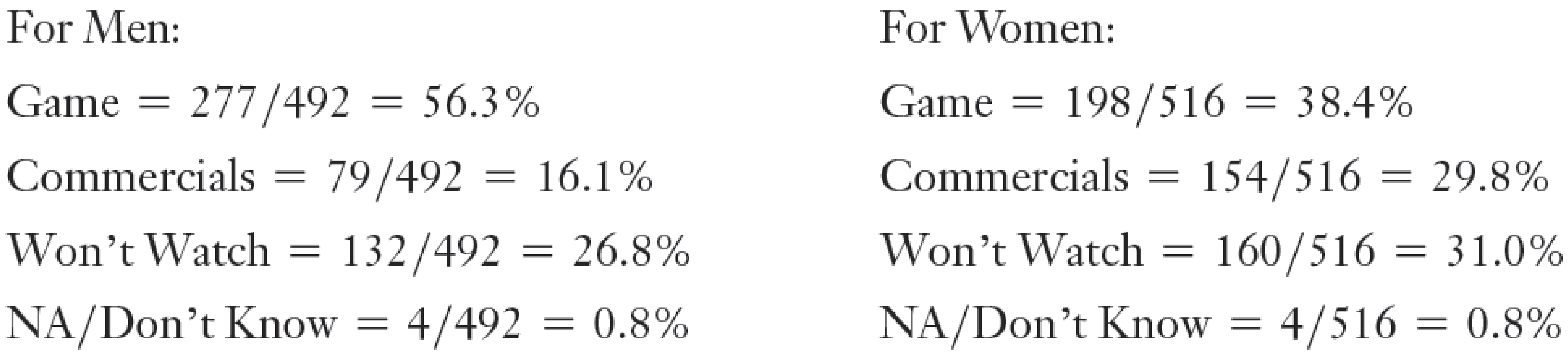
Does it seem that there is an association between what viewers are interested in watching and their sex?

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.3 Exploring Two Categorical Variables: Contingency Tables

Example: Super Bowl

First, find the conditional distributions of the four responses for each sex:



For Men:

Game = 277/492 = 56.3%

Commercials = 79/492 = 16.1%

Won't Watch = 132/492 = 26.8%

NA/Don't Know = 4/492 = 0.8%

For Women:

Game = 198/516 = 38.4%

Commercials = 154/516 = 29.8%

Won't Watch = 160/516 = 31.0%

NA/Don't Know = 4/516 = 0.8%

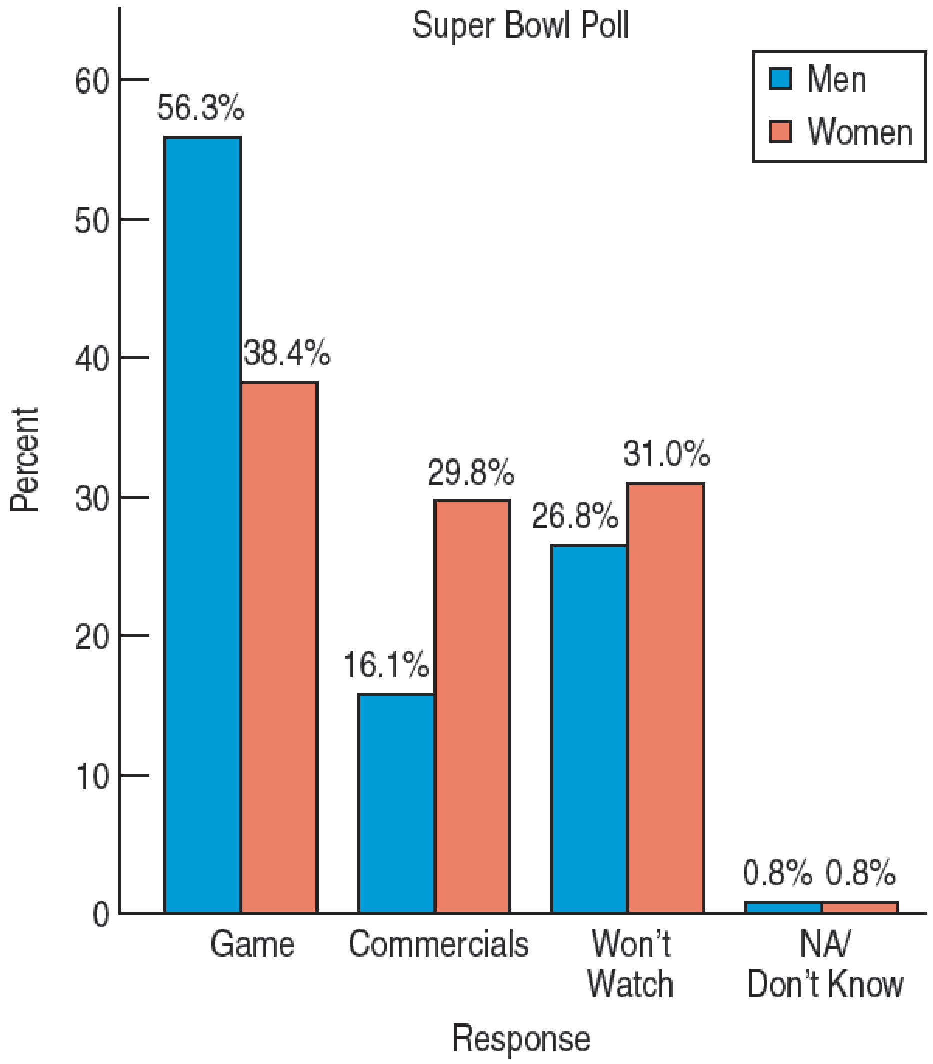
Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.3 Exploring Two Categorical Variables: Contingency Tables

Example: Super Bowl

Next, display the two distributions with side-by-side bar charts:

Based on this poll, there appears to be an association between the viewer's sex and what the viewer is most looking forward to.



Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.4 Segmented Bar Charts and Mosaic Plots

To further visualize conditional distributions, we can create segmented bar charts and mosaic plots.

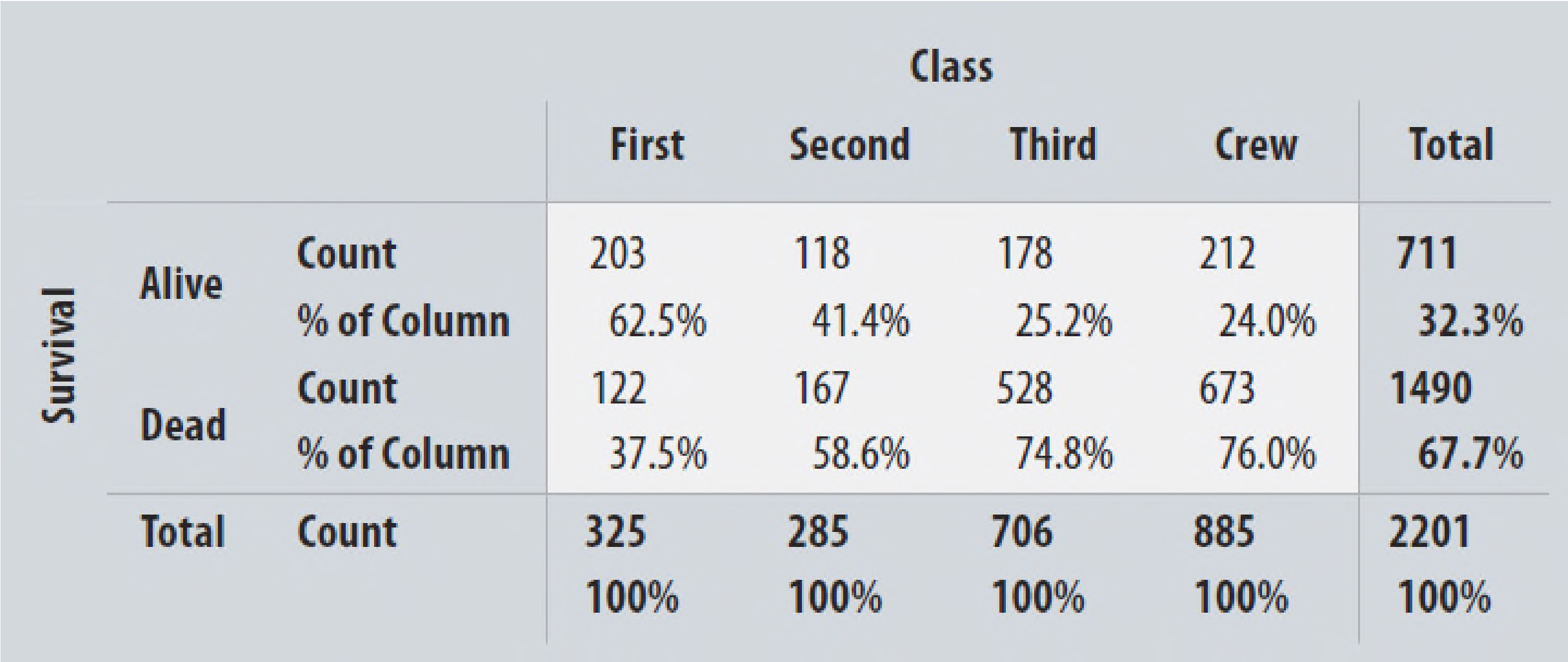
A segmented bar chart treats each bar as the “whole” and divides it proportionally into segments corresponding to the percentage in each group.

A variant of the segmented bar chart, a mosaic plot, looks like a segmented bar chart, but obeys the area principle better by making the bars proportional to the sizes of the groups.

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.4 Segmented Bar Charts and Mosaic Plots

Everyone knows what happened in the North Atlantic on the night of April 14, 1912 as the *Titanic,* thought by many to be unsinkable, sank, leaving almost 1500 passengers and crew members on board to meet their icy fate. Here is a contingency table of the 2201 people on board:

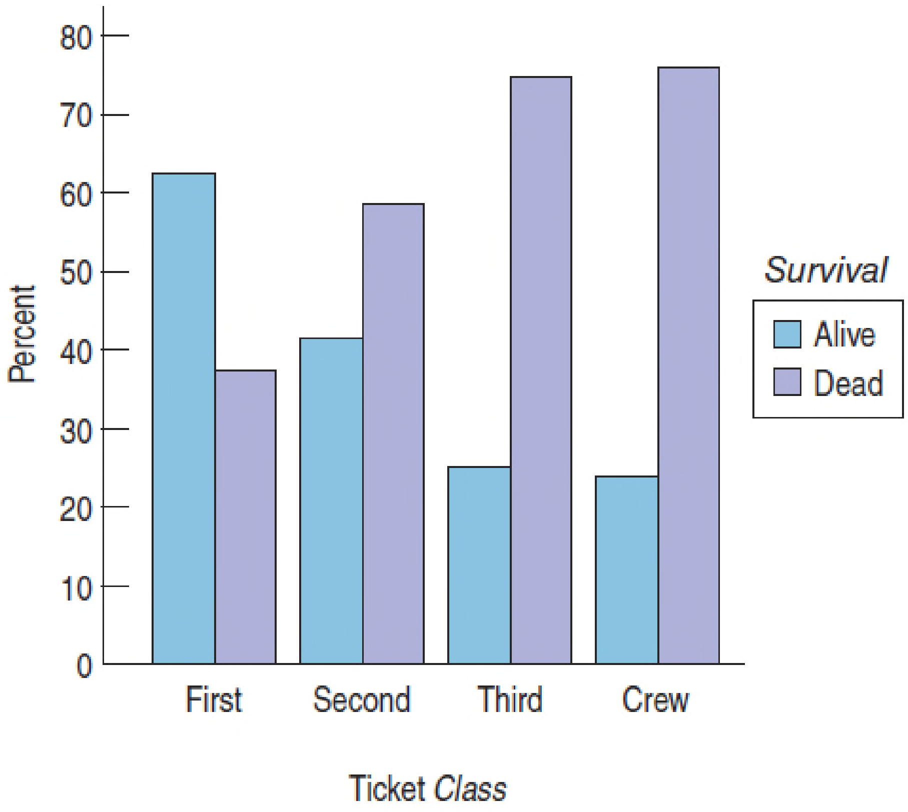


			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	711
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
	Total	Count	325	285	706	885	2201
			100%	100%	100%	100%	100%

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.4 Segmented Bar Charts and Mosaic Plots

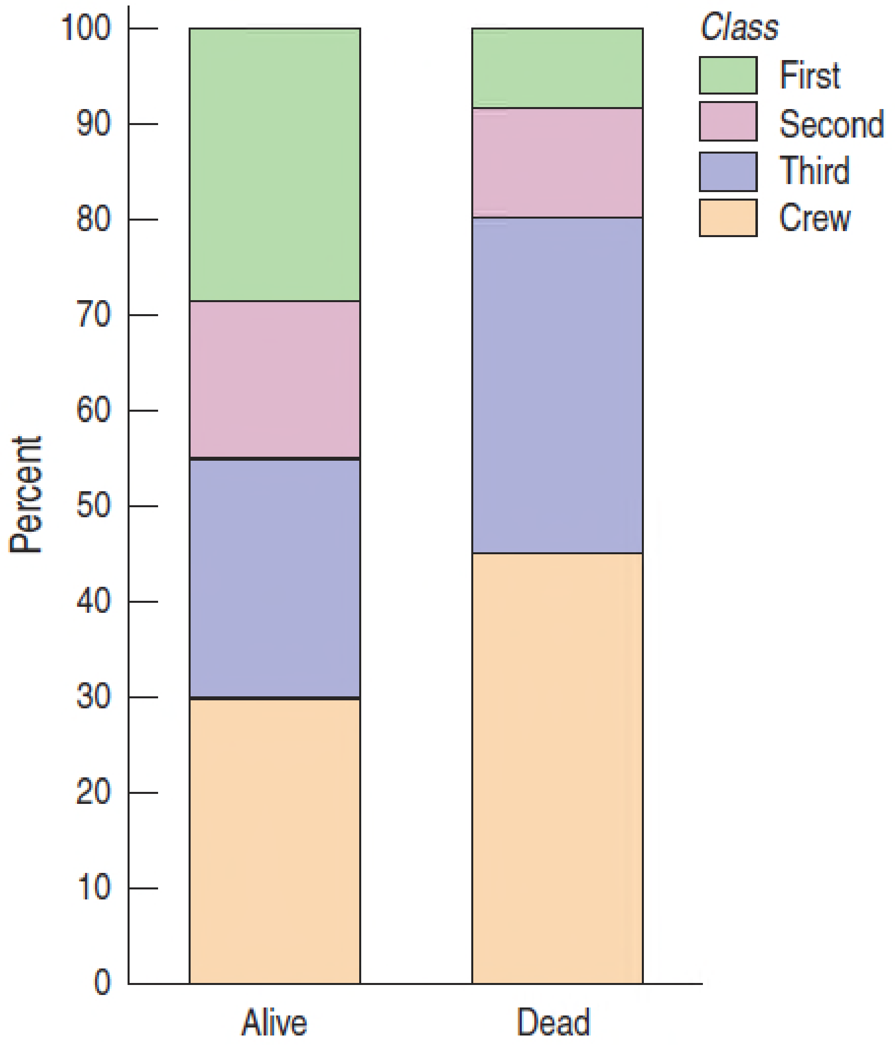
Here is a side-by-side bar chart showing the conditional distribution of *Survival* for each category of ticket *Class*:



Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.4 Segmented Bar Charts and Mosaic Plots

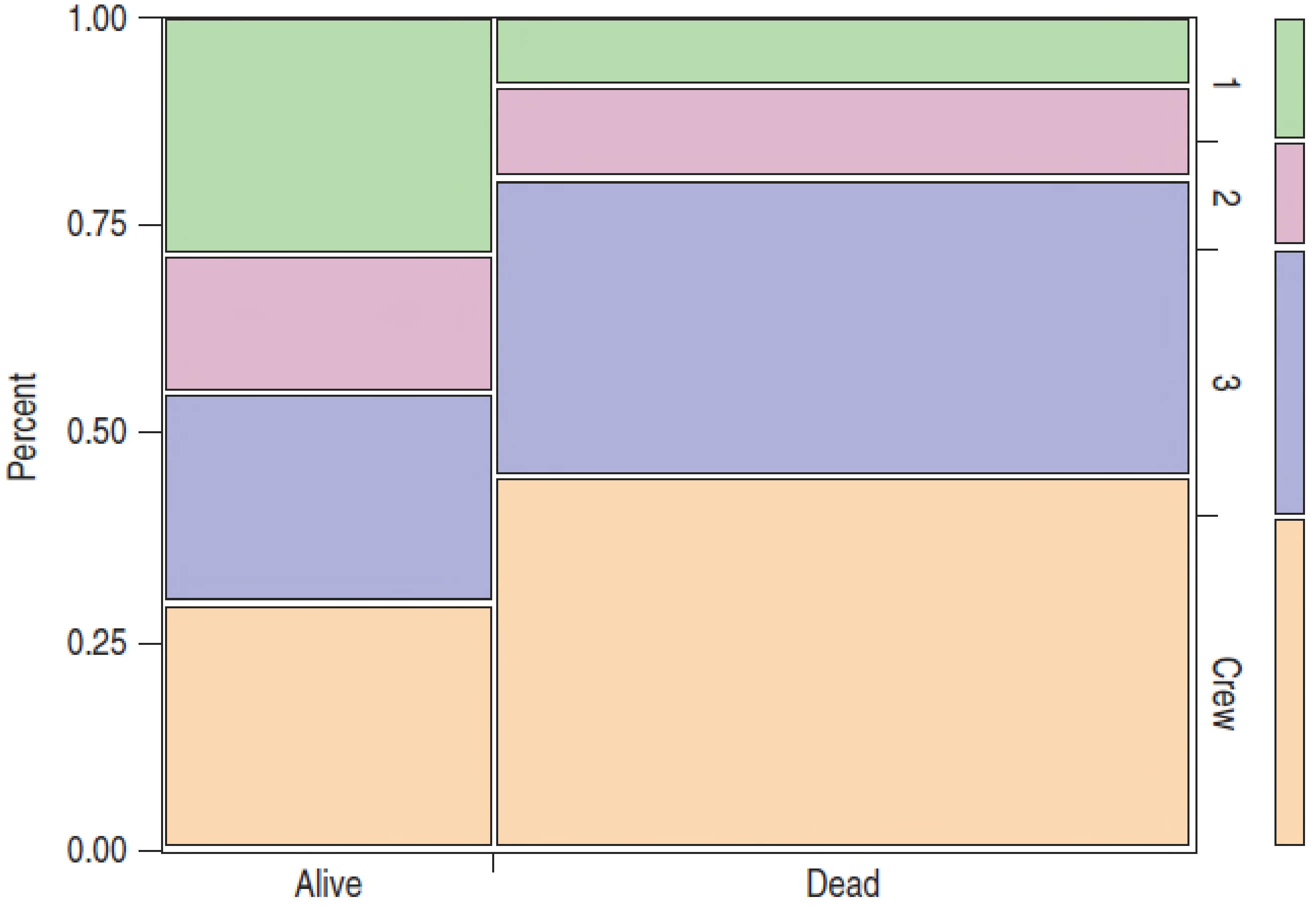
Here is a segmented bar chart. We can clearly see that the distributions of ticket Class are different, indicating again that survival was not independent of ticket class:



Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.4 Segmented Bar Charts and Mosaic Plots

Finally, here is a mosaic plot for Class by Survival.



Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.5 Simpson’s Paradox

Simpson’s Paradox

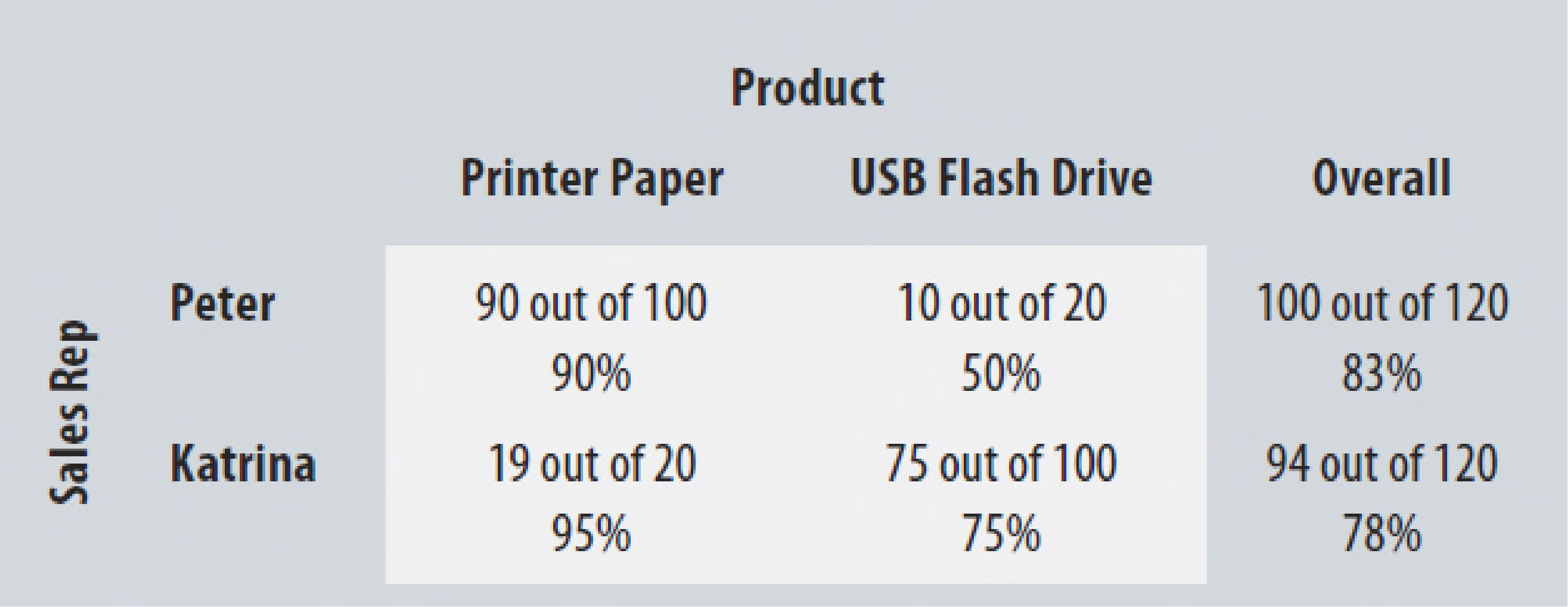
Combining percentages across very different values or groups can give confusing results. This is known as *Simpson’s Paradox* and occurs because percentages are inappropriately combined.

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.5 Simpson’s Paradox

Example

Suppose there are two sales representatives, Peter and Katrina. Peter argues that he’s the better salesperson, since he managed to close 83% of his last 120 prospects compared with Katrina’s 78%. Let’s look at the data more closely:



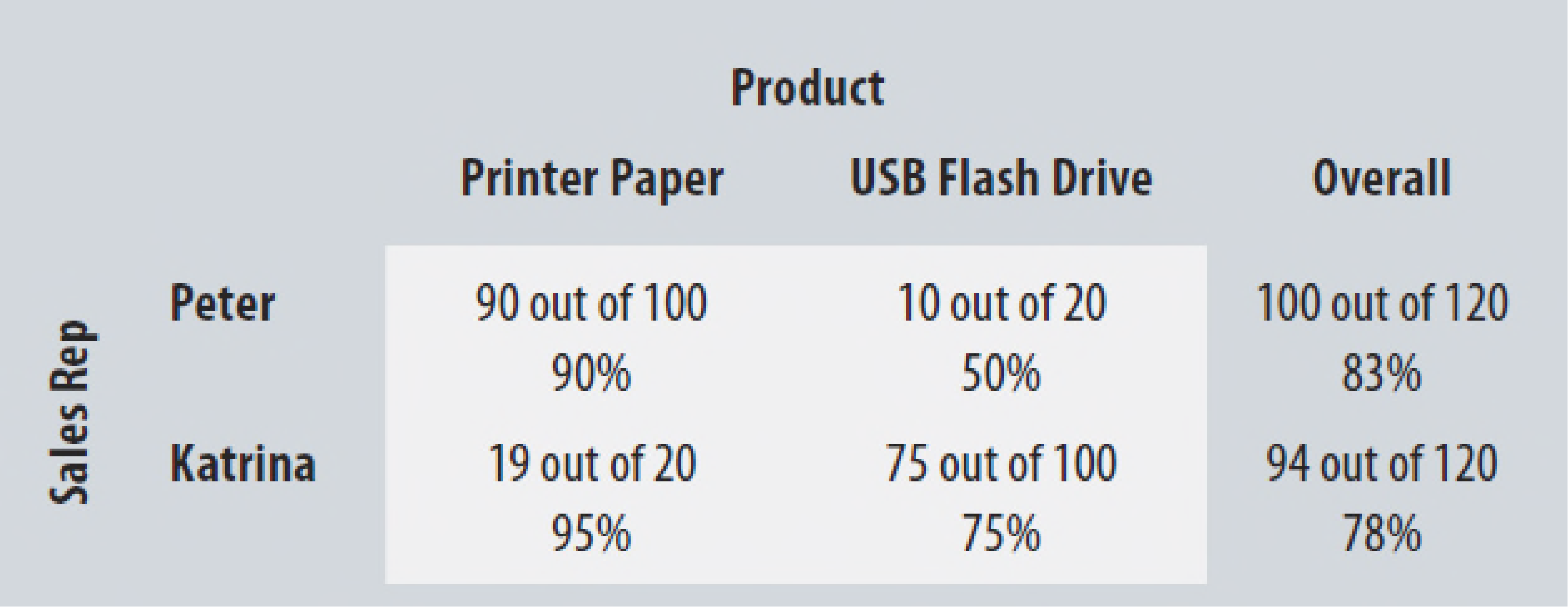
		Product		
		Printer Paper	USB Flash Drive	Overall
Sales Rep	Peter	90 out of 100 90%	10 out of 20 50%	100 out of 120 83%
	Katrina	19 out of 20 95%	75 out of 100 75%	94 out of 120 78%

Copyright © 2017, 2015 Pearson Education. All rights reserved.

2.5 Simpson’s Paradox

Example

Katrina is outperforming Peter in both products, but when combined, Peter has a better overall performance. This is an example of Simpson’s paradox, and results from inappropriately combining percentages of different groups.

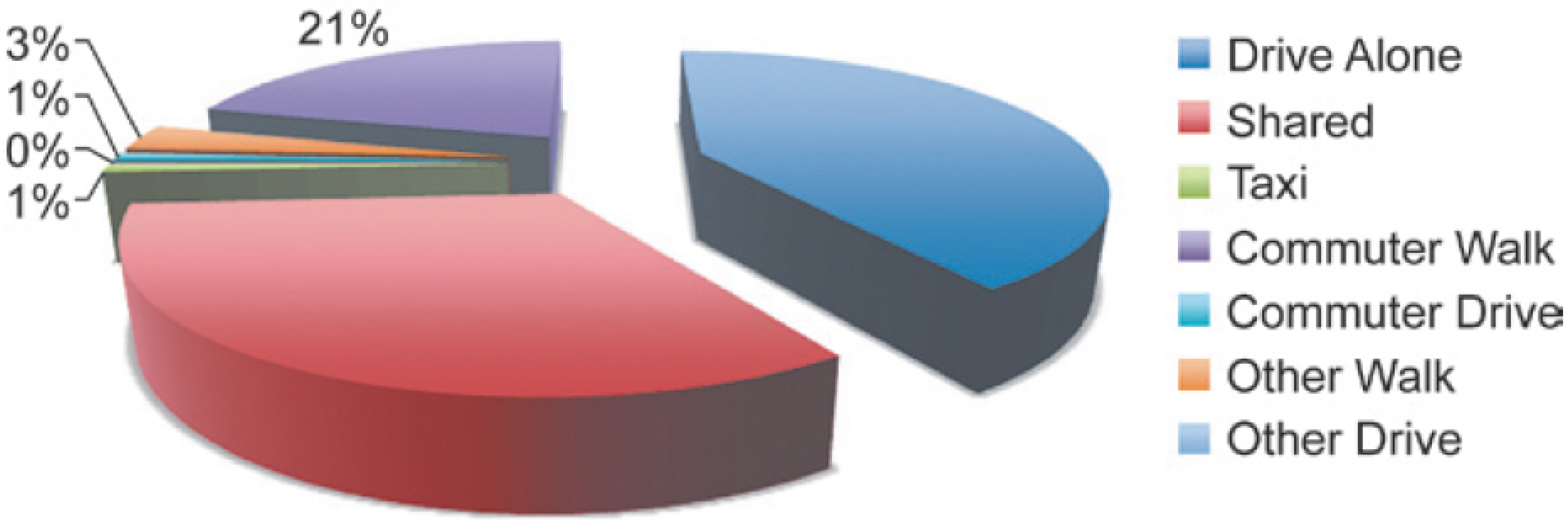


		Product		
		Printer Paper	USB Flash Drive	Overall
Sales Rep	Peter	90 out of 100 90%	10 out of 20 50%	100 out of 120 83%
	Katrina	19 out of 20 95%	75 out of 100 75%	94 out of 120 78%

Copyright © 2017, 2015 Pearson Education. All rights reserved.

What Can Go Wrong?

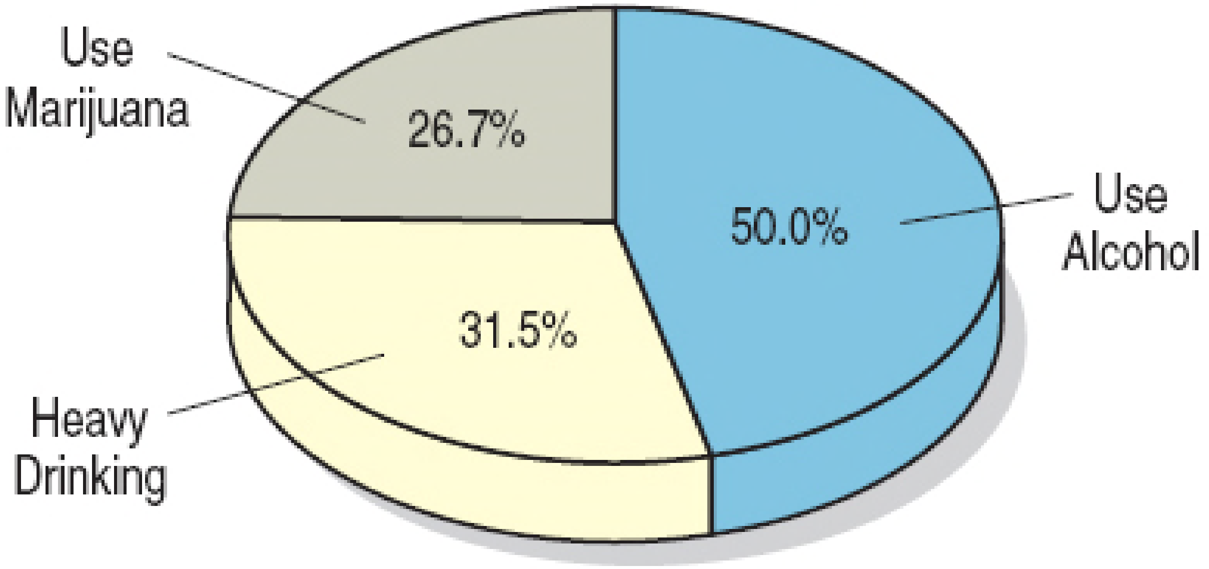
- **Don’t violate the area principle. “Shared” rides is actually 33%.**



Copyright © 2017, 2015 Pearson Education. All rights reserved.

What Can Go Wrong?

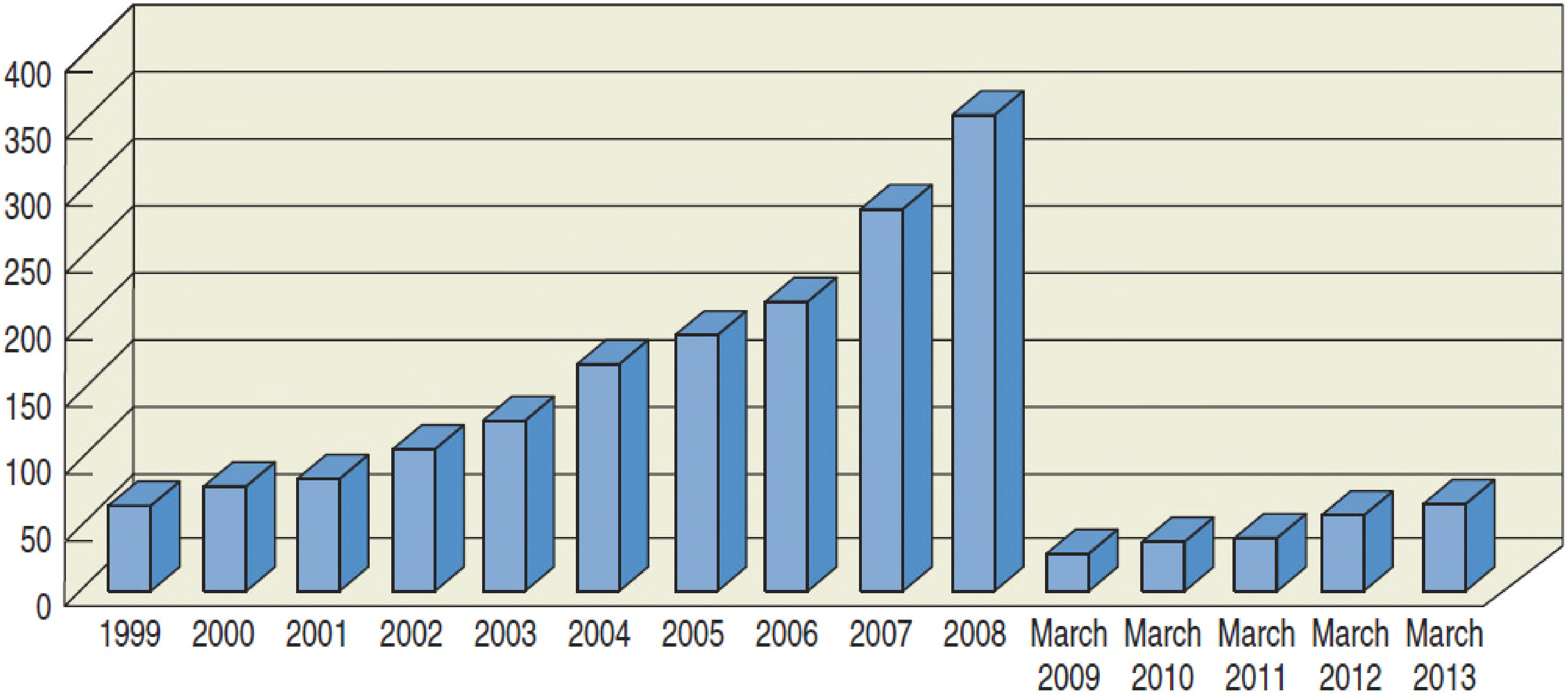
- **Keep it honest.**
  - The pie chart below is confusing because the percentages add up to more than 100% and the 50% piece of pie looks smaller than 50%.



Copyright © 2017, 2015 Pearson Education. All rights reserved.

What Can Go Wrong?

- **Keep it honest.**
  - The scale of the years change from one-year increments to one-month increments. This is misleading.



Copyright © 2017, 2015 Pearson Education. All rights reserved.

What Can Go Wrong?

- **Don’t confuse percentages – differences in what a percentage represents needs to be clearly identified.**
- **Don’t forget to look at the variables separately in** contingency tables and through marginal distributions.
- **Be sure to use enough individuals in gathering data.**
- **Don’t overstate your case.** You can only conclude what your data suggests. Other studies under other circumstances may find different results.

Copyright © 2017, 2015 Pearson Education. All rights reserved.

What Have We Learned?

Make and interpret a frequency table for a categorical variable.

- We can summarize categorical data by counting the number of cases in each category, sometimes expressing the resulting distribution as percentages.

Make and interpret a bar chart or pie chart.

- We display categorical data using the area principle in either a bar chart or a pie chart.

Make and interpret a contingency table.

- When we want to see how two categorical variables are related, we put the counts (and/or percentages) in a two-way table called a contingency table.

Copyright © 2017, 2015 Pearson Education. All rights reserved.

What Have We Learned?

Make and interpret bar charts and pie charts of marginal distributions.

- We look at the marginal distribution of each variable (found in the margins of the table). We also look at the conditional distribution of a variable within each category of the other variable.
- Comparing conditional distributions of one variable across categories of another tells us about the association between variables. If the conditional distributions of one variable are (roughly) the same for every category of the other, the variables are independent.

Copyright © 2017, 2015 Pearson Education. All rights reserved.

Visit TestBankDeal.com to get complete for all chapters