Chapter 2: Data

Mix and Match

1. Variable Name: brand of car; Type: nominal; Cases: drivers

2. Variable Name: household income; Type: numerical; Cases: households

3. Variable Name: color preference; Type: nominal; Cases: consumers in focus group

4. Variable Name: customer counts; Type: numerical; Cases: outlets of retail chain

5. Variable Name: item size; Type: ordinal; Cases: unknown (could be stocks in stores or purchase amounts)

6. Variable Name: shipping cost; Type: numerical; Cases: unknown (could be a time series or could be the costs for various items or destinations)

7. Variable Name: stock price; Type: numerical; Cases: companies (though the question is vague)

8. Variable Name: number absent; Type: numerical; Time Series Frequency: days

9. Variable Name: Sex; Type: nominal; Cases: respondents in survey

10. Variable Name: Education; Type: ordinal; Cases: customers

True/False

11. False. Zip codes are numbers, but these numbers are used only for identification and would not have any numerical meaning.

12. True.

13. False. Cases is another name for the rows in a data table.

14. True.

15. True.

16. False. A row holds an observation.

17. False. A Likert scale is used for ordinal data.

18. True.

19. False. Aggregation collapses a table into one with fewer rows.

20. True.

Think About It

21. (a) The data are cross sectional.
(b) The variables are whether the employee opened an IRA (nominal) and the Amount saved (numerical with dollars as the units).
(c) Did employees respond honestly, particularly when it came to the amount they reported to have saved?

22. (a) The data are cross sectional.
(b) The variables are Reaction to increase (categorical, or perhaps ordinal if asked to rate the chance of moving to another bank), Current balance and other aspects of the customer that would be useful additions to the data. (Bank may not care if it loses unprofitable customers.)
(c) How many customers responded to the questionnaire? Were their responses about leaving the bank sincere?

23. (a) The data are cross sectional.
(b) The variable is the Service rating (ordinal most likely, using a Likert scale).
(c) With only 450 replying, are the respondents representative of the other guests?

24. (a) The data are cross sectional.
(b) The variables are whether a coupon was used (nominal) and Purchase amount (numerical with dollars as the units).

(c) How were these homes chosen? Was there a time limit on redemption?

25. (a) The data are a time series.
(b) The variable is the Exchange rate of the US dollar to the Canadian dollar (numerical ratio of currencies).
(c) Are the fluctuations in 2016 typical of other years?

26. (a) The data are a time series.
(b) The variable is the Average time spent on the lot for ten car models (numerical for each model).
(c) Did dealers accurately report this information? Were all dealers surveyed, or just some of them? If it's a survey, did it concentrate more in some regions than others?

27. (a) The data are cross-sectional.
(b) The variables are the Quality of the graphics (ordinal from bad to good) and the Degree of violence (ordinal from none to too much).
(c) Did some of the participants influence the opinions of others?

28. (a) The data are cross sectional.
(b) The variables are Income (numerical with units in dollars), Sex (nominal), Location (nominal), Number of cards (numerical count) and Profit (numerical with dollars as the units, derived from other data).
(c) Why were these accounts sampled and not all of them?

29. (a) The data are cross sectional (though they could be converted to a time series).
(b) The variables are Name (nominal), Zip code (nominal), Region (nominal), Date of purchase (nominal or numerical, depending on the context; the company could compute the average length of time since the last purchase), Amount of purchase (numerical with dollars as the units) and Item purchased (nominal).
(c) Presumably the region was recorded from the zip code.

30. (a) The data are a time series.
(b) The variable is Vehicle type (nominal or ordinal as compact, regular, large and SUV)
(c) The mix of cars on the weekend may not be the same as on a weekday. Do employees get an accurate count since they have other things to do as well?

4M Economic Time Series

(a) Answers will vary, but should resemble the following.
By merging the data, we can see how sales of Best Buy move along with the health of the general economy. If sales at Best Buy rise and fall with disposable income, we might question the health of this company if the government predicts a drop in the amount of disposable income.

(b) A row in the data from FRED2 describes the level of disposable income in a month whereas a row in the company-specific data is quarterly, summarizing a quarter (3 months).

(c) The columns are both numbers of dollars, but with different multipliers. The national disposable income is in billions (so the value for January 2010 means that consumers have $11.041 trillion annually to spend). The quarterly sales are in millions (so Best Buy's net sales in the first quarter of 2010 were $3.036 billion).

(d) We can aggregate the monthly numbers into a quarterly number such as by taking an average (FRED2 will do this for you if you want to return to the web site). Alternatively, we could take the quarterly number and spread it over the months. That's a bit hard to do, so the first path is more common.

(e) Name the columns Net Sales ($ billion) and Disp Income ($ trillion) and scale as shown previously. That avoids lots of extraneous zeros if you were, for example, to label them all as dollars. The dates might best be recorded in a single column as, say, 2010:1, 2010:2, and so forth, or in the style shown in the following table.

(f) Here's the merged data table for 2010:

Quarter	Net Sales ($ billion)	Disp Income ($ trillion)
Jan-2010	$3.036	$33.124
Apr-2010	$3.156	$33.593
Jul-2010	$3.233	$33.860
Oct-2010	$4.214	$34.277

(g) Sales at Best Buy rocket up in the fourth quarter (30% higher during the holiday season), but consumers don't have that much more money to spend. Looks like some people spend a lot more during the holidays, no surprise there!

4M Textbooks

(a) Various sources report that books cost about $100 per class. In 2003, U.S. Senator Charles E. Schumer of New York released a study showing that the average New York freshman or sophomore pays $922 for textbooks in a year. So reducing the cost 5% would save $46.10 a year and by 10% would save $92.20 a year.

(b) Your table should have headings like these. You should use the names of the stores you shopped at if different from these. The first two columns are nominal, with the first identifying the book and the second giving the label. The two columns of prices are both numerical.

Book Title	Type	Price at Amazon	Price at B&N

(c) These will vary. Presumably, you've got five textbooks from your current classes. Hopefully, you've also got some other personal books. For popular books, you might consider books on one of the best-seller lists or those at the top of the lists offered on-line.

(d) You may have to change the list of books, particularly for textbooks. Some on line sites have a limited selection of these.

(e) You should include all of the relevant costs. Some Internet retailers add high shipping costs.

(f) Again, answers will vary depending on the choice of books and the choice of stores. The key to notice is the value of comparison. Because you've got two prices for the same books, you can compare apples to apples and see whether one retailer is systematically cheaper than the other

Chapter 3: Describing Categorical Data

Mix and Match

In each case, unless noted, bar charts are better to emphasize counts whereas pie charts are better to communicate the relative share of the total amount.

1. Proportion of autos: pie chart is the most common; a bar chart or Pareto chart can also be used.

2. Types of defects: Pareto chart (a bar chart with the categories sorted in order of the most common defect)

3. Coupons: bar chart or Pareto chart (these are counts) or perhaps a table (only three values)

4. Type of automobile: bar chart or Pareto chart (counts) or pie chart (shares)

5. Destination: bar chart or Pareto chart (counts) or pie chart (shares)

6. Hanging up: Pareto chart (counts)

7. Excuses: Pareto chart (counts)

8. Brand of computer: bar chart (counts) or pie chart (shares)

9. Software: pie chart (shares) or perhaps a table (only three values)

10. Camera: bar chart (counts), pie chart (shares), or a table (only three values)

11. Ratings: Bar chart or table (only four values). Because the values are ordinal, avoid a pie chart.

12. Loans: Bar chart or table (only three values). Because the data is ordinal, it should not be put into a pie chart – even though the plot shows shares.

True/False

13. True, but only in general. For variables with few categories, a frequency table is often better, particularly when the analysis requires knowing the detailed frequencies.

14. False. The measure of variation has to be within one category.

15. False. The frequency is the count of the items.

16. False. A relative frequency is a proportion.

17. True. It would be false if the variable were ordinal; you should not put the shares of an ordinal variable into a pie chart.

18. False. The proportion must match the relative frequency.

19. True.

20. False. It has fewer bars.

21. True.

22. False. The median only applies to ordinal variables and identifies the category of the middle value.

Think About It

23. The message is that customers tend to stick with manufacturers from the same region. Someone trading in a domestic car tends to get another domestic car whereas someone who trades in an Asian car tends to buy another Asian car. There's not a lot of switching of loyalties. The more subtle message, one that is disturbing to domestic car makers, is that those who own Asian cars are more loyal (78% buy another Asian car compared to 69% who stick with a domestic car). That makes it hard for domestic manufacturers to win back customers, even if they improve the quality of their cars.

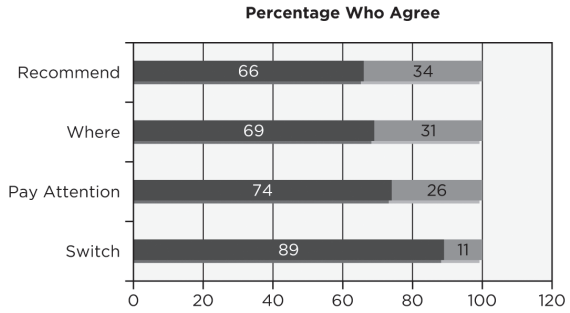
24. The answer is yes. Since lighting makes up 37% of the use of electricity, reducing the demand for electricity by using more efficient bulbs can have a substantial impact. Compact fluorescent bulbs produce the same amount of light with much less, say one-quarter, of the electricity used by an incandescent bulbs. Less energy also implies less heat and lower cooling costs. That said, the benefit of these savings for utilities is less pronounced because these savings happen mostly at night, not during the times of peak load that occur during the daytime.

25. This is a bar chart if you think about the underlying data as labeling the dollars held in these countries. The intent of the plot is to show the relative sizes of these counts, comparing the shares of U.S. debt held in these countries.

26. No, this is not a bar chart in the sense of this chapter. The chart uses bars to show a very short time series with five data points, the annual revenue in 2011-2015. Hence, it is a time plot that uses bars to show the data.

27. (a) No, these categories are not mutually exclusive. These percentages summarize four dichotomous variables, not one variable.

(b) Divided bars such as these might work well. This style is commonly used in reporting opinion poll results in the news. Sorting the values so that the percentages are in order also makes for a cleaner presentation.



28. (a) No. Each customer could report several of these items, so the categories are not mutually exclusive.
(b) A figure such as the divided bars used in Exercise 27 would be useful to illustrate the varying shares.

29. No. These percentages only list the percent of executives that report each problem. The categories are not mutually exclusive; some of the executives listed several issues.

30. The percentages do not add to 100; we need another category (which has a 9% share of the market).

31. A bar chart would not be appropriate for this situation since the variable, the amount spent by the last 200 customers, is a continuous variable.

32. This grouping data into categories is one step in constructing a frequency table. A histogram should be used since it reports a continuous variable, not a bar chart (a bar chart should be used when you deal with category variable).

33. The bar chart would have one very long bar (height 900) and five shorter bars of height 20 each. The plot would not be very useful, other than to show the predominance of one category.

34. A pie chart would devote 90% of its area to the main category and divide the remaining area into five small slices, each with equal area of 2%.

35. The bar chart would have five bars, each of the same height.

36. The bar chart. It would be hard to tell in the pie chart that the slices were of the same size (however, if the slices were labeled with the percentages it would be the same).

37. A bar chart is preferred because the categorical nature of the variable. A pie chart also can be used to present this data. A frequency table is not appropriate in this situation.

38. With so many categories (the 51 states, including Washington D.C.), some aggregation by region might be useful. Alternatively, it might be good to highlight the most common states, and combine the rest together into a separate, other category. A bar chart or pie chart could be used, and a frequency table would be fine if there were only a few states represented.

39. The mode is Public. There's no median for this chart since this is nominal data.

40. The East is the modal location. To find the median size, notice there are 50 sizes given, so the median is the size in position 25 or 26. Both lie in the category 10,000 to 19,999. The percentages of enrollment categories, not counts, should be shown in a pie chart.

41. The manufacturers want to know the modal preference because it identifies the most common color preference. Color preferences cannot be ordered, the median color preference can't be defined.

42. A median rating of Excellent implies that at least one-half rated the service as Excellent. A modal rating of Excellent implies that this is the most common rating, but far fewer than half might have picked this rating.

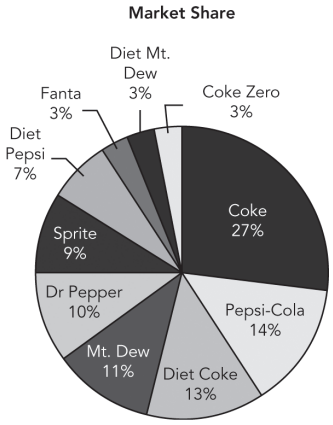
43. The radius of the circle for consumer electronics, for example, would have to be sqrt (10.5/9.1) times the radius of the circle for cell phones, sqrt(10.5/6.3) times the radius for chips, and sqrt(10.5/5.7) times the radius for LCD panels.

44. For instance, render as dollar bills with area determined by the amounts.

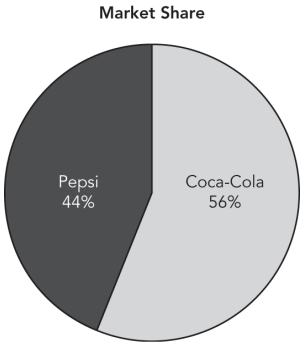
You Do It

45. (a) It probably accumulates case sales by brand over some period, such as daily or weekly. It is unlikely that every case is represented by a row.

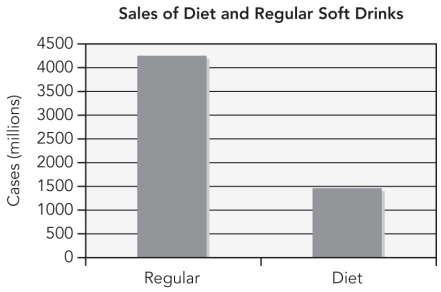
(b) A pie chart emphasizes shares.



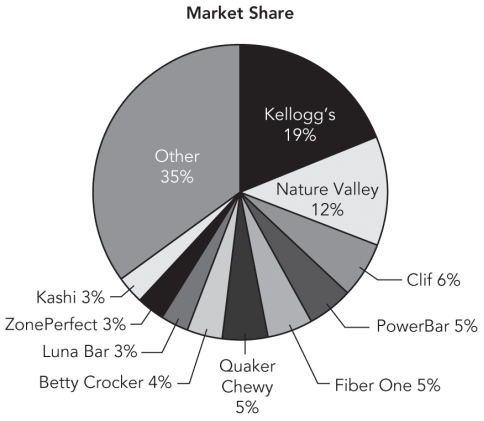
(c)



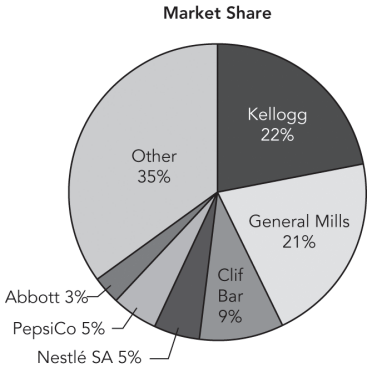
(d)



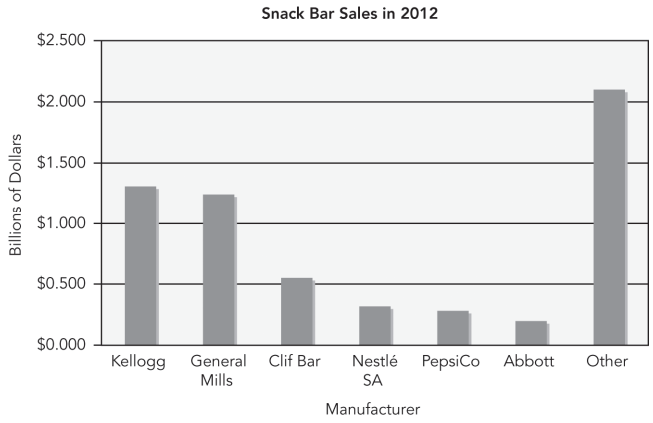
46. (a)



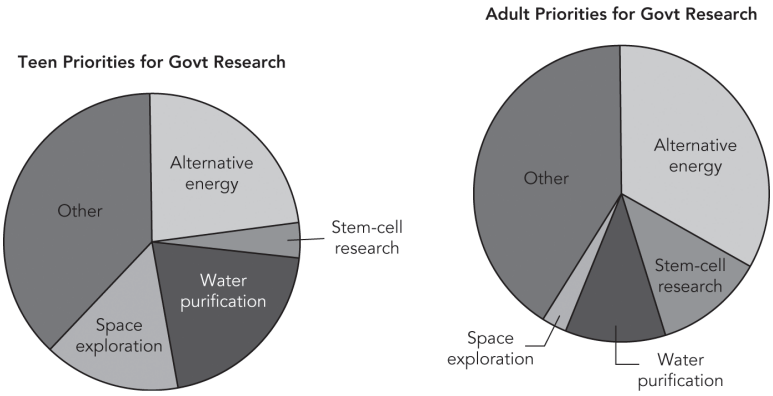
(b)



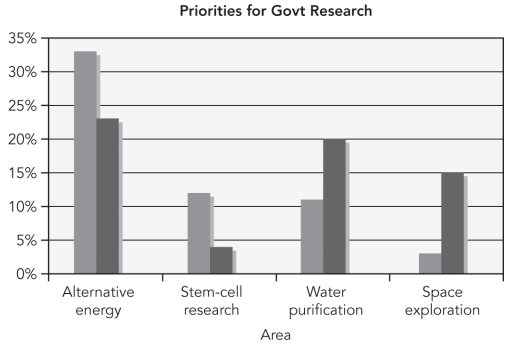
(c)



47. (a) The Other category forms an additional row in the tables so that each column adds up to 100%. The addition of this extra row makes up a big part of both pie charts.



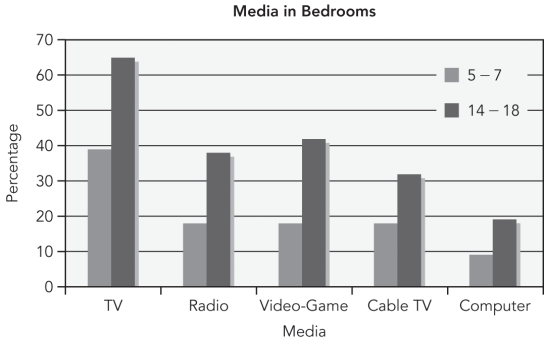
(b) The side-by-side bar chart works well for this. Notice that we no longer need the Other category that dominates the pie charts.



(c) No, because the categories would no longer partition the cases into distinct, non-overlapping subsets. A pie chart should only be used to summarize mutually exclusive groups.

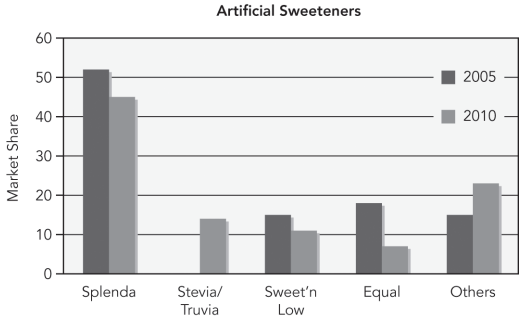
48. (a) The totals within a row do not sum to 100%. The columns give the proportion with different types of media, and these can sum to more than 100% as well. The represented categories do not divide the homes into different groups; a bedroom could have all of these media.

(b) The side-by-side bar chart shows more of every type of media in the rooms of older children.

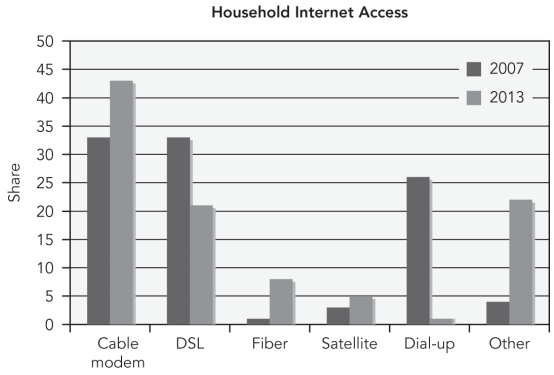


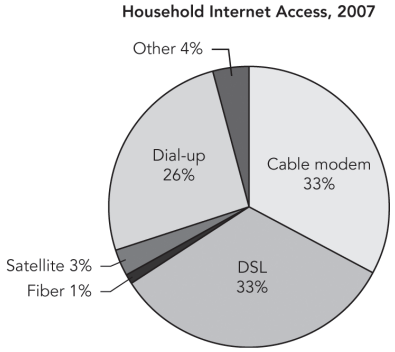
(c) The big adoption of games appears to happen in the 8-13 age range.

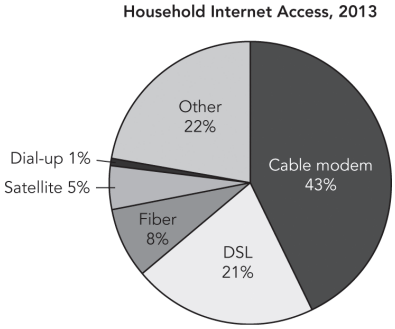
49.



50. (a) The pie charts are favored slightly over the side-by-side bar charts.

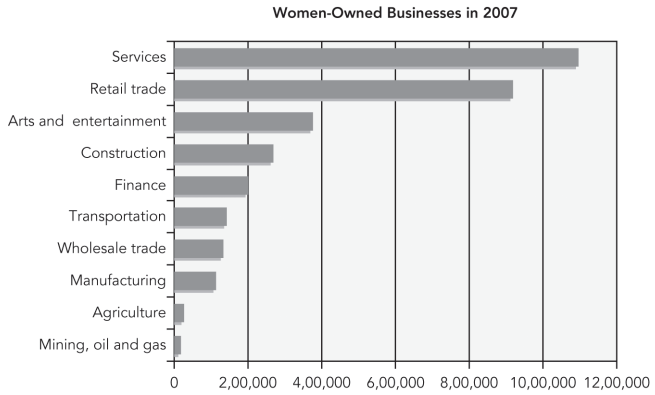


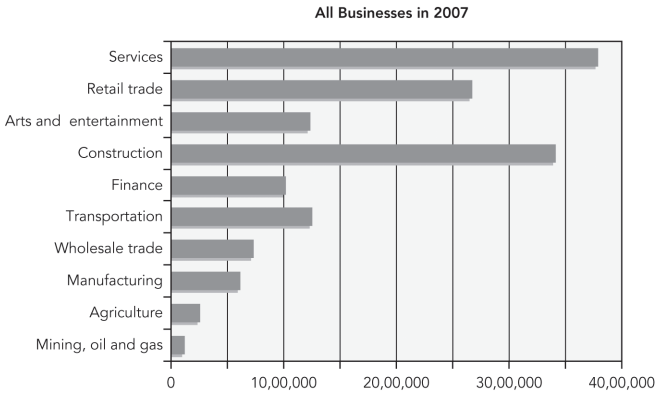




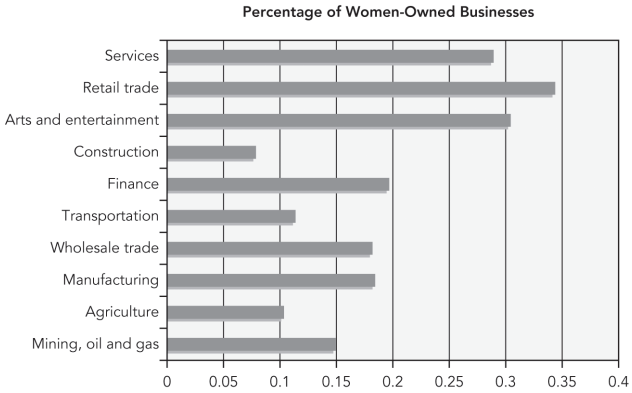
(b) No; to compare, count, not percentages, are needed.

51. (a)





(b)

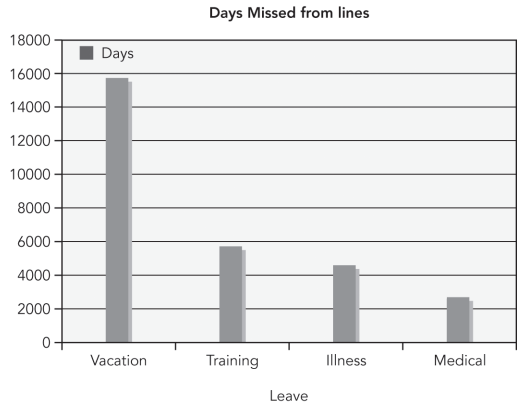


52. (a) No, the percentages do not add; some respondents gave more than one reason.
(b) No, unless you think there is a natural way to prioritize the reasons.
(c) A bar chart, with the length given by the percentage.

53. (a) Use a table with the two rows and the percentages (or proportions)

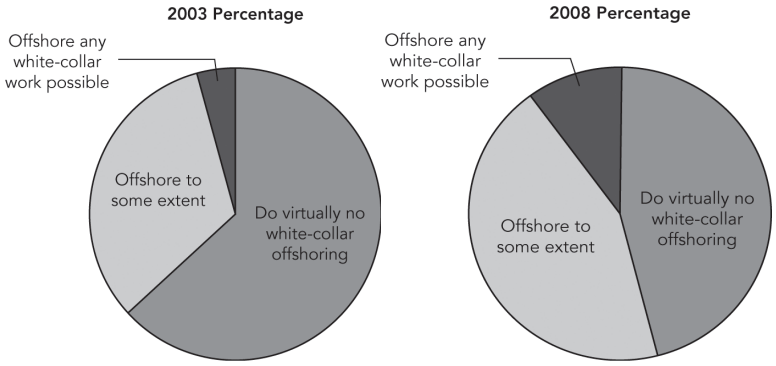
Unexpected illness	4,463	15.8%
Planned leave	23,735	84.2%

(b) A Pareto chart shows the categories in order of size.

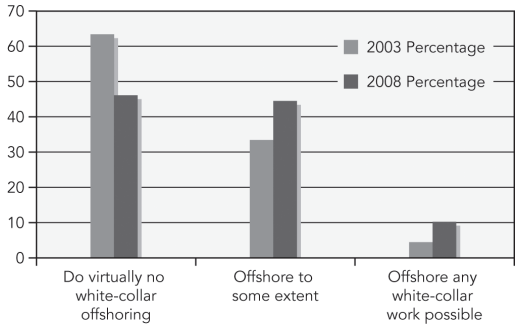


54. (a) Kraft plus Cadbury (15.2%) becomes the mode.
(b) Yes, if the Hershey items are much less expensive than those of other brands.

55. (a) Yes, pie charts are fine because the responses are mutually exclusive and sum to 100%.



(b) Various answers are possible. The following layout is reasonable.

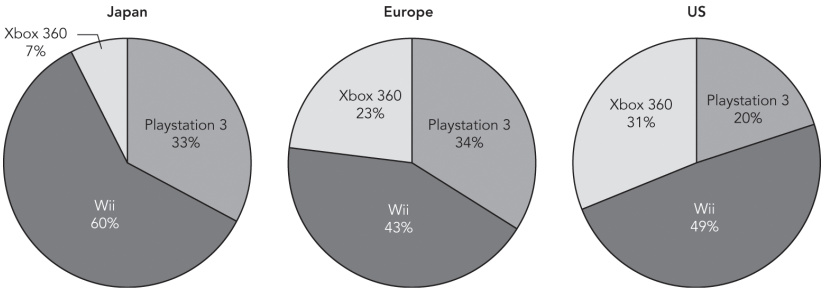


(c) The bar chart facilitates comparison. The pie chart makes the relative shares more apparent. For example, the 2003 pie shows a predominant share for taking no action, the only choice anticipated to fall in 2008.

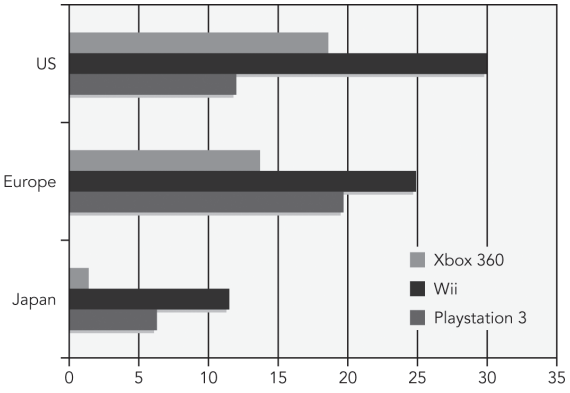
(d) The mode and median agree (virtually none) in 2003, but differ in 2008 as responses shift from the more consistent response to a tendency to do more off shoring.

56. (a) The radius for each region should be proportional to the total sales in that region, as suggested in this table:

	Japan	Europe	US
Playstation 3	6.3	19.7	12
Wii	11.5	24.9	30
Xbox 360	1.4	13.7	18.6
	19.2	58.3	60.6
radius	1	1.742543639	1.7765838
diamenter	2	3.485087278	3.553167601



(b)



57. (a)

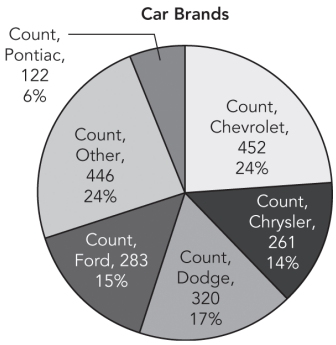
BUICK	Frequency 16	Relative Freq 0.00849
CHEVROLET	452	0.23992
CHRYSLER	261	0.13854
DODGE	320	0.16985
FORD	283	0.15021

(b) Chevrolet.

(c) The pie chart has too many slices.

(d)

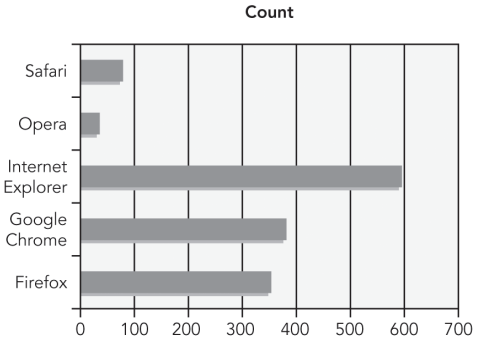
Brand	Count
Chevrolet	452
Chrysler	261
Dodge	320
Ford	283
Other	446
Pontiac	122

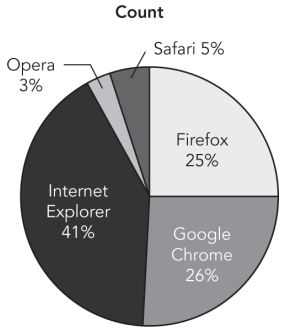


58. (a) Frequency table

Browser	Count
Firefox	354
Google Chrome	382
Internet Explorer	596
Opera	36
Safari	79

(b) The pie chart is easiest since it shows the percentages.





(c) Internet Explorer is losing share and Google Chrome has grown.

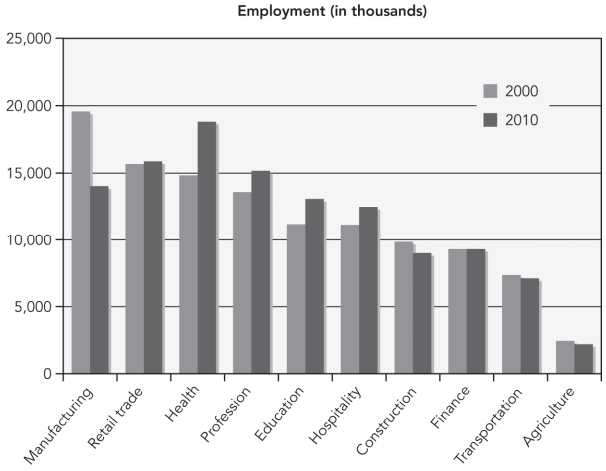
59. 4M Growth Industries

(a) It could use trends suggested by the table to indicate how to shift its sales force from declining industries to those that appear stronger and growing.

(b) A bar chart would show the counts and make it simpler to compare the counts within a year. A pie chart would emphasize the shares among industries in the two years.

(c) By looking at the changes from 2000 to 2010, you can see which industries are growing and which are shrinking. You could also use a side-by-side chart, but a chart of the differences does the subtraction for us.

(d) Grouped bar chart, order by size in 2010 to emphasize the change in manufacturing.



(e) Changes are shown in the differences in adjacent bar heights.

(f) One might choose to show just those that change, but also useful to see those that remained steady. Ten works fine.

(g) One can see the increase in health care and services and the big loss in manufacturing and in construction (mortgage debt crisis).

(h) Percentage shares are not evident, and the plot hides other types of employment.

60. This table shows the distribution of the hosts.

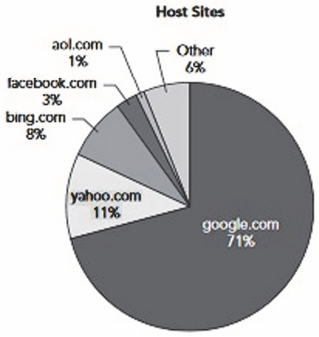
Host	Count	Proportion
google.com	3561	0.7122
yahoo.com	535	0.107
bing.com	384	0.0768
facebook.com	156	0.0312
aol.com	78	0.0156
imdb.com	43	0.0086
ask.com	41	0.0082
comcast.net	38	0.0076
humblebundle.com	33	0.0066
buyholidaygift.com	27	0.0054
reddit.com	23	0.0046
gevella.com	22	0.0044
yidio.com	20	0.004
youtube.com	20	0.004
redbubble.com	19	0.0038

Answers vary for this question. A sample answer is:

Motivation Amazon would benefit by forming a relationship with newly popular hosts that send shoppers to its Web sites.

Method Use bar or pie charts to compare the distribution of hosts in 2014 to that seen in the text example.

Mechanics A pie chart such as the following works well here to show the domination of Google. It is useful to combine the smaller hosts into an "Other" category. (Compare to Figure 3.6 in the text.)



Message Google dominates the search market and, not surprisingly, is the major host that sends shoppers to Amazon. Google was the third most popular host in 2004, but in 2014 it dominates, sending more than 70% of the visitors who use a host. An important caveat notes that these are shoppers who use a host, omitting those who come to Amazon directly.

Chapter 4: Describing Numerical Data

Mix and Match

1. g

2. a

3. h

4. i

5. j

6. b

7. d

8. c

9. e

10. f

True/False

11. False. The box is the median, with its lower edge at the 25% point (lower quartile) and its upper edge at the 75% point (upper quartile).

12. False. They are outliers, but that's no reason to automatically remove them from the data.

13. True.

14. False. The multiple modes may indicate meaningful subsets in the data and should not be hidden by making the bins artificially wide.

15. True.

16. True.

17. False. The Empirical Rule applies only to numerical variables that have a symmetric, bell-shaped distribution.

18. False. 5% are larger than 2 or less than -2.

19. True.

20. False. The IQR is the difference from the 25% to 75% points.

21. True. In the absence of variation, all of the data values are the same, and hence the mean and median are equal to this common value.

22. False. The variance is the average squared deviation, so adding more observations does not necessarily cause the variance to grow. The range, however, can only stay the same or increase as more data are added to a variable.

Think About It

23. You cannot tell from the median. There could be, for example, one very, very long song that filled the Shuffle by itself. Because this one large song does not affect the median, the median could be small but the songs would not fit.

24. (a) No. All you need is the total amount of room required in megabytes, and you can get that as 1,000 times the mean.
(b) The size of the SD is not relevant.

25. The histogram of incomes in the United States is very heavily right skewed. It's hard to get very far below zero (people with negative incomes have received tax credits), but the upper limit is in the stars.

26. Because of the right skewness of U.S. incomes, the mean income is larger than the median income.

27. The payments on an adjustable rate mortgage are more variable than those on a fixed rate mortgage. If interest rates climb, then the required payments climb as well. In this context, variation is "bad" in the sense that the homeowner might see a large increase in the monthly payment if interest rates climb. Fixed rate mortgages hold the payment fixed, eliminating the variation in payment.

28. The purpose of interval training is to increase the variation in the workout. In this context, larger amounts of variation are "good", assuming you think that interval training is more helpful than a steady but hard effort.

29. The range is very sensitive to outliers because it is the difference between the two most extreme values. The presence of large (or small) outliers increases the size of the range.

30. The range tends to increase with the sample size. It is the distance between the most extreme values and cannot get smaller as more cases are added. Because it is an average, the variance (and consequently the SD) does not have this property and tends to stabilize as the number of rows grows.

31. Mortgage payments have a larger SD because these payments are so much larger than allowances. It's likely, though, that the coefficient of variation may be larger for the allowances. The mean mortgage payment is much larger than the mean allowance!

32. The prices of used cars are going to be much more variable, so s^2 is larger for these. Car prices vary by the thousands of dollars, whereas items in a market vary by the 10s of dollars (and usually much less). The coefficient of variation for the two might in fact be larger for the grocery; it sells a much wider variety of items. The new car dealer is only selling cars; the grocery is selling everything from gum to filet mignon.

33. (a) The group with the largest mean, music only.
(b) No, because you cannot recover the total amount from the median.
(c) No, because we do not know that every shopper bought the same amount; we'd expect substantial variation in the data with overlap between groups.

34. (a) Saturday-Sunday.
(b) The shown SDs indicate substantial overlap. For example, the mean sales amount on Saturday is about $400 more than on Sunday; the SD for Sunday alone is more than twice this amount. Even though the data are likely somewhat skewed, the Empirical Rule suggests considerable overlap.
(c) Given the hint about consecutive days, you'd like to see if there's a sequential pattern in sales, such as increasing during the summer holiday months.

35. No. If the distribution is bell shaped, the Empirical Rule suggests this is a common occurrence. Even if it's not bell shaped exactly, 1 SD is not far from the mean.

36. No. One-sixth is the value from the Empirical Rule. Expect fewer for right-skewed data because these tend to bunch up on the left close to the mean, as in the histogram in the text of the lengths of songs.

37. (a) Right skewed, with a peak at zero (for those without an iPod) and tailing off to the right.
(b) Right skewed with one mode, from moderate prices to very large orders.
(c) Bell shaped around the target weight (or perhaps a weight above the target).
(d) It depends, but this distribution may be bimodal if there is a mix of male and female students.

38. (a) Either uniform from about 18 to 22 or a bit bell shaped depending on the community near campus.
(b) Right skewed, with some at zero (buying for relatives, perhaps) to those with lots of kids.
(c) Slightly right skewed (if the business is not very busy), but with a definite bell shape near the center at the typical level of business.
(d) Bimodal distribution. The four weeks before Christmas we would expect to see a lot of packages shipped whereas August is typically the slow vacation season.

39. (a) 11. Add the heights of the two bins located between 4 and 5.
(b) The mean is slightly less than 5% (4.88). The outliers that charge no tax (or very small) pull the mean to the left.
(c) The median is larger. The outlier to the far left pulls down the mean, but has less impact on the median.
(d) The rates are rounded to values like 6.5 or 5.5, producing the isolated peaks.
(e) The SD is about 2. The mean is near 5%. If the SD were 5, then the mean ± 1 SD would hold all of the data. It would take a range of about ±3 SDs to make that happen.

40. (a) The interval is from $700 to $800.
(b) The mean and median are both approximately $900.
(c) The mean is slightly larger.
(d) Highly urban states with greater population density, wider intervals obscure the mode on the right.
(e) The highest is not exceptionally distinct. The largest premium is in New Jersey, closely followed by New York and the District of Columbia.
(f) Closer to $200.

41. (a) The mean is $34,000 compared to the median at $27,000. You can get the median from the centerline of the boxplot. You know that the mean is larger because of the skewness, but it is hard to guess how much larger. It would have to be very skewed, however, for the mean to exceed the upper quartile (which is less than $50,000).
(b) The IQR is about $30,000.
(c) Usually, the SD is smaller than the IQR (about 2/3 of the size), but the skewness and presence of outliers changes that. The two are about the same size here, but the SD is slightly larger.
(d) Only the labels on the *x* axis would change, dropping 3 zeros from each. Otherwise, the figure is identical.

42. (a) The median (the line in the center of the box) is 25%, and the mean is slightly larger (those outliers and peak at the right) at 33%.
(b) The IQR (the length of the box) is 24%.
(c) The SD = 25% is larger in order to accommodate the outliers and tall bin at the right. Outliers have a large impact on the SD, more so than on the mean.
(d) The tall bin occurs above 100% because quite a few households report paying more in rent than they report as earned income. Either someone is getting rent support from a family assistance program (for example) or some income is not being completely reported.

43. The pricing errors will lead to more variation in prices and a more spread-out histogram.

44. No. In fact, most gasoline sold contains little if any ethanol. The content must only average 2.78%; this is not a requirement for any particular gallon of gasoline.

45. The distribution of income is very right skewed, with the upper tail reaching out to very high incomes. In this case, the mean will be larger than the median.

46. Families with high incomes saw an increase, but incomes fell for a majority of families.

47. (a) Multiply each of these summary statistics by 60, as shown below.

Summary	*File Size MB*	*Song Length Sec*
Mean	3.8	228
Median	3.5	210
IQR	1.5	90
Standard deviation	1.6	96

(b) The mean and median would increase by 2 MB, but the IQR and SD would remain the same.
(c) Such a song is longer than all of the others (see the text discussion). Hence, the median would stay about where it is (perhaps shifting up to the next larger value in the sorted list of sizes) and the IQR would remain the same.

48. (a) Divide each of these summary statistics by 1.2, as shown below.

Summary	*Price ($)*	*Price (€)*
Mean	30,300	25,250
Median	29,200	24,333
IQR	5,700	4,750
Standard deviation	4,500	3,750

(b) The mean and median would increase by $500, but the IQR and SD would remain the same.
(c) The median would be slightly larger; the IQR would remain the same.

49. The shape will be the same as in Figure 4.1, but the labels on the x-axis will change to 60, 120, 180, and so forth. The count axis and bin heights would be the same.

50. (a) The shape would be the same, but the labels on the x-axis would change ranging from about €16,666 to €37,500. The count axis and bin heights would be the same.

(b)The histogram would shift to the right along the x-axis by $500.

51. (a) The mean and SD are $18,000 and $10,000, but these do not capture the bimodal nature of the data.
(b) The data are bimodal with cluster centers having means of about $7,500 and $30,000.
(c) It's hard to see, but in fact, there's a clue that the data is multi modal in the boxplot. Because the groups are of comparable size, the box of the boxplot is long relative to the size of the whiskers on either side. Unless you're looking carefully, however, you'd never recognize that this is a signal for a bimodal shape.
(d) The schools in the cluster with mean near $7,500 are public; the others are private schools.

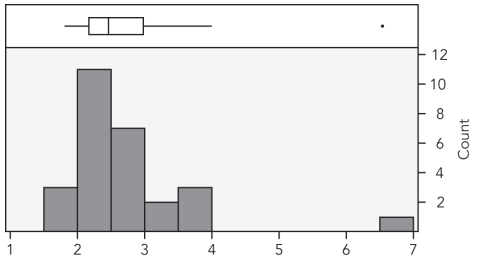
52. (a) The mean is 130 seconds with SD = 13.5 seconds. These are poor summaries and not suited to the bimodal shape.
(b) The histogram is clearly bimodal.
(c) Because the groups have rather different sizes, the boxplot in this example (unlike the prior example) clearly shows the clusters. The boxplot is very small, located over the one larger pile of times. A cluster of outliers identifies the location of the second, smaller clump of times.
(d) The bimodal shape is caused by a change in the length of the race! The races since 1896 have been shorter than those run earlier.

You Do It

53. (a) Multimodal, with modes at 3.5 and 2.5 liters.
(b) Roughly unimodal, with a dip between 4-5 liters.
(c) Yes; wide bins conceal the popularity of engines near 2 and 3 liters.
(d) $\bar{y} \approx 3.82, s \approx 1.35$ $\bar{y}$ lies at the center of the distribution; most engines are between $\bar{y} \pm 2s$ liters in size.
(e) $s / \bar{y} \approx 0.35$; engine sizes are more variable relative to the mean than for manufactured items like MMs.
(f) Very large or small engine.

54. (a) The histogram is unimodal, slightly right-skewed.
(b) The boxplot locates the median and quartiles; the histogram shows unimodal shape.
(c) $\bar{y} \approx 20.13, s \approx 4.76$ $\bar{y}$ lies at the center of the distribution, and most cars obtain between 10 to 30 MPG ($\bar{y} \pm 2s$).
(d) $s / \bar{y} \approx 4.76/20.13 \approx 0.24$; the variation relative to the mean is large compared to manufactured items meant to be similar.
(e) Scion IQ; no
(f) 82/413 meet the goal; many of these vehicles are trucks.

55. (a) This histogram and boxplot follow.
(b) The large outlier is the song "Hey Jude", which goes on for 7 minutes and uses 6.6MB.



(c) This table shows these summaries, with and without "Hey Jude".

	Mean	*Median*
With "Hey Jude"	2.73MB	2.47MB
Without	2.59	2.43

If we exclude "Hey Jude", the mean falls from 2.73 down to 2.59 MB, about 5%. The median hardly changes. The median, because it relies only on sorting the values, does not account for their magnitude. As a result, the median is not sensitive to the presence of outliers and is often said to be more robust than the mean.
(d) The median is well positioned in the center of the large cluster. The mean is to the larger side and gives a less useful notion of the center of most of the data.
(e) For this task, the mean is more useful. Suppose you know only that there are 27 songs and that the median size is 2.47 MB. It would be tempting to think that you could fit all 27 songs from the Beatles #1 album in $27 \times 2.47 = 66.69$ MB. Unfortunately, you'd run out of room. The mean takes the size of all of the songs into account, and from the mean, we'd know that the total space is $27 \times 2.73 = 73.71$ MB.

56. (a) This table summarizes the distribution, with and without "Hey Jude", the large positive outlier to the right of the distribution. The calculation of the IQR may vary slightly depending on your software's calculation of the quartiles, but should be close to the values shown here.

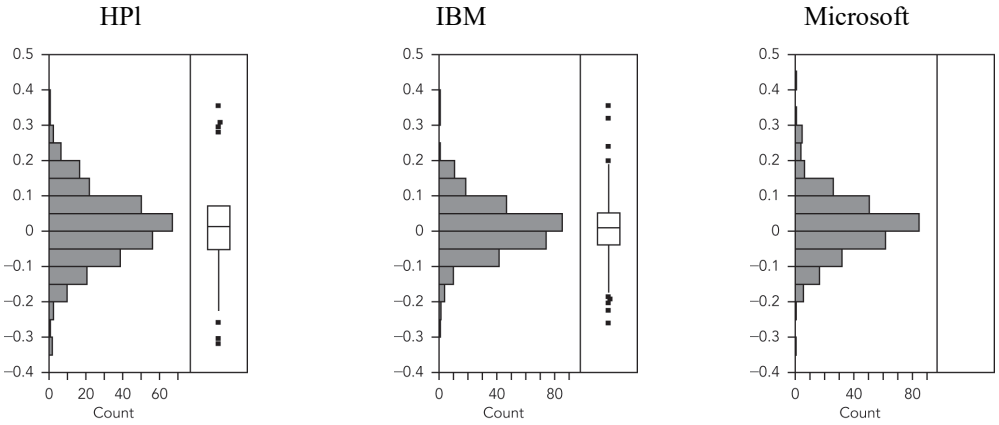
	With Outlier	**Without Outlier**
Maximum	6.5525	3.9932
Quartile	2.9795	2.9619
Median	2.4654	2.4276
Quartile	2.1596	2.1543
Minimum	1.8212	1.8212
IQR	0.82	0.81
Mean	2.7343786	2.5875276
Std. Dev.	0.9503484	0.5777002
N	27	26

(b) By setting aside Hey Jude and recalculating the summary measures, we can see exactly how the outlier affects the summaries. Because it is based on percentiles, not averaging, the IQR hardly changes at all. The SD falls about 60% of the size with all 27 songs.
(c) The SD changes more, proportionally. The mean falls from 2.73 to 2.59 (95% of its prior value), but the SD drops from 0.95 down to 0.58 (about 60% of its prior value).

57. (a) Median=$220 million, mean = $2.2 billion, SD = $9.6 billion
(b) Heavily right skewed; outliers conceal most of the data.
(c) Three most extreme outliers are ATT, Verizon, and Microsoft.
(d) Dominated by very large telecoms and Microsoft.

58. (a) The number of employees is missing for 18 companies.
(b) Median=$244,000 per employee, mean=$337,000 per employee, SD= $370,000 per employee.
(c) No; very right skewed.
(d) Flint Telecom; only 7 employees.
(e) 15 or 15 times median.

59. (a)

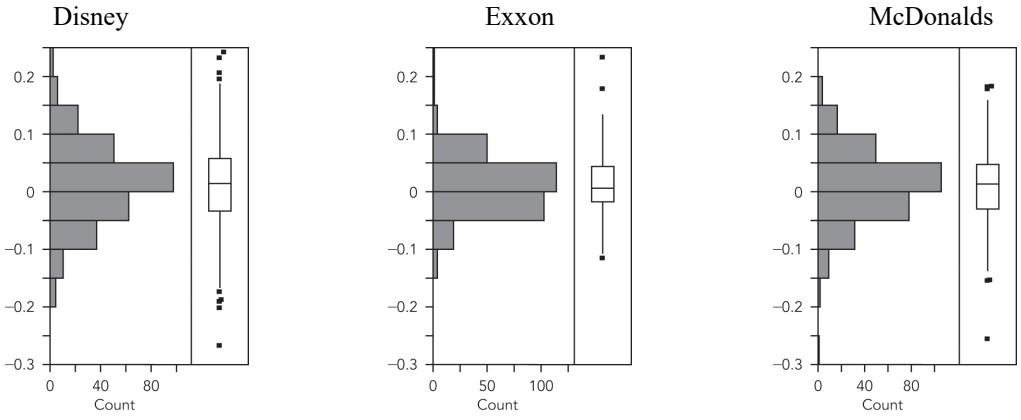


(b)

	HP	IBM	Microsoft
Mean	0.01359	0.01099	0.01994
Std. Dev.	0.10408	0.07931	0.09346
Coef. of var.	7.66	7.22	4.69

(c) The values of c_v indicate that returns on HP and IBM are much more variable relative to the mean level when compared to Microsoft. You cannot see that distinction in the histograms, and it may be due to how close the mean return on IBM is to zero.
(d) Large positive mean for the return and small SD for the return are ideal; the closer the c_v is to zero (large mean and small SD), the better.
(e) Only partially. Microsoft has the highest average return, but not the largest SD.

60. (a) The three histograms appear below. All are bell shaped with a few outliers. Each seems to have a slightly positive mean, with more variation in the returns on Disney and McDonalds, and less variation in the returns on Exxon.



(b) The means and SDs appear below. Because of the bell shaped distributions, the Empirical Rule works well for these distributions.

	Disney	Exxon	McDonalds
Mean	0.01130	0.01024	0.01135
Std. Dev.	0.07396	0.04683	0.06196
Coef. of var.	6.55	4.57	5.46

(c) The values of c_v indicate that returns on Disney are much more variable relative to the mean return than those on the others, then McDonalds, with Exxon the most steady. As in exercise 57, the c_v may be unreliable because the means are so close to zero.
(d) Loosely, Disney and McDonald's have distinctly higher average returns as well as larger SDs than ExxonMobil.

61. (a) The mean and SD of these returns appear in the solution for Exercise 59. The Sharpe ratios are below. Microsoft looks best: it is more volatile than IBM but has much larger average return.

	HP	**IBM**	**Microsoft**
Sharpe Ratio	0.106	0.107	0.187

(b) The mean is the Sharpe ratio for HP.
(c) The Sharpe ratio is used to compare returns on different stocks. Standardizing by forming *z*-scores is most useful (with the Empirical Rule) for identifying outliers and judging the relative size of different observations of the *same* variable.

62. (a) The mean and SD of these returns appear in the solution for Exercise 60. The Sharpe ratios are below. ExxonMobil is best: it has the smallest average return, but its return is much larger relative to its variation.

	Disney	**ExxonMobil**	**McDonalds**
Mean	0.01130	0.01024	0.01135
Std. Dev.	0.07396	0.04683	0.06196
Sharpe Ratio	0.119	0.165	0.143

(b) The Sharpe ratio for Disney (0.119).
(c) No. Exxon had the largest loss. Like any summary, the Sharpe ratio describes an average level of performance, but it's not going to tell you the ordering of the three for every month. There's variation in the returns, and this month, Exxon was the loser, even though on average it came out on top.

63. 4M Financial ratios
Motivation

(a) One number is a more convenient, plus it is easily interpreted like a percentage.

(b) Percentage of total assets.

(c) The retailer might want to have a high return on assets compared to competitors.

(d) Many are possible, such as profit/(stockholder value) or profit/employee. A single score makes it easier to define a goal rather than having a two-part goal.

Method

(e) Histogram, boxplot.

(f) Bell-shaped distributions justify use of empirical rule.

Mechanics

(g) Both are right skewed, more so for Total Assets.

(h) Wal-Mart.

(i) More bell-shaped, with a smattering of outliers.

(j) No. Wal-Mart has a typical return on assets (about 9%).

Message

(k) Bell-shaped with mean 4% (median 5%) and SD 10%. Half are in the range 0.7% to 8.8%.

(l) Yes, but not outstanding. The z-score is $(0.10 - 0.038)/0.098 \approx 0.6$. The return on assets is larger than the upper quartile, but not an outlier.

64. M Credit Scores

Motivation

(a) The average ignores variation in the data and is affected by outliers.

Method

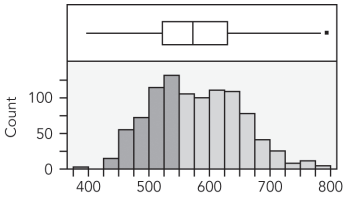
(b) Measures of variation are most important.

Mechanics

(c) Below.

Message

(d) There is considerable variation in the credit scores of these borrowers. The new threshold would exclude 381 out of the 963 past loans, almost 40% of the borrowers. The lender may not want to exclude so many customers.



Quantiles

100.0%	maximum	795
75.0%	quartile	630
50.0%	median	573
25.0%	quartile	522
0.0%	minimum	397

Chapter 5: Association between Categorical Variables

Mix and Match

1. j

2. f

3. e

4. g

5. a

6. h

7. i

8. c

9. d

10. b

True/False

11. True, at least in principle. If the percentages in the table match those in the margins, there's little association. The counts might not match perfectly due to rounding, but they should be very close.

12. False. We have to see the cells of the table, not just the marginal counts. Bar charts show only the marginal distribution, not the presence of association.

13. True.

14. False. Chi-square might be large just because of the size of the table and the number of observations.

15. True.

16. True.

17. False. The value of chi-square is the same if the rows and columns are reversed.

18. False. Association does not imply causation.

19. False. Association is not the same as cause and effect. We cannot interpret association as causation because of the possible presence of a lurking variable. Some managers may operate under very different conditions.

20. False. You can only guess the presence of a lurking variable if you know the context of the problem and suspect that the initial table hides a lurking variable.

21. True.

22. False. It is not associated with defective items.

Think about It

23. (a) Finding association would mean that employed and retired respondents found different rates of satisfaction in resolving the disputed charge.
(b) The two variables are not associated. Roughly equal proportions of both groups of callers (about 70%, 697/1,000 and 715/1,000) were satisfied with the outcome of their call.

24. (a) Finding association would mean that the proportion of loans repaid was different for smaller loans than for larger loans.
(b) Yes. The rate of defaults is much higher among large loans (10/50 = 20%) than for the smaller loans (20/950 ≈ 2%).

25. A lack of association makes this division much simpler, because then the choice of the best color does not affect the packaging. If the two are associated, then the preferred style of packaging depends on the color. In this case, the two divisions would have to coordinate their efforts more closely.

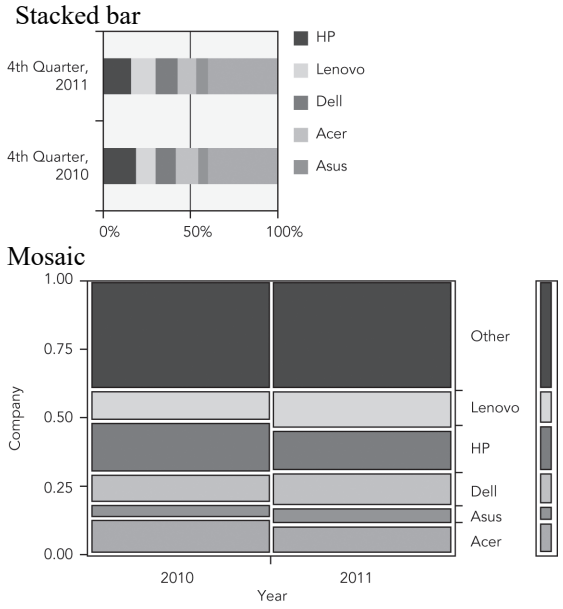
26. They hope to find association and that the association implies that customers who saw the ads choose the brand featured in the ad more often than those who did not see the ads.

27. (a) The administration group. The red segment is largest for this group.
(b) The variables are associated because the composition of the segments differs for the three groups. In particular, we can see that family issues are a more prominent reason for missing work in the administration group than the other two.

28. (a) The variables are associated, primarily because of the increase in the share of complaints about delivery received on Friday. It would appear that customers hoped for a package to arrive before the weekend.
(b) Because the chart shows percentages, we cannot tell the number of complaints that arrive each day. A manager cannot tell, for example, whether there are a lot of issues with deliveries on Friday or these just make up a large share of a relatively small number of complaints.

29. (a) Asia Pacific. This is the widest column in the plot (about 29% of the total).
(b) Latin America and Middle East/Africa. The red share of these is more than half. (Use the scale at the left of the plot).
(c) 80% of the Asia/Pacific market would be larger.
(d) No. The figure emphasizes the share of brand within the region, not the other way around. (it's about one-quarter).
(e) Yes. The conditional distribution depends on the location. One manufacturer dominates in one region (Japan Tobacco in Asia Pacific) whereas another is dominant elsewhere (Phillip Morris in North America).

30. (a) The two plots convey small changes in market share overall, with HP losing share but remaining the leading vendor.



(b) The mosaic plot shows that total shipments decreased from 2010 to 2011; this aspect is subtle because the drop is only about 1.4 percent (from 93.45 to 92.17 million).

31. If the choice of color is not associated with the style of the vehicle, then yes. Otherwise, it may be the case that the choice of color depends on the type of vehicle.

32. The table will show association unless none of the shoppers who were sent a coupon makes a purchase. The table has the following form:

	Used coupon	Did not use coupon
Sent coupon	*a*	*b*
Did not send coupon	0	*c*

The lower left-hand cell of the table must be zero. Unless the number of used coupons is also zero, the conditional distribution in the second row will not match the marginal distribution and the two variables are associated. Structural zeros in a contingency table (here, it's not possible to use a coupon if you do not have one) complicate the analysis of tables because they force a type of artificial association.

33. These are most likely associated. Children are less likely to be in stores at night.

34. Yes. S&*P* would be bankrupt itself if its ratings were not associated with the presence or absence of default. AAA bonds are much like U.S. Treasuries with no chance (or very little) of default. CCC bonds historically have about a 20% annual default rate.

35. Association is not the same as causation. It could be the case, as the eminent statistician R. A. Fisher argued, that both smoking and cancer share a common cause such as some underlying genetic characteristic that causes cancer and also produces a desire to smoke. His objections have not stood the test of time, and most now accept the relationship between smoking and cancer as causal.

36. No. We know that a wide variety of evolving viruses cause the flu. The weather is associated with the flu because the virus finds it easier to move from one person to the next in colder weather when people tend to spend more time in crowded spaces with others.

37. (a) $V = 0$. The rows are proportional, and the two variables are not associated with each other.
(b) $V = 0$. Because it is designed to measure the association between categorical variables (which do not define an ordering), Cramer's V does not depend on the order or arrangement of the rows in the table. You can rearrange the rows and columns; the value of V remains the same.
(c) The lack of association means that when ordering paints, the store should order the same fraction of gloss in each color (2/7 low gloss, 1/7 medium, and 4/7 high gloss).

38. (a) $V = 1$. The two variables are redundant; once you know the choice of color, you know the finish (and vice versa).
(b) $V = 1$. Cramer's V does not depend on the order or arrangement of the rows in the table.
(c) Unlike the previous question, the gloss type is determined by the color choice. The store would probably order very little low gloss red and very little high gloss blue.

You Do It

39. (a) Type of Day By Grade of Gasoline

Count	Premium	Plus	Regular	Total
Weekday	126	103	448	677
Weekend	63	29	115	207
Total	189	132	563	884

(b) Among weekday purchases, 126/677 = 19% are for premium, 103/677 = 15% are for plus, and the remaining 66% are for regular.
(c) 126/189=67% of premium purchases are on weekdays, and 33% on weekends.
(d) No. These conditional distributions are not directly comparable. One refers to the type of gas given that the purchase happens on the weekday, whereas the second refers to the timing of the purchase given that premium gas was bought. To illustrate association, one can, for example, compare the conditional distribution of purchases of regular gas to the answer in part c. For regular purchases 448/563 = 80% occur on weekdays, compared to 67% of premium purchases.
(e) Weekends. Though more premium gas is sold during the week, there's a greater concentration of premium sales on weekend days.

40. (a) The marginal counts are
Size By Style

Count	Beach	Button-down	Polo	Total
Large	22	103	65	190
Medium	28	65	82	175
Small	36	18	27	81
Total	86	186	174	446

The marginal distribution of the styles is 186/446 button down, 174/446 polo and 86/446 beach prints. The marginal distribution of sizes is 81/446 small, 175/446 medium, and 190/446 large.
(b) Within the polo column, the conditional distribution is 27/174 = 16% small, 82/174=47% medium and 65/174 = 37% large.
(c) Within the row for large shirts, the proportions are 103/190 = 54% button down, 65/190 = 34% polo, and 22/190 = 12% print.

(d) The table shows association. The manager should evidently stock larger sizes in the button-down shirts and smaller sizes in the beach prints. Polos are most popular in the middle size.

41. (a) The expected counts for the table are

(189·677)/884=144.74 (189·207)/884=44.26
(132·677)/884=101.09 (132·207)/884=30.91
(563·677)/884=431.17 (563·207)/884=131.83

and from these we arrive at the overall chi-square for the table:

$\chi^2 = 13.32$ and $V = \sqrt{\frac{13.32}{884 \cdot 1}} = 0.12$.

(b) V is rather small, indicating weak association between the type of gas and the timing of the purchase. There is some association (regular tends to be bought more during the week than premium), but the association is not strong.

42. (a) The expected frequencies under lack of association from the margins are

(81·186)/446=33.78 (81·174)/446=31.60 (81·86)/446=15.62
(175·186)/446=72.98 (175·174)/446=68.27 (175·86)/446=33.74
(190·186)/446=79.24 (190·174)/446=74.13 (190·86)/446=36.64

These produce the following contributions to chi-square:

$\frac{(18-33.78)^2}{33.78} = 7.37$ $\frac{(27-31.60)^2}{31.60} = 0.67$ $\frac{(36-15.62)^2}{15.62} = 26.60$
$\frac{(65-72.98)^2}{72.98} = 0.87$ $\frac{(82-68.27)^2}{68.27} = 2.76$ $\frac{(28-33.74)^2}{33.74} = 0.98$
$\frac{(103-79.24)^2}{79.24} = 7.12$ $\frac{(65-74.13)^2}{74.13} = 1.12$ $\frac{(22-36.64)^2}{36.64} = 5.85$

Adding these up gives $\chi^2 = 53.34$, and $V = \sqrt{\frac{53.34}{446 \cdot 2}} = 0.24$.

(b) The association is weak but nonetheless noticeable. The largest share of prints is sold in smaller sizes, whereas the largest share of button-downs are in larger sizes.

43. he expanded table is as follows with the marginal totals:
Question Wording By Satisfaction

Count	Very dissatisfied	Somewhat dissatisfied	Somewhat satisfied	Very satisfied	Total
Dissatisfied	23	20	69	128	240
Satisfied	10	12	82	139	243
total	33	32	151	267	483

	Question uses *satisfied*	**Question uses *dissatisfied***	**Total**
Very satisfied	139	128	267
Somewhat satisfied	82	69	151
Somewhat dissatisfied	12	20	32
Very dissatisfied	10	23	33
Total	243	240	483

(a) The table shows the marginal counts.
(b) Combine the first two rows of the table. If the question used the word satisfied, (139+82)/243 = 90.95% were satisfied. If the question used the word dissatisfied, then the percentage satisfied dropped to (128+69)/240 = 82.08%.
(c) Phrase the question in a positive sense using the word satisfied. Wording puts the notion of being satisfied in the customers mind rather than encouraging the customer to think of reasons that he or she might not be satisfied.

44. (a) This table adds the marginal counts.

	Stockholders	**Non stockholders**	**Total**
Very likely	18	26	44
Somewhat likely	41	65	106
Not very likely	52	68	120
Not likely at all	19	31	50
Unsure	8	13	21
Total	138	203	341

(b) (18+41)/138 = 42.75% of stockholders anticipated another drop.
(c) Among non-stockholders, the percentage expecting another drop is slightly higher, (26+65)/203 = 44.83%
(d) We might have expected the direction of this effect, with those who own stock being more optimistic about the potential of the market. The difference in this poll is quite small, however.

45. (a) The table shows weak association. The expected counts from the margins are:

$(267\cdot243)/483=134.33$ $(267\cdot240)/483=132.67$
$(151\cdot243)/483=75.97$ $(151\cdot240)/483=75.03$
$(32\cdot243)/483=16.10$ $(32\cdot240)/483=15.90$
$(33\cdot243)/483=16.60$ $(33\cdot240)/483=16.40$

Accumulating the squared deviations divided by these counts, we obtain

$$\chi^2 = \frac{(139-134.33)^2}{134.33} + \ldots + \frac{(23-16.40)^2}{16.40} = 8.67 \text{ and } V = \sqrt{\frac{8.67}{483\cdot 1}} = 0.13.$$

The association is weak.

(b) The combined table with marginal totals is:

	Question uses *satisfied*	**Question uses *dissatisfied***	**Total**
Satisfied	221	197	418
Dissatisfied	22	43	65
Total	243	240	483

The expected counts are

(418·243)/483=210.30 (418·240)/483=207.70
(65·243)/483=32.70 (65·240)/483=32.30

and $\chi^2 = \frac{(221-210.30)^2}{210.30} + \ldots + \frac{(43-32.30)^2}{32.30} = 8.14$ with $V = \sqrt{\frac{8.14}{483 \cdot 1}} = 0.13$. Chi-square is slightly smaller, but V is the same.

46. (a) The expected counts are:

(44·138)/341=17.81 (44·203)/341=26.19
(106·138)/341=42.90 (106·203)/341=63.10
(120·138)/341=48.56 (120·203)/341=71.44
(50·138)/341=20.23 (50·203)/341=29.77
(21·138)/341=8.50 (21·203)/341=12.50

Adding the squared deviations after dividing by these expected counts gives $\chi^2 = \frac{(18-17.81)^2}{17.81} + \ldots \frac{(13-12.50)^2}{12.50} = 0.73$ and $V = \sqrt{\frac{0.73}{341 \cdot 1}} = 0.05$. There is almost no association between ownership and how they responded to the survey.

(b) The merged table with 3 rows has these counts, along with the shown margins.

	Stockholders	**Non Stockholders**	**Total**
Likely	59	91	150
Unlikely	71	99	170
Unsure	8	13	21
Total	138	203	341

The expected counts from the margins are:

(150·138)/341=60.70 (150·203)/341=89.30
(170·138)/341=68.80 (170·203)/341=101.20
(21·138)/341=8.50 (21·203)/341=12.50

The resulting value of χ^2 is $\frac{(59-60.70)^2}{60.70} + \ldots \frac{(13-12.50)^2}{12.50} = 0.25$, smaller than the already small value in part a. $V = \sqrt{\frac{0.25}{341 \cdot 1}} = 0.03$, also smaller than the prior value.

47. (a) Marginally, given a balanced number of cases in each industry, the marginal proportion of men is 43% ((34%+40%+38%+60%)/4=43%). None of the industries has this proportion, and so each conditional distribution does not match the marginal distribution. There is association.

(b) The table shows association because the proportion of men working in each industry depends on the industry. Some industries have a higher proportion of male employees than female employees.

(c) If n = 400 with 100 in each row, then the expected counts are

	Men	**Women**
Advertising	43	57
Book publishing	43	57
Law firms	43	57
Investment banking	43	57

Accumulating the squared deviations divided by these values, the overall $\chi^2 = \frac{(34-43)^2}{43} + \ldots \frac{(40-57)^2}{57} = 16.5$ and Cramer's $V = \sqrt{\frac{16.5}{400 \cdot 1}} = 0.20$. If n = 1,600 with 400 in each row, then the expected counts are four times larger. Chi-square is also four times larger, but V is unchanged. Chi-square increases with n, but V does not.

48. (a) Yes. The percentages that default differ among the rows.
(b) If the score were not associated with default, then the score would not be useful as a predictor of whether the customer would repay the loan. Stores could not use the score to gauge the risk of making a loan.
(c) The association is weak because there are very few cases with differing percentages. The cell counts based on the indicated numbers in each row are given in the following table.

	Defaulted	Repaid	Total
Risky	15	35	50
Uncertain	22	78	100
Acceptable	180	8,820	9,000
Perfect	17	833	850
Total	234	9,766	10,000

The expected cell counts are then:

(50·234)/10,000=1.17	(50·9,766)/10,000=48.83
(100·234)/10,000=2.34	(100·9.766)/10,000=97.66
(9,000·234)/10,000=210.60	(9,000·9.766)/10,000=8,789.40
(850·234)/10,000=19.89	(850·9.766)/10,000=830.11

Based on these counts, $\chi^2 = \frac{(15-1.17)^2}{1.17} + \ldots + \frac{(833-830.11)^2}{830.11} = 341.53$ and Cramer's $V = \sqrt{\frac{341.53}{10{,}000 \cdot 1}} = 0.18$.
(d) The association would be stronger were there more cases in the higher risk rows of the table. These are the rows with differing proportions. For example, moving just 50 cases from the acceptable to the risky row increases Cramer's V to 0.22.

49. (a) The shown proportions differ in the rows of the table, but this is not the type of association that we have studied in this chapter. See part b.
(b) This table is *not* a contingency table. The 1,527 respondents are not put into cells on the basis of two measurements for each. Think of the how the data table is shaped. For each of the 1,527 rows (the survey respondents), we have five columns. One column is the rating assigned to banks, the second is the rating assigned to big business, and so forth. Each respondent has a value in each row of the table. The cells are not mutually exclusive.

50.
(a) Differences in the shown column percentages indicate association. The association shows that investors in the US have a might higher preference for the US market than investors outside the US.
(b) No. These are not counts.
(c) Cannot compute chi-squared without column marginal counts.

51. (a) The results show clear association, with the appearance that support influences the outcome. Of those that get support, 24 out of 47 (51%) produced a supportive article (percentages within the first row). None of those who were unsupported wrote a favorable paper.
(b) The association is so strong as to make it difficult to find a lurking variable that would mitigate what is shown in the table. That said, one could suggest that the authors were already in the progress of their research prior to funding. Companies only fund results that were headed their way. The support might have come after the research, rather than before.

52. Reaction to cell phone

	Male	Female	Total
Favorable	36	18	54
Ambivalent	42	7	49
Unfavorable	29	9	38
Total	107	34	141

(a) Yes. Only 36/107 = 33.64% of men were favorable, compared to 18/34 = 52.94% of women. Though fewer women were favorable, fewer women were asked.
(b) First find out why so few women participated in the survey. Then, assuming the reactions are representative, market toward women first.
(c) A lurking variable may be some factor that affects why more men than women participated in the survey, such as the proportion that carry a phone or the time of day or day of the week of the mall survey. Perhaps even the behavior of the person collecting the data; this individual might cause different reactions in men and women.

53. (a) Use the column percentages. American is doing better; its on-time arrival rate overall is 3525/(3525+783) ≈ 0.818 whereas US Airways is 4,083/(4,083+1,002) ≈ 0.803.
(b) Yes. The on-time percentage at Los Angeles is 2,670/(2,670+536) ≈ 0.833 whereas it's 3,000/(3,000+835) ≈ 0.782 in Philadelphia. This difference matters because most of the American flights go to Los Angeles, whereas US Airways flies to Philadelphia.
(c) Yes. If we compare the arrival rates at each destination, US Airways comes out better: for example, 2,188/2633≈83.1% versus 482/573≈83.4% in Los Angeles and 206/283≈72.8% versus 2,794/3,552≈78.7% in Philadelphia.

54. (a) Use the column percentages. American is doing better; its on-time arrival rate overall is 1,536/(1,536+416) ≈ 0.787 whereas Delta's is very slightly smaller at 11,769/(11,769+3,343)≈ 0.779.
(b) Yes. The on-time percentage at Atlanta is 11,512/(11,512+3,334) ≈ 0.775 whereas it's 1,007/(1,007+244) ≈ 0.805 in Las Vegas. This difference matters because a higher percentage of American flights go to Las Vegas, whereas Delta flies most often to its hub in Atlanta.
(c) Yes. If we compare the arrival rates at each destination, Delta comes out better: 77.6% versus 76.1% in Atlanta, 80.7% versus 80.4% in Las Vegas, and 84.2% versus 79.5% in San Diego.

4M Discrimination in Hiring

Motivation

(a) *Employees.* The employees would expect to find association between age and action, namely that the company was laying off a higher proportion of older employees.

(b) *Company*. The data for a company that does not discriminate on the basis of age would show that the two variables are not associated. Alternatively, the company might argue that a lurking factor (ability) undermines this analysis. If indeed the company did lay off older employees, then it might argue that these workers were less capable of doing the necessary work.

Method

(c) The percentages within the rows would be more useful in order to convey (at least to most observers) the presence of discrimination. If the percentage laid off rises steadily with age, then there would be some evidence that the policy of the firm was more harmful to older employees.

(d) Cramer's *V* is more useful than chi-square alone because Cramer's *V* has the more interpretable range from 0 to 1. It summarizes in a single value the differences in rates of layoffs within the rows of the table.

Mechanics

(e) The expected cell counts are as follows

(805·68)/1968 ≈ 27.82	(805·1900)/1968 ≈ 777.19
(646·68)/1968 ≈ 22.32	(646·1900)/1968 ≈ 623.68
(392·68)/1968 ≈13.54	(392·1900)/1968 ≈ 378.46
(125·68)/1968 ≈ 4.32	(125·1900)/1968 ≈ 120.68

The following table shows the row percentages as well as the contributions to chi-square from each cell of the original table.

Age	Laid Off	Retained	Percentage Laid Off
<40	$\frac{(18-27.82)^2}{27.82} \approx 3.46$	$\frac{(787-777.19)^2}{777.19} \approx 0.12$	0.022
40-49	$\frac{(14-22.32)^2}{22.32} \approx 3.10$	$\frac{(632-623.68)^2}{623.68} \approx 0.11$	0.022
50-59	$\frac{(18-13.54)^2}{13.54} \approx 1.47$	$\frac{(374-378.46)^2}{378.46} \approx 0.05$	0.046
60 or more	$\frac{(18-4.32)^2}{4.32} \approx 43.33$	$\frac{(107-120.68)^2}{120.68} \approx 1.55$	0.144
		Overall	0.035

From the entries in the table, we calculate chi-square to be 3.46+...+1.55 = 53.2 and Cramer's $V = \sqrt{\frac{53.2}{1968 \cdot 1}} \approx 0.164$. Cramer's V shows that the association is weak. The row percentages show you why. The percentage laid off is mostly confined to those of age 60 or more. Notice that the major contribution to chi-square is from the cell of those laid off who were 60 or more. Otherwise the rates are relatively similar to the overall rate (about 3.5%).

Message

(f) The data suggest that the layoffs discriminate against older employees. The data show weak association between age and whether an employee was laid off. The overall rate of layoffs is about 3.5%. The rate is less than this for employees who are younger than 50. The rate is higher than this for those employees who are older. Among employees who are 60 or older, the rate climbs to more than 14%.

(g) The presence of a lurking factor could mean that the association found in this table is not due to age, but rather due to some other factor that is related to age, such as the ability of the employee to do the work. For example, if the older employees are not able to use, say, a new type of computer-controlled machine, the company might argue that the layoffs had nothing to do with age and were motivated by ability instead. The cited reference has further details in this context.

4M Picking a Hospital.

Motivation

(a) More information is good so long as the patients are able to make sense of the information. Too much information confuses some and does not help them make the best choice. If the two variables (*Stage* and *Hospital*) are not associated, there would be little to gain from showing the outcomes for both types of cancers. And, by combining the two columns, we might get a better sense of the overall performance of the hospital. In other words, show the whole table if there is association, but limit the information to the margins if there isn't any.

Method

(b) She needs to see the conditional probabilities based on her diagnosis. The marginal probability combines the outcomes across both stages. If the variables are associated, that can be misleading.

Mechanics

(c) These are simply the ratios based on the outcomes in the table. In this summary, we've rounded the percentages.

	Early Stage	**Late Stage**	
CH	10%	57%	29%
UH	5%	44%	37%

(d) Late stage cases make up 21/51 = 41.2% at CH but 405/503 = 80.5% at UH.

(e) The university hospital has a lower death rate for either type.

(f) The community hospital.

Message

(g) Go to the university hospital regardless of the type of cancer. Even though the university hospital has a higher death rate overall, this is an artifact of its patient mix. The university hospital treats many more late-stage cancers than the community hospital, and hence it has a higher marginal rate of deaths per surgery (37% versus 29%).

(h) The marginal information is not adequate. Unless the data is accompanied by information on the mix of patients, one cannot judge the quality of care. (Indeed, most ratings that are used to publicly compare hospitals use a variety of statistical adjustments to make the patient mix comparable.)

Chapter 6: Association between Quantitative Variables

Mix and Match

1. (a) i
(b) ii
(c) iii
(d) iv

2. (a) iv
(b) ii (This plot ignores the white space rule.)
(c) i
(d) iii

3. (a) iv
(b) iii
(c) i
(d) ii

4. (a) iii
(b) i
(c) ii
(d) iv

True/False

5. True.

6. False. Those with highest incomes would be found at the top of the plot. They might happen to be located more to the right (positive association), but they need not be there.

7. False, in general. The pattern could also be in a negative direction.

8. True.

9. True.

10. False. The pattern would be nonlinear since the growth in sales slows as the amount of advertising increases. The line would have to flatten out and become a curve.

11. False. The pattern would be linear, with $y \approx 0.1\,x$. where y denotes revenue and x denotes sales.

12. False. These are properties of the marginal distributions. The scatterplot and association are properties of the conditional distribution of y given x (or vice versa).

13. False. The value of the stock would fall along with the economy. We'd rather have one that was negatively related to the overall economy as a hedge against a recession.

14. True. The correlation is the covariance times the product of SDs, and these are both positive.

15. True. Untangle the standardized variables and you'll see that the correlation line in the original units predicts y to be $\bar{y} + r\frac{s_y}{s_x}(x - \bar{x})$. If $r = 1$, the slope is the ratio of SDs.

16. False. The correlation should be near zero unless the phone company has somehow managed to assign high phone numbers to customers living in wealthy areas who charge expensive items.

17. False. The correlation between x and y is the same as the correlation between y and x.

18. True.

19. True. The correlation is not affected by changing the scale.

20. True. The matrix is symmetric.

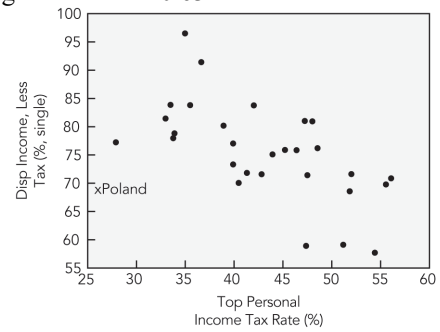
Think About It

21. (a) Sales: Total cost is the response and number of items is the explanatory variable. Expect to see a positive direction, linear, with lots of variation because of the varying costs of items bought.
(b) Productivity: Items produced is the response and hours worked is the explanatory variable. Expect to see positive direction, linear (with perhaps some curvature for long hours), and moderate variation.
(c) Football: Weight is the explanatory variable and time is the response. Expect negative direction (those little guys need to be quick!), probably linear but with lots of variation in the times.
(d) Fuel: Number of miles is the explanatory variable and gallons left is the response. Expect negative direction, linear, with small variation around the trend (assuming we drive similarly after each fill-up).
(e) Investing: The number recommending is the explanatory variable and the subsequent price change is the response. We expect (being skeptical) little or no pattern, but many would expect positive association.

22. (a) Electricity: Temperature is the explanatory variable and kilowatt hour is the response. Expect to find a relationship with a positive direction (higher temperature leads to higher use), lots of variation, and linear.
(b) Long-distance calls: Time is the explanatory variable and cost is the response. Expect positive direction, linear, with very low variation. The correlation would be 1 if we paid a fixed cost per minute of calling time.
(c) Freight: Weight is the explanatory variable and fuel is the response. Positive, linear or perhaps bending (with weight mattering more for smaller planes), with lots of variation due to other factors (such as wind speeds and flight distance).
(d) Advertising: Length is the explanatory variable and number who recall is the response. Expect little or no relationship, but it might be slightly positive. A good ad does not have to be long to be memorable.
(e) Health: Exercise is the explanatory variable and fat is the response. Expect a negative pattern with considerable variation depending on, among other things, sex and fitness level.

23. (a) There's positive association, but it seems rather weak. It's hard to say whether it's linear from the figure; the relationship is too weak.
(b) The actual correlation is 0.38.
(c) The cluster increases the correlation. The correlation without this cluster is about half the size; $r = 0.19$.
(d) No, when the outliers are excluded, the association is too weak to arrive at this conclusion.

24. (a) There is a moderate or weak negative linear association between the two.
(b) That the direction is negative makes sense; higher tax rates leave less to spend.
(c) The correlation is -0.53.
(d) We'd pick Poland, as shown below, but others might differ. The correlation becomes more clearly negative when this case is excluded, falling further to -0.63.



25. The correlation is not affected by changing the scale from either dollars to thousands or dollars to euros.

26. Yes, because the correlation does not depend on the scales of the two variables, but we'd like to see a plot.

27. No, because the correlation does not depend on the center of the data. You can add and subtract a constant from a variable without changing the correlation.

28. Not at all because the correlation does not depend on the center of the data. You can add and subtract a constant without changing the correlation's value.

29. The slope of the linear relationship measured by the correlation is r. Hence, the predicted z score must be smaller than the observed z score because the absolute value of r is less than 1. This phenomenon is often called regression to the mean.

30. The presence of a linear association with $r = 0.4$ means, for example, that we would expect an employee with $z = -1$ (1 SD below the mean on attendance this year) to be only 0.4 SDs below the mean next year. Perhaps the incentive program works, but that would not explain the fall in relative performance for those who did well this year ($z = +1$ dropping to $z = +0.4$). (See also the discussion of Exercise 29.)

31. (a) Association is hard to judge from time plots.
(b) Yes, but still no clear pattern.
(c) Weak. ($r = 0.24$)
(d) A scatterplot is more useful for judging correlation, but the scatterplot hides the timing of the data.
(e) No. Association is not causation.

32. (a) Association is present, but not consistently positive or negative. Sentiment has been lowest with inflation both very high and very low.
(b) Yes. In the past, sentiment improved when inflation was low (negative association).

(c) More negative. (overall, $r = -0.47$)

33. The correlation is larger among stones of the same cuts, colors and clarities. These are the other factors that add variation around the correlation line. By forcing these to be the same, we get a more consistent pattern with fewer lurking variables.

34. This averaging has a great impact on the correlation. Assume that the scatterplot that shows the daily values has a positive, linear pattern, with the average number produced on the *y* axis and the number of employees on the *x* axis. By averaging, the manager removes the effects of some lurking factors and finds a larger correlation than would be found with the daily data. Beware of correlations computed with aggregated data.

35. Cramer's *V* measures association between *categorical* variables. Because the levels of categorical variables cannot in general be ordered, it does not make sense to speak of the *direction* of the association.

36. No. Cramer's *V* detects *any* association, not just linear association. Because *V* measures association between categorical variables, it would not make sense to measure linear association. The levels of the categorical variables cannot in general be ordered or drawn as though on a line.

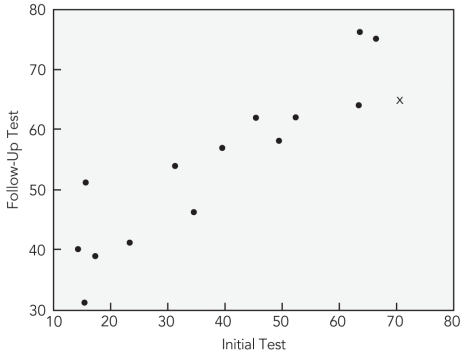
37. You have a 1-in-4 chance of guessing the original. That's not a pattern that you've found; that's luck.

38. No, because the marginal histograms are all the same. Scrambling the data alters the pairing of the *x* and *y* coordinates, but does not change the marginal distributions.

39. (a) Most would expect moderate association, but answers vary.
(b) The data show no linear association.
(c) $r = -0.03$. The data have no linear association.
(d) No. The line would be flat or horizontal.

40. (a) Expected negative association: lower winter temperature suggests more snow and need for four-wheel drive.
(b) Strong linear association, with more variation in states with lower temperature.
(c) Highest in Wyoming and N Dakota; lowest in Florida.
(d) $r = -0.85$.

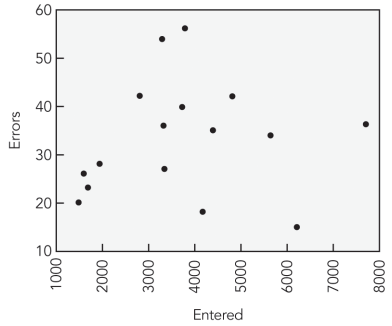
41. (a) Yes. Assuming that the employees differ in skill.
(b) The scatterplot shows strong, positive, linear association. Employees who scored well the first time tend to score relatively well the second time.



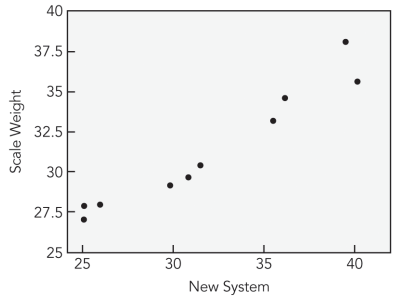
(c) The correlation is $r = 0.903$. The correlation is suitable because the association is linear. We'd expect the relative position on the second test to be lower, closer to 1.8 (twice the correlation) on the second test.
(d) The employee in question is marked by the x in the figure. The correlation line, with slope 0.9 implies that we expect the scores on the second test to be a little closer to the mean than the scores on the first test. The decline for this employee seems consistent with the pattern among the others and it would be inappropriate to judge this employee as becoming less productive.

42. (a) Expect some association, but the races very different in length and separated by almost of a month of racing.
(b) The association is moderate, with the winning times (smallest, Cancellara and Martin) being outliers.
(c) $r = 0.48$
(d) Not exceptional (6^{th} place in the first race but 35^{th} in the second).

43. (a) The scatterplot is below, with Entered as the explanatory variable and Errors as the response. We think of errors as the result of entering data rather than the other way around. There's little or no pattern in these data.
(b) 0.094.
(c) The correlation does not depend on the units of the data and would not change.
(d) The correlation indicates a very weak association between the number of data values entered and the number of errors. Evidently, those who enter a lot of values are simply faster and more accurate than those who enter fewer. (They are more accurate because they have entered more items but kept about the same number of errors.)
(e) There's virtually no association between the number entered and the number of errors.



44. (a) The scatterplot below shows the scale measurements on the x axis and the weights from the new system on the y axis. The plot shows a strong pattern with little variation around the positive, mostly linear trend. There may also be some nonlinearity, with the pattern getting steeper as the scale weight increases. With so few cases, it is hard to be confident in the accuracy.

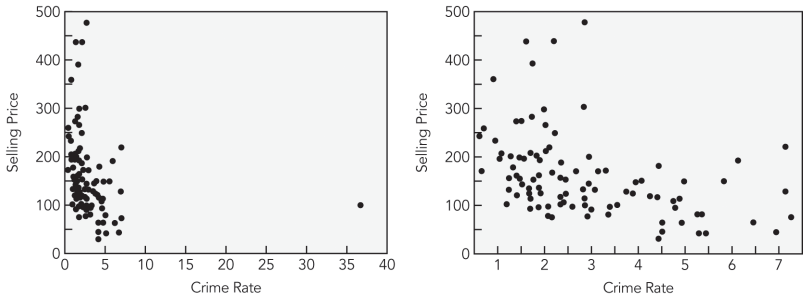


(b) $r = 0.965$.
(c) No. The correlation is not affected by the scale of the data.
(d) The correlation tells that the two measurements are very highly related, but not a 1-to-1 match. Officials will have to decide if the new system is accurate enough for checking truck weights.

45. (a) Horsepower on the *x*-axis with combined MPG on y-axis.
(b) Moderate negative association with some curvature and outliers such as the Honda CR-Z.
(c) $r = -0.80$
(d) Yes it conveys moderate association, but does not measure the curvature.
(e) Estimated Combined MPG ≈ 32.4 – 0.31 * 200 ≈ 26.2 MPG. Seems reasonable for cars with 200 HP, but omits curvature.

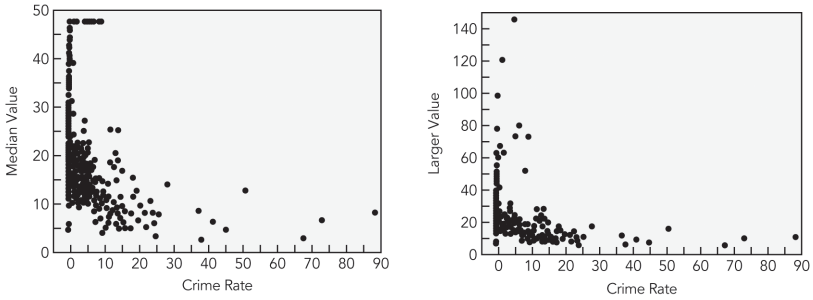
46. (a) Weight on the *x*-axis with city MPG on the y-axis.
(b) Moderate negative association, with evidence of curvature at higher weights.
(c) $r = -0.78$
(d) Estimated City MPG ≈ 37.2 − 0.0040*4000 ≈ 21.2 MPG.

47. (a) The plot on the left below plots the price on the crime rate. The outlier is Center City, Philadelphia, and it is unusual in terms of the crime rate, but not the selling price.



(b) The correlation using all of the data is $r = -0.25$.
(c) The refocused scatterplot on the right shows a great deal of variation around a weak, negative trend that appears to bend. The price seems to drop off faster on the left (few crimes) than the right (more crimes), either that or there are new outliers (such as the cluster of expensive districts at the upper left).
(d) The correlation without Center City, Philadelphia, is much stronger than previously found, −0.43.
(e) No, we cannot for several reasons. First, this is aggregated data. We do not see the prices for individual homes, only for communities. Second, correlation measures association, not causation.

48. (a) The scatterplot on the left below shows all of the data. Notice the group of points in the upper left corner as well as the number of values piling up near zero crime rate. The cluster of values at 50 ($50,000) occur because of censoring; the census in 1970 truncated housing values in tracts at $50,000. These would be higher.



(b) The correlation is $r = -0.39$.
(c) The extended values include a few outliers values in the upper left corner of the scatterplot on the right. These outliers suggest a less linear relationship, and the calculated value of the correlation moves closer to zero, $r = -0.28$.
(d) Areas with very large crime rates have rather low housing values, but the relationship is not shown to be causal because correlation does not imply causation. There is clearly association, however, as the areas with larger home values all have relatively less crime.

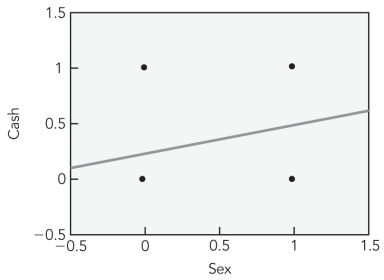
49. (a) The value of chi-square is 8.466. The table below shows the counts, expected count (under no association), and the contribution of that cell to chi-square. Cramer's V is then sqrt(8.466/150) ≈ 0.23757.

Count Expected Cell χ^2	Male	Female	Total
Cash	50 42 1.5238	10 18 3.5556	60
Credit	55 63 1.0159	35 27 2.3704	90
Total	105	45	150

(b) The correlation is approximately 0.24. The following table shows the details.
(c) The two are exactly the same, though that will not be the case when the correlation is negative. Cramer's V is always positive, whereas the correlation picks up the direction. The squared correlation is always the same as the squared value of V.

Variable	**Mean**	**Std. Dev.**	**Correlation**
Sex	0.7	0.459793	0.237566
Cash	0.4	0.491539	

50. a) The scatterplot (shown here with the correlation line added) consists of just four points. Each point represents a cell in the prior contingency table. We cannot see the volume of data because of overprinting.

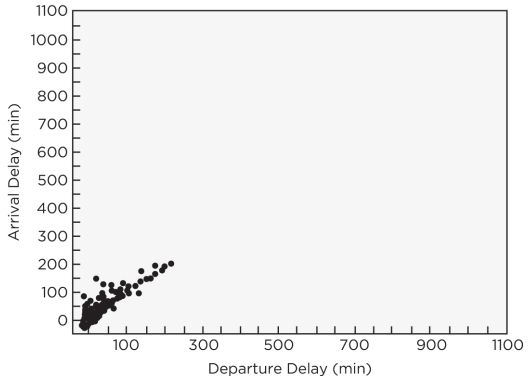


(b) A very useful improvement is to make the size of the point vary with the amount of overprinting. Alternatively, you can add a little random variation around the points to have the same effect.
(c) Two points determine a line, and you can't find anything other than linear association with two dummy variables. Hence, Cramer's V and correlation are both measuring any type of association in this situation.

51. This exercise and the next show that macroeconomic time series are often highly correlated with each other because they all measure aspects of a growing economy.
(a) Yes, both grow steadily.
(b) Yes, shows very strong positive correlation. The data occupy little of the content of the plot, so we may miss subtle features of the data.
(c) $r = 0.99$
(d) No, association is not causation. Both have grown.

52. (a) Yes, but the patterns of growth differ; also debt fell in later quarters.
(b) Strong association, but white space suggests plot is not concentrating on showing variation. Care is warranted.
(c) $r = 0.98$. The series are strongly associated, but the association is not along a single line.
(d) No, association is not causation.

53. (a) Yes, although one might hope that there would be small correlation, anticipating that the pilots would make up for delays along the way.
(b) The scatterplot shows strong positive linear association, with one pronounced outlier (row 437, a flight on Northwest from Billings, MT to Minneapolis).



(c) The correlation is 0.958.
(d) The correlation is noticeably smaller without the outlier. It's still rather positive, but has fallen to 0.907.
(e) The correlation would be the same. Correlation has no units, and so it is the same regardless of the time units.

54. (a) Use CO2 for the *x* axis.
(b) Difficult to see because of outliers (white space rule)
(c) $r = 0.73$
(d) Answers will differ, but China, U.S., and Japan are clear outliers, and probably the Russian Federation and India as well. Without these 5, $r = 0.83$.

55. 4M. Correlation in the Stock Market

Motivation

(a) If the stocks are strongly related, then if one goes up, they all go up. That would be fine, but if one goes down, then they all also go down. Consequently, owning several highly related investments is just like having put all of your money into one of them.

(b) They'd ideally like to find negatively associated investments. Then if one went down, another would go up. (This is the motivation behind hedging or diversifying one's investments.) As in this example, most stocks tend to be positively associated, making this strategy require other approaches.

Method

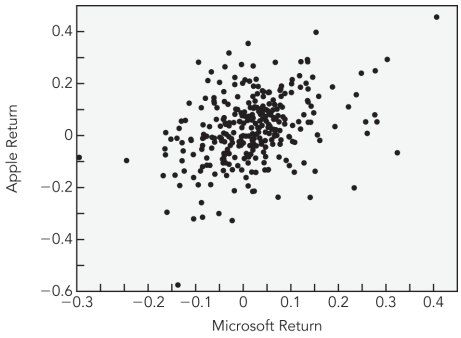
(c) Six possible pairs.

(d) To check for linear relationships and outliers, the investor needs to see scatterplots.

(e) If the returns are not related to time, then time cannot influence the relationship between the returns. Plots of the returns over time will answer such questions.

Mechanics

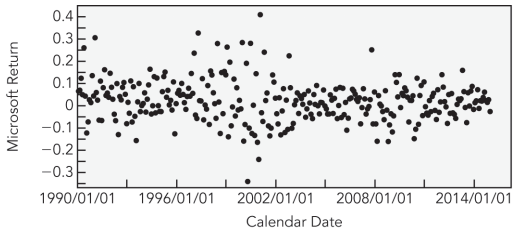
(f) No, order is irrelevant. The returns are weakly to moderately associated, with vaguely linear patterns. For example, this scatterplot shows returns on Apple and Microsoft.



(g) The correlation matrix is

	Apple Return	HP Return	IBM Return	Microsoft Return
Apple Return	1.0000	0.3740	0.3400	0.3710
HP Return	0.3740	1.0000	0.4540	0.4090
IBM Return	0.3400	0.4540	1.0000	0.4490
Microsoft Return	0.37100	0.4090	0.4490	1.0000

(h) Timeplot is



Message

(i) The returns on stock in these three companies are positively associated, with considerable variation around the pattern. The returns tend to move up and down together, but not in lock step. The association among these returns indicates that they all tend to rise or fall together, offering less diversification than an investor might like.

56. 4M. Cost Accounting

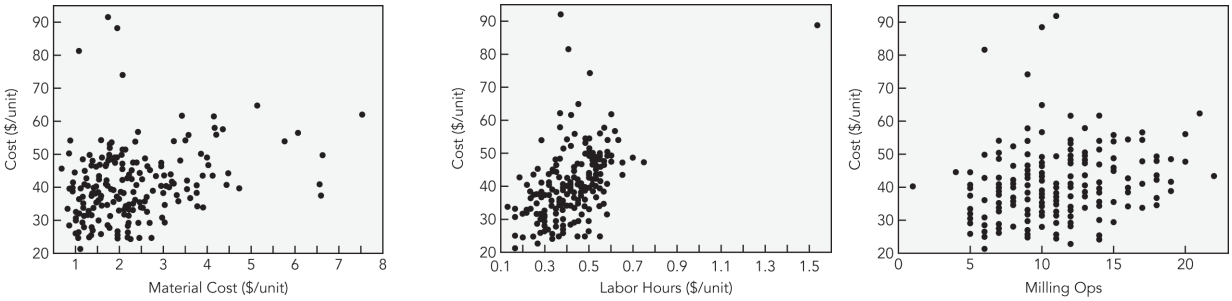
(a) If the company can identify properties of the order that are highly associated (either positively or negatively), it can use these to improve the informal method used to estimate the price of the order. Better, more accurate prices would then make it possible to offer competitive prices with less chance of losing money on a bid.

(b) The response is the cost per unit. That's what we'd like to be able to estimate, and we naturally think of this final cost as the result of inputs to the manufacture such as material costs, labor costs, and machining costs.

(c) If the associations are linear, correlation can identify which of the inputs is most associated with the final cost per unit. Correlations that are very positive indicate factors associated with increasing costs, whereas those with negative correlation indicate factors that are associated with decreasing costs. Management wants to know the magnitude and the direction of the effect.

(d) Scatterplots are needed to verify the checklist for the correlation, so that we can avoid missing nonlinear patterns or the effects of outliers.

(e) The scatterplots are given below. The strongest linear association appears to be between the final cost and the number of labor hours. The outlier will make this correlation even larger. Note the presence of several other outliers; four tasks appear to produce very expensive blocks. For one of these, we can attribute the expense to the large number of labor hours. These variables do not explain the high costs of others.



(f) The correlation matrix for these data is given next. The highest correlation (as seen in the scatterplots) is between the response and labor hours (0.5086).

	Cost ($/unit)	**Material Cost**	**Labor Hours**	**Milling Operations**
Cost ($/unit)	1.0000	0.2563	0.5086	0.1801
Material Cost	0.2563	1.0000	0.0996	0.3400
Labor Hours	0.5086	0.0996	1.0000	0.3245
Milling Operations	0.1801	0.3400	0.3245	1.0000

(g) The four outliers appear to weaken the correlation between the response and material costs, increase the correlation with labor hours, and perhaps weaken the small amount of association between the response and the number of milling operations.

(h) This table repeats the calculation of the correlation without the four outliers. (These are the four runs with highest costs, in rows 19, 94, 106, and 173 of the data table.) The correlation with material costs is noticeably higher (0.3655 compared to 0.2563). The correlation with labor is slightly smaller, and the correlation with milling operations is higher.

	Cost ($/unit)	**Material Cost**	**Labor Hours**	**Milling Operations**
Cost ($/unit)	1.0000	0.3655	0.4960	0.2674
Material Cost	0.3655	1.0000	0.1340	0.3354
Labor Hours	0.4960	0.1340	1.0000	0.4054
Milling Operations	0.2674	0.3354	0.4054	1.0000

(i) The amount of labor is most associated with the response, and the relationship appears linear.

(j) The size of the correlation implies that much of the variation in cost per unit cannot be attributed to labor alone and is due to other factors (such as material cost and milling operations). Management cannot accurately anticipate the cost of an order using just one of these explanatory variables alone.

SIA 1

Pfizer case

1. Stock splits occur at the sudden drops. You could verify this by finding out the number of Pfizer shares being trading on the stock market.
2. (a) Most of the content of the plot is empty; this plot hides variation in the early years.
   (b) No, the data have a strong trend that would be hidden in the histogram.
3. Yes, this is simple variation.
4. Percentage changes capture only most recent variation, whereas cumulative values are on a scale compressed by long-term growth.
5. (a) Bell-shaped.
   (b) A six percent drop is z=(− 6.0 − 1.46)/7.45 ≈ − 1 SDs below the mean. From the empirical rule, the chance is about ½ × 1/3 = 1/6. In the data, 119 such events occur out of 809 months, or 119/809 ≈ 0.147, or about 15%.
   (c) − 1.45 − 2*7.45 = − 16.45%, so the VaR is 1000×.1645 = $164.5. That's the most that could be lost, ruling out the worst 2.5% of occurrences.

Executive Compensation case

1. Verify the distribution is bell-shaped.
2. No. Skewness remains after excluding the extreme values.
3. The correlation is r = 1. Log base e of a number is 2.3 times time log base 10.
4. Closer to the median. Avg $\log_{10}$ revenue ≈ 0.216. 10^0.216 ≈ 1.64. The median total revenue is $1.49 billion and the average is $6.84 billion.

Chapter 7: Probability

Mix and Match

1. j, probabilities multiply for independent events

2. g, disjoint events have no outcomes in common

3. h

4. f

5. b

6. c

7. d, rearranged to have the intersection on the left hand side.

8. i

9. e

10. a. These events are dependent. For independent events, the product of the probabilities matches the probability of the intersection.

True/False

11. False. The sample space consists of all possible sequences of yes and no that he might record. That's $2 \times 2 \times 2 \times 2 \times 2 = 2^5 = 32$ elements in the sample space.

12. False. Independence refers to whether the outcome for one shopper changes the probability of observing another carrying a bag.

13. False. These are not disjoint events. For example, the outcome that all five shoppers have a bag {yes, yes, yes, yes, yes} lies in both **A** and **B**.

14. False. P(**B** *and* **C**) = **P(C**), not P(**B**). The last two shoppers having a bag do not imply that the last three have a bag; however, if the last three have a bag it follows that the last two have a bag.

15. True.

16. False. P(**A** and **C**) = p^5.

17. False, both events could happen. Hence, they are not disjoint. If events are disjoint, only one of them can occur.

18. False, only in the long run does the relative frequency match the probability. Two days of experience is hardly enough to qualify for the long run.

19. False, only if the data lacks patterns does the relative frequency tend to the probability in the long run.

20. True. The ratings must range from 1 to 10. Hence a rating of 11 is not possible.

21. False. The intersection **B** and **C** is a subset of the event **A**. For example, **A** also occurs if 4 of the 6 candidates rating 8 or better come from Monday, with 2 on Tuesday.

22. True. The events **A** and **B** are disjoint, so we can use the Addition Rule for Disjoint Events.

Think About It

23. Graphs a and b. There is a clear pattern in c (up then down). There appears to be a slight pattern in d. (gradually flattening, more variable upward trend). There's no pattern in a or b, although there are outliers in b.

24. Graphs a, b and d. The visual test for association only finds a clear pattern in c. There might be a pattern in a, but the outliers make it hard to be sure. No evident pattern appears in b or d.

25. (a) Intersection: {fresh}
(b) Union: S = {frozen, refrigerated, fresh, deli}
(c) Complement: $\mathbf{A}^C$ = {deli}

26. (a) **W** *and* **V** = {B, BB, BBB}
(b) All but R and D: **W** *or* **V** = {AAA, AA, A, BBB, BB, B, CCC, CC}
(c) $(\mathbf{W}$ *or* $\mathbf{V})^C$ = {R,D}

27. (a) Big (waist) and tall.
(b) This would mean that the choice of waist size is independent of the length of the pant leg. That is, a tall man is just as likely to be very thin as a short man.
(c) (**B** *or* **T**). The intersection (**B** *and* **T**) are pants made for large people, those with a thick waist and long legs. The union includes tall people that are skinny as well as short people with a large waist.

28. (a) Yes. The Ferrari is not in **D**, but it's surely in **C**.
(b) Probably not. If you know **C** happens, then among these brands, it's probably not a domestic model.
(c) Zero, because the intersection is empty.

29. An intersection. The company wants both attributes (engineering, foreign language), not one or the other.

30. Intersection, because the customer must go to this department *and* buy something.

31. a is true, but not a result of the Law of Large Numbers. b is false unless you have to learn every probability from data. Only c is a consequence of the law of large numbers, and it requires that the trials lack a pattern.

32. a and b. The LLN applies regardless of the probability.

33. Not likely. The intensity of traffic would change over the time of day and the day of the week. Also, the outcomes (whether more than 50 cars pass by) would likely be dependent. Imagine what would happen during a traffic jam or very late at night.

34. Perhaps. One would be concerned that flights to some destinations might pack more heavily, such as international flights or flights to vacation spots. Periods when these flights are loading would produce changes in the nature of the data.

35. (a) Yes. The results appear to lack a disruptive pattern. The clicks are relatively few and far between, but there does not seem to be a pattern in where they occur.
(b) No. We can only approximate (or estimate) the chance for clicking on the ad from this sequence (because it lacks a visible pattern), but we'd likely get a different answer if we looked at another 100 tosses. We need to see the whole sequence, continuing infinitely long.

36. No, the data appears to have a strong pattern. Looks like the rainy season occurs in the later part of the data.

37. Go for the 3-point shot. That gives a 30% chance of winning the game. The 2-point strategy gives only a $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ chance of winning the game, assuming that the outcomes are independent so that we can multiply the probabilities.

38. Go for the expanded contract in the initial meeting, assuming that its goal is the expanded contract. There's a 40% chance of approval, compared to only $\frac{3}{4} \times \frac{1}{2} = 3/8 = 0.375$ chance for the two-stage approach. This calculation requires the assumption of independence.

39. Pure speculation, and almost certainly a subjective probability based on this analyst's experience when the market feels like the current situation. Of course, a more quantitative analyst might be using data.

40. Not likely. Some of the best shooters have no memory and keep firing away, knowing that hot and cold streaks are just things that can happen randomly.

41. (a) The Law of Large Numbers only means that the long run proportion matches the probability in the long run. Additionally, if the accidents are dependent, accidents one day rattle the pilots so much that they do not fly as well the next day, then the LLN does not apply because the process does not consist of independent trials.
(b) No. More likely, accidents form a simple sequence, and one is not safer the day after an accident. You might suspect, however, that after an accident, everyone from mechanics to pilots is more watchful and it might be safer.

42. (a) No. The chance for an accident is constant unless the construction enters a dangerous phase. Accidents are, by definition, events that occur spontaneously though a random collection of circumstances. Sabotage, for example, would not be an accident.
(b) We'd rather visit the site that has gone a long time without an accident. That suggests to us that they are very safe, not that they are due for an accident. We'll take the data as evidence that the probability of an accident at the 100-day site is smaller than at the 14-day site.

You Do It

43. (a) 1. $\boldsymbol{S}$ = {blue, orange, green, yellow, red, brown}
2. P(blue *or* red) = P(blue) + P(red) = 0.24 + 0.13 = 0.37
3. P(not green) = 1 – P(green) = 1 – 0.16 = 0.84

(b) 1. S = {triples of three colors}
2. P(blue and blue and blue) = P(blue)^3 = $0.24^3 \approx 0.0138$
3. P(any color and any color and red) = $1 \cdot 1 \cdot 0.13 = 0.13$
4. P(at least one is blue) = 1 – P(none is blue) = $1 - 0.76^3 \approx 0.561$

44. (a) 1. P(2 or less) = 0.55
2. P(more than 5) = 1 – (0.55 + 0.32) = 0.13

(b) 1. P(more than 2 *and* more than 2) = $0.45^2 = 0.2025$
2. P(one has more than 5 years)
= P(more than five and not more than 5) + P(not more than five and more than 5)
= 2 · P(more than five) P(five or less) = $2 \cdot 0.13 \cdot 0.87 = 0.2262$
3. P(at least one with more than 5) = 1 – P(both have 5 or less) = $1 - 0.87^2 = 0.2431$
which is the answer to the prior question plus the probability that both have more than 5 years of experience ($0.13^2 = 0.0169$).

45. (a) Assuming the complaints produced by calls are independent and are equally likely to occur for each call to the foreign call center, then
P(next 3 complain) = P(complain 1st *and* complain 2nd *and* complain 3rd)
= P(complain 1st) · P(complain 2nd) · P(complain 3rd) = $0.62^3 \approx 0.238$

(b) P(complain 1st *and* complain 2nd *and* not complain 3rd) = $0.62^2 \cdot 0.38 \approx 0.146$

(c) $3 \times 0.146 = 0.438$ (There are three sequences that produce no complaint: on the first, second or third call.)

(d) P(none of 10 complain) = $0.38^{10} \approx 0.0000628$.

46. (a) Assuming that the chance for a breakdown is constant and that breakdowns occur at random and independently, then
P(fine on Monday) = P(fine during shift 1 *and* fine during shift 2 *and* fine in shift 3)
= P(fine during shift 1) P(fine during shift 2) P(fine in shift 3)
= $0.85^3 = 0.614125$

(b) P(fine on Monday *and* fails on shift 1 on Tuesday) = P(fine on Monday) P(fails on shift 1 on Tuesday)
= $0.85^3 \cdot 0.15 \approx 0.0921$.

(c) P(breakdown during 3 shifts) = 1-P(runs fine for day) = $1 - 0.85^3 \approx 0.386$

(d) This event has 3 equally likely, disjoint pieces. The breakdown can happen in any of 3 shifts, and the other two shifts must run without a failure. The probability for any one of these is
P(breakdown in shift 1 and OK shift 2 and OK shift 3) = $0.15 \times 0.85^2 \approx 0.108$
The probability of a breakdown in any one shift is thus $3 \cdot 0.15 \times 0.85^2 \approx 0.325$.

47. (a) If we assume that components fail independently, the probability that the computer works is the probability that all components work, or $(999/1000)^{100} \approx 0.905$. This means that the probability of a system failing is about 10%.

(b) If we want the probability of the system working to be 0.99, then we require that the probability p of a component working (still assuming independence) is $p^{100} = 0.99$.
Solving for p by using logs, $\log p = (\log 0.99)/100$, we find that $p = 0.9999$. The rate of defects must be reduced to 1 in 10,000.

(c) $P(F_1 \text{ or } F_2 \text{ or } \ldots F_{100}) \le 100 \cdot 0.001 = 0.1$. The actual probability of this event found in part a is, assuming independence, slightly smaller at 0.095. The bound is tight in this example because the events have such small individual probability that there's not much double counting.

48. (a) Let W_i denote the event that the i^{th} coupon is a winner. Then the probability of winning on the seven purchased items is

$$\begin{aligned} P(W_1 \text{ or } W_2 \text{ or } \ldots W_7) &= 1 - P(W_1^C \text{ and } W_2^C \text{ and } \ldots \text{ and } W_7^C) \\ &= 1 - P(W_1^C)\, P(W_2^C) \ldots P(W_7^C) \text{ (by independence)} \\ &= 1 - 0.95^7 \\ &\approx 1 - 0.698 \\ &= 0.302 \end{aligned}$$

(b) Looking back at the prior calculations, we need to find k so that $1 - 0.95^k = 0.5$, which gives $k \approx 13.5$. So, they need to buy at least 14 items.

(c) Boole's inequality is not so accurate as in the prior question because the probability of each event is larger (even though they are fewer in number).

$$P(W_1 \textit{ or } W_2 \textit{ or } \ldots \textit{ or } W_7) \le 7 \times 0.05 = 0.35$$

In this case, it's off by about 15%.

49. By assuming independence (*i.e.*, not getting rattled), the probability of answering them all correctly is $0.8^6 \approx 0.262$.

50. No. She knows more than she is saying. With a 25% chance getting them each right by guessing, the probability (again, assuming independence) of getting them all correct is only $0.25^6 \approx 0.000244$. That makes it doubtful that she was guessing.

51. (a) The probability of staying for more than a year and having a college education is $0.75 \times 0.75 = 9/16$. This calculation presumes independence, which is questionable here. It could be the case, for example, that those with college education are less likely to stay because of other opportunities.

(b) This probability is zero, with no qualifications needed. The probability of staying for two years (which is zero) is at least as large as the probability for staying for two year *and* being college educated.

52. (a) $0.86*0.93 \approx 0.800$; requires independence.

(b) $0.83^4 \approx 0.475$

(c) $0.72^4 \approx 0.269$; probability of all four changed more.

53. (a) $169/365 \approx 0.463$

(b) India lowest ($109/1250 \approx .087$), US highest ($152/319 \approx .476$)

(c) $(1 - (152/319))*(1 - (109/1250))*(1(- 60/250))*(1 - (44/122)) \approx 0.232$

54. (a) For black males, the rate is 264 per 100,000, so that the probability from this study is 264/100,000 = 0.00264.

(b) For Japanese-American women, the probability of developing cancer is 50/100,000 = 0.0005. Assuming that the outcomes for these women are independent, the probability that none develops cancer is $0.9995^4 \approx 0.998$. Hence, the probability of cancer for any one of these is 0.002, *less* than the chance for one black male (part a).

(c) The assumption of independence is unlikely to hold if the women have a common genetic disposition that favors the development of cancer. Were they all in the same family, one might suspect the appearance of lung cancer in one to the indicative of the incidence in others.

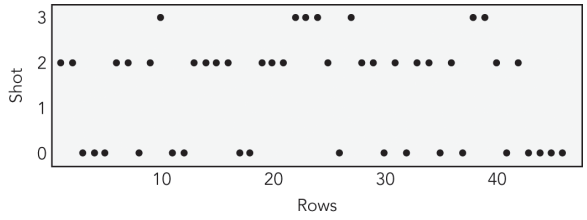
55. (a) The customer must get 40% or 50% off. These are disjoint events, so the probabilities sum to 1/8.

(b) The clerk was right to be surprised. We can calculate that the probability of 3 in a row, independently, is $(1/32)^3 \approx 0.00003$. Not much chance of 3 in a row if the events are truly independent.

(c) Assume that the purchase amount is unrelated to the discount (the card is scratched off at the time of purchase only). Break the overall event into disjoint pieces like this:
P(save more than \$20) = *P*(50% coupon and buy sweater) + *P*(more than 10% and buy suit). The only way to save more than \$20 when buying the sweater is to have a 50% coupon. For those spending \$200, the only those with the 10% coupon save \$20 or less. Now use independence:
P(50% coupon and buy sweater) = 1/32 × ½ and *P*(more than 10% and buy suit) = ½ × ½ .
Hence the overall probability is *P*(save more than \$20) = 1/64 + ¼ ≈ 0.266.

56. (a) Use the complements rule (and independence): *P*(at least one winner) = 1 – *P*(all lose) = 1 – (¾)^3 ≈ 0.578
(b) Think of the outcomes in the sample space as 3 consecutive letters *N*, *F*, or *S* (for nothing, fries and sundae). The sample space consists of all possible sequences of three of these letters, *{NNN}, {NNF}, {NNS}, {NFN}…*and so forth. The outcomes that meet this condition are those that are arrangements of *N*, *F*, and *S*. There are 6 such arrangements, {NFS, NSF, SNF, FNS, FSN, SFN}
each with the same probability: *P*({*N*,F,S}) = (3/4) (1/8) (1/8)= 3/256 ≈ 0.0117.
Thus, as these are disjoint outcomes, *P*(one fries *and* one sundae *and* nothing) = 18/256 ≈ 0.0703.
(c) To win \$5 or more, the family must get either 3 sundaes or 2 sundaes and fries. These are disjoint, so we can add the probabilities. *P*(win \$5 or more) = *P*(win \$5) + *P*(win \$6).
Now use independence: *P*(win \$6) = *P*(3 sundaes) = (1/4)^3 = 1/64
and calculations similar to those in part b: *P*(win \$5) = 3 × *P*({SSF}) = 3 × 1/64.
Adding, *P*(win \$5 or more) = 4/64 = 0.0625.

57. (a) The plot of the sequence of shots taken does not indicate anything out of the ordinary, especially if you have observed the surprising randomness in the experiments of the previous two questions.
(b) One explanation for the cooling down at the end involves fouls by the opposing team. Bryant scored 12 points on free throws in the 4th period alone (18 points for the game). Misses associated with fouls are counted in the box score as misses just the same. Bryant was fouled on the last 3 shots (which were misses).



58 4M Racetrack odds

(a) No. It would tend to break even on wagers. It needs a profit to pay expenses.

(b) It pays less at 10 − 1 than at 11 − 1 if Mubtaahij wins.

(c) 1/11, 1/16, 1/13, 1/31, 5/8, 1/6, 1/21, 1/7.

(d) 1/11+1/16+1/13+1/31+5/8+1/6+1/21+1/7 ≈ 1.245

(e) Racetrack odds are not true odds and allow the race track to take a share of the amount wagered to pay for expenses.

4M Auditing a Business

(a) By sampling a few transactions and thoroughly investigating these, the auditor is likely to do a better job than by trying to check *every* transaction. It might not even be possible to check every transaction: the firm might generate more transactions in a day than the auditor could check in a day. If the auditor finds no anomalies in, say, 30 or 40 transactions, it's unlikely that there's a lot of fraud going on. Of course, a careful thief could get away with this.

(b) At random. Without further input, we need to make sure that we get items from all over the business, without assuming some are more likely to be fraudulent. If, however, managers suspect a problem in certain situations, we should check these first.

(c) We'd end up knowing a lot about this one division, but nothing about the others unless we're willing to believe that fraud is equally likely over the entire firm.

(d) Let the event S_i denote sales fraud is detected for the i^{th} audited receipt. Hence, the event S_i^C denote no fraud was detected for this sale. If we assume that the results for the receipts are independent, then

$$\begin{aligned} P(\text{no sales fraud detected}) &= P(S_1^C \textit{ and } S_2^C \textit{ and } \ldots \textit{ and } S_{25}^C) \\ &= P(S_1^C)\, P(S_2^C) \ \ldots\ P(S_{25}^C) \qquad \text{(independent)} \\ &= 0.98^{25} \\ &\approx 0.603 \end{aligned}$$

(e) The probability of finding no fraud among the 25 inventory checks is (following the prior steps)

$$P(\text{no inventory fraud detected}) = 0.97^{25} \approx 0.467$$

The probability of finding no fraud of either type is then

$$\begin{aligned} P(\text{no fraud detected}) &= P(\text{no sales fraud } \textit{and} \text{ no inventory fraud}) \\ &= P(\text{no sales fraud})\, P(\text{no inventory fraud}) \\ &\approx 0.603 \cdot 0.467 \approx 0.282 \end{aligned}$$

Hence, the chance of finding some fraud is about $1 - 0.282 = 0.718$.

(f) Because inventory fraud is more common, we should audit 100 of these transactions. The chance of finding no fraud among these is smaller 0.97^{100} than the chance of finding a sales fraud. The choice to audit only inventory transactions depends on knowing that there's a higher rate of fraud in these transactions rather than sales transactions. If that knowledge is wrong - or suspect - we could be making the wrong choice. The point, however, is that you should audit the area where you suspect a problem. Just don't assume there's no fraud elsewhere.

(g) No. There could be fraud and we missed it. Direct calculations in part e show that even if there's a 2% chance of sales fraud and a 3% chance of inventory fraud, an evenly divided audit of 50 transactions has a 28% chance of finding only legitimate transactions.

Chapter 8: Conditional Probability

Mix and Match

1. f

2. d

3. c

4. e

5. b

6. a

True/False

7. False. The statement asserts that $P(\mathbf{A}) > P(\mathbf{A}|\mathbf{S})$. This might happen, but need not be the case.

8. False. The statement asserts that $P(\mathbf{A}|\mathbf{S}) = P(\mathbf{S}|\mathbf{A})$. This may or may not occur.

9. False. The statement asserts that $P(\mathbf{A})>\mathrm{P}(\mathbf{S})$, but these are only marginal probabilities and the ordering does not imply dependence.

10. True. $[P(\mathbf{A}|\mathbf{S}) = P(\mathbf{A}\ \textit{and}\ \mathbf{S})/\mathrm{P}(\mathbf{S})] > [P(\mathbf{S}|\mathbf{A}) = P(\mathbf{S}\ \textit{and}\ \mathbf{A})/P(\mathbf{A})]$ because the fractions have the same numerator but $P(\mathbf{A}) > P(\mathbf{S})$.

11. False. She also needs a joint probability.

12. True. Independence requires that the probabilities multiply, and order is not important when multiplying.

13. False. Independence implies that one event does not influence the chances for the other.

14. False. The expression should be $P(\mathbf{A}_1|\mathbf{A}_2) = P(\mathbf{A}_1)$.

15. False. These are disjoint events, so the probability of the intersection must be 0.

16. False. These would only be equal if independent and $P(\mathbf{A}) = P(\mathrm{W})$.

17. True. Think about how the mosaic plot looks when there is independence.

18. False. The indicated percentages are conditional probabilities, $P(\mathbf{W}|\mathbf{M}) = 0.2$ and $P(\mathbf{W}|\mathbf{A}) = 0.4$.

19. True.

20. True. The events are dependent because the conditional probability does not match the marginal probability; $P(\text{W}|\mathbf{M}) = 0.75 \neq 0.3 = P(\text{W})$.

Think About It

21. Independent. It is unlikely that seeing a Honda, for example, would make us suspect that the next car was also a Honda (unless there's a parade of Hondas which is unlikely on an interstate highway).

22. Dependent. Severe winter weather is likely to produce a collection of related accidents of the same type. We would expect most to be fender benders due to the icy conditions.

23. Dependent. The number of visits today is probably influenced by the same factors that influenced the number of visits yesterday.

24. Dependent. Many in fact claim that the *best* predictor of the price of a stock tomorrow is the price today.

25. Most likely independent, unless we know that there is some particular sale in progress that has drawn shoppers looking for the same item.

26. Most likely independent, unless a caravan of large SUVs or hybrids arrive together at the station.

27. Independent. If **A** happens, then the chance for **B**, $P(\mathbf{B}|\mathbf{A})$, remains 1/4 because **B** is ¼ of the area of **S**.

28. Dependent. If **A** occurs, then the chances for **B** grow to $P(\mathbf{B}|\mathbf{A}) = ½$ which is larger than the marginal probability for **B**, $P(\mathbf{B}) = 1/4$.

29. (a) Classify everyone as a drug user. In this way the sensitivity of the test must be 1, P(test says drug use | use drugs) = 1.
(b) The problem is that the test will not be very specific; the test will have many false positives. In particular, any clean person who takes the test will be falsely accused of using drugs.

30. The company needs to determine the marginal probability of healthy people who take the test. (See the text discussion of Bayes' rule and reversing conditional probabilities.)

31. No. The statement of the question first gives $P(\mathbf{F}|\mathbf{Y}) = 0.45$ and then $P(\mathbf{Y}|\mathbf{F}) = 0.45$. In order to be independent, $P(\mathbf{F}|\mathbf{Y}) = P(\mathbf{F})$. We are not given marginal probabilities and hence cannot determine that the events are independent. To see some pictures, have a look at the Venn diagrams for Exercises 27 and 28. In both cases, $P(\mathbf{A}|\mathbf{B})=P(\mathbf{B}|\mathbf{A})$, but in Exercise 27 the events are independent whereas in Exercise 28 they are dependent.

32. These must be the same. Because the conditional probabilities match,
$P(\mathbf{F}|\mathbf{Y}) = P(\mathbf{F}\ \textit{and}\ \mathbf{Y})/P(\mathbf{Y})$ is the same as $P(\mathbf{Y}|\mathbf{F}) = P(\mathbf{F}\ \textit{and}\ \mathbf{Y})/P(\mathbf{F})$
Hence, canceling the common term (the intersection), $P(\mathbf{F}) = P(\mathbf{Y})$.

33. (a) 0.42 = P(working affected grades | have loan). The sample space might be the collection of recent college grads.
(b) You cannot tell. In general, you cannot obtain $P(\mathbf{B}|\mathbf{A})$ from $P(\mathbf{A}|\mathbf{B})$ without the marginal probabilities. In this example, you don't know the proportion who worked in college.

34. (a) P(sold possessions | have debt) = 0.33, P(sold possessions | no debt) = 0.17. The sample space is the collection of recent college graduates (or perhaps the collection of those covered in this survey).
(b) To find P(sold possessions), you need the marginal proportion that have debt after graduating or the marginal proportion that have not had debt after graduating. With these, you could use a table or Bayes' Rule to reverse the order of conditioning.

You Do It

35. (a) The choice is up to you, but a tree is easy to use in this context for two main reasons: the probabilities are given in a sequential form (as conditional probabilities), and the options for the two types of vehicles are *not* the same. You cannot get a sunroof on a truck from this brand, but you can get one on a car.
(b) P(sunroof) = P (car) × P (sunroof | car) = 1/2 × 1/4 = 1/8.

36. (a) A table works well to collect this information. You can also use a tree and sum the relevant events (those that include clothing) from the final leaves of a tree to answer the question in part b.
(b) The following table shows the results. The problem states marginal probabilities for the columns, which are then divided among clothing and camping supplies (conditionally on the type of order). Adding the marginal row probabilities gives P(clothing *or* (clothing *and* supplies)) = 0.6 + 0.2 = 0.8.

	In Store	Online	Telephone	
Clothing alone	0.25	0.20	0.15	0.60
Camping supplies	0.00	0.15	0.05	0.20
Both	0.00	0.15	0.05	0.20
	0.25	0.50	0.25	1

37. P(drink *and* popcorn) = P(drink) P(popcorn | drink) = $0.7 \times 0.3 = 0.21$

38. P(\$200 or more) = P(major repair *and* \$200 or more) = P(major repair) P(\$200 or more | major repair) = $0.3 \times 0.5 = 0.15$. In this example, P(**A**) = P(**A** *and* **B**) because **A** is a subset of **B**. (Draw the Venn diagram to see what's happening.)

39. You need to reverse the conditioning. This problem can be solved by Bayes' Rule. A table organizes the information easily without the need to remember the formula and provides a check on the calculations. If a player fails (the first column), then there's a $0.50/0.55 \approx 0.91$ chance that it has the flaw.

	Fails in Six Months	Does Not Fail	Total
Has flaw	0.50	0	0.50
No flaw	0.05	0.45	0.50
Total	0.55	0.45	1

40. You need to reverse the conditioning. Organize the given information in a table (as above in the prior solution) or use the formula for Bayes' Rule directly as done below.

$$P(\text{director}|\text{MBA}) = P(\text{MBA } \textit{and} \text{ director})/P(\text{MBA}) = (P(\text{MBA}|\text{director}) \times P(\text{director}))/P(\text{MBA})$$
$$= 0.6 \times 0.15 / (0.6 \times 0.15 + 0.35 \times 0.85) \approx 0.23$$

41. We assume that any order of selection of the parts is equally likely.
(a) P(both good) = P(first good) × P(second good | first good) = (7/12) (6/11) = 7/22
(b) P(at least one is good) = 1 – P(none is good)
= 1 – P(first bad) P(second bad | first bad)P(third bad | first two bad)|
= 1 – (5/12)(4/11)(3/10) = 21/22

(c) P(four good) = (7/12)(6/11)(5/10)(4/9) = 7/99
(d) P(four bad then good) = (5/12)(4/11)(3/10)(2/9)(7/8) = 7/792 ≈ 0.00884

42. Because the clothes are in a jumble, we assume that she is equally likely to choose items of any size. (That is, she cannot recognize the shirts in the pile.)
(a) P(neither is medium) = P(first not medium) P(second not medium | first not) = (16/20)(15/19) = 12/19
(b) P(third shirt is first medium) = P(first two are not) P(third is |first two are not) = (12/19)(4/18) = 8/57
(c) P(four extra large) = (6/20)(5/19)(4/18)(3/17) = 1/323 ≈ 0.003096
(d) P(at least one medium) = 1 – P(no medium) = 1 – (16/20)(15/19)(14/18)(13/17) ≈ 0.624

43. (a) 1/12
(b) 1/11
(c) The events are dependent. Let the event **A** denote finding the system with the missing component first, and **B** denote finding it second. Then $P(\mathbf{B}|\mathbf{A})$ is 0 because he cannot find it on the second system if he already found it on the first. In lay language, he has 12 possible choices first, but only 11 possible choices second. He's more likely to find it second since he's eliminated one of the extra choices.

44. (a) 0.15
(b) 0.15. The same.
(c) The answers should match if you believe that absences are independent over the collection of employees. That is, the presence or absence of the first employee has no impact on the status of the second. This might be reasonable so long as it's not flu season. Alternatively, if you believe that absences are related (it is flu season), then you'd expect dependence, so if one employee is absent, others will be as well. You can see the impact this has on staffing levels.

45. (a) Dependent. The probability that the first assignment goes to a man is 25/50, but this event changes the probability for the second assignment (because there is one less man who can receive the assignment). The probability that the second assignment also goes to a man given that the first is to a man is 24/49 < 1/2.
(b) P(fifth lead to a man | first four leads go to men) = 21/46
(c) You can do this one several ways. Let **A** denote the event that at least four leads go to men, and let **B** denote the event that men receive all five leads. The question asks for $P(\mathbf{B}|\mathbf{A}) = P(\mathbf{B}\ \textit{and}\ \mathbf{A})/P(\mathbf{A}) = P(\mathbf{B})/P(\mathbf{A})$ (**B** is a subset of **A**). Let's find P(**A**) first. If W stands for a lead given to a woman and M for a lead given to a man, then the event **A** means that the final result must be one of these.

WMMMM, MWMMM, MMWMM, MMMWM, MMMMW, MMMMM

The chances for any of the events with a woman getting one lead are the same. For example, multiplying the conditional probabilities, P(MMWMM) = (25/50)(24/49)(25/48)(23/47)(22/46) ≈ 0.02985
and P(MMMMW) = (25/50)(24/49)(23/48)(22/47)(25/46) ≈ 0.02985.
Now that you have seen two, you can see why all five are the same. No matter where we put the W, the probability is always

$$\frac{25\times 24\times 23\times 22\times 25}{50\times 49\times 48\times 47\times 46}$$

The probability for the last outcome is

P(**B**) = P(MMMMM) = (25/50)(24/49)(23/48)(22/47)(21/46) ≈ 0.02508

Hence the probability of four or more men is (the events are disjoint, and so add)

$5 \times 0.02985 + 0.02508 = 0.17433$

The probability of five men given at least four men is then

$P(\mathbf{B})/P(\mathbf{A}) = 0.02508/0.17433 \approx 0.1439$

46. (a) Let **A** denote that the Type A component is within specifications and **B** denote that the Type B component is within specifications. Assuming that the random selection eliminates any type of dependence, then the events **A** and **B** are independent, so $P(\mathbf{A}\ \textit{and}\ \mathbf{B}) = 0.95 \times 0.90 = 0.855$.

(b) It's tempting to guess that the chance that it's A is 1/3, with 2/3 going to B since B has twice as many out-of-spec components as A. It's just not right.

We are given that either A is out of spec (the event $\mathbf{A}^C$) or B is out of spec ($\mathbf{B}^C$) or both are out of spec ($\mathbf{A}^C$ *or* $\mathbf{B}^C$). The question asks for

$P(\mathbf{A}^C|\mathbf{A}^C \textit{ or } \mathbf{B}^C) = P(\mathbf{A}^C \textit{ and } (\mathbf{A}^C \textit{ or } \mathbf{B}^C))/ P(\mathbf{A}^C \textit{ or } \mathbf{B}^C) = P(\mathbf{A}^C)/P(\mathbf{A}^C \textit{ or } \mathbf{B}^C)$. From the statement of the question and the Complements Rule, $P(\mathbf{A}^C)=1- P(\mathbf{A}) = 0.05$. Similarly, $P(\mathbf{A}^C \textit{ or } \mathbf{B}^C) = 1- P(\mathbf{A} \textit{ and } \mathbf{B}) = 1 - 0.855 = 0.145$. Hence, $P(\mathbf{A}^C|\mathbf{A}^C \textit{ or } \mathbf{B}^C) = 0.05/ 0.145 \approx 0.345$.

47. (a) P(Service | Man)
(b) $32/(32+17) \approx 0.653$
(c) Easiest for mining and construction; hardest for management and professional.
(d) $(14+20)/200 = 0.17$

48. This problem can be solved by Bayes' Rule because we are given P(Defective|Supplier) and need to reverse the conditioning. It may be helpful to organize the information into a table such as the one that follows. The marginal sums of the columns are given in the problem, and splitting out the conditional probabilities gives the values inside the table and ultimately the row marginal totals. Once you have this table, the questions are easily answered.
(a) P(Defective) = 0.06
(b) P(Supplier A | Defective) = 1/6

	Supplier A	**Supplier B**	**Total**
Defective	0.01	0.05	0.06
OK	0.49	0.45	0.94
Total	0.5	0.5	1

49. A table makes it easy to organize the information. The values in the following table in bold come directly from the statement of the exercise. The rest come from making the values add up correctly.
(a) P(Luggage in SF) = 0.82
(b) P(Arrived late in Dallas|No luggage in SF) = 0.1/0.18 = 5/9

	Arrive on Time	**Arrive Late**	**Total**
Luggage	**0.72**	**0.1**	0.82
No luggage	0.08	0.1	0.18
Ttal	**0.8**	0.2	1

50. This is an application of Bayes' Rule. It may be easier to compute by filling in a table like this one which organizes what is known. The numbers shown in bold are calculated from the given information. Once you have the table, it's easy to see that the probability of a foreign agent given a complaint occurs is

P(Foreign | Complains) = 0.372/0.496 = 3/4

	Caller complains	**Does not complain**	**Total**
Foreign	$0.6\times0.62 =$ **0.372**	**0.228**	0.6
American	$0.4\times0.31 =$ **0.124**	**0.276**	0.4
Total	**0.496**	**0.504**	1

51. It helps in problems such as this to organize the information in a table. We are given the information shown in bold in the table below. We are also given *P*(No internet | Computer) = 0.08. The other probabilities are derived as follows:
P(No internet) = 1 – *P*(Internet) = 0.29
P(Computer) = *P*(Computer and Internet) + *P*(Computer and No internet)
.77 = .71+ *P*(Computer and No internet)
P(Computer and No internet)=.06

	Internet Access	**No Internet**	**Total**
Have computer	**0.71**	0.06	0.77
No computer	0	0.23	0.23
Total	**0.71**	0.29	1

It then follows that *P*(No computer | No internet) = 0.23/0.29 = 0.79.

52. (a) These statistics imply dependence. The chance for having this illness overall is 0.24, but the chance among smokers is 0.60. Hence *P*(Illness) ≠ *P*(Illness | Smoke)
(b) *P*(Smoke | Illness) = 0.06/(0.06+0.135)=0.31

4M Scanner Data
The following table shows the conditional distributions within columns.

		Number of Dogs Owned					
	Joint prob. Col probability	0	1	2	3	More than 3	Total
Number of dog food items purchased	**0**	4.87 7.35	2.17 7.89	0.25 4.51	0.02 3.75	0.00 0.91	7.31
	1 to 3	16.98 25.61	7.34 26.66	1.04 18.86	0.04 9.58	0.02 10.00	25.42
	4 to 6	11.82 17.83	5.16 18.73	0.93 16.85	0.06 12.92	0.02 10.00	17.99
	7 to 12	11.60 17.50	4.69 17.01	1.13 20.54	0.12 27.08	0.05 25.45	17.59
	More than 12	21.03 31.72	8.18 29.71	2.16 39.24	0.21 46.67	0.11 53.64	31.69
	Total	66.30	27.54	5.51	0.45	0.20	100

Motivation

(a) One would expect dependence. Presumably, shoppers who own more dogs will be inclined to buy more dog food. (That would be positive dependence in the sense of Chapter 6.)

(b) Most likely it means that the shopper owns a dog, but did not indicate that when applying for the shopper's card or perhaps the dog was acquired after filling out the application. A shopper might also buy the food for a friend, relative, or neighbor. Finally, the self-reported data may not be accurate.

Method

(c) The most useful to compute first would be column probabilities. These describe shopping habits given the number of pets owned. Assuming that the self-reported data is reliable (note part b above), the store could then anticipate the likelihood of customers wanting to buy relatively large or small amounts of pet food and use this insight to produce, for example, specific incentives.

(d) No. These *joint* probabilities are small because relatively few shoppers report having more than 3 dogs. Within this column (conditional on high ownership), the most common number of items is more than 12.

(e) Most likely a rounding error or careless computation mistake. As probabilities, they must add to 1.

Mechanics

(f) The table shown above adds the marginal probabilities in the gray border. In the table, probabilities are shown multiplied by 100 as percentages.

(g) From the conditional probabilities shown in the table above (proportions within a column, shown second in each cell), *P*(more than 3 items | no reported dogs) = 17.83 + 17.5 + 31.72 = 67.05%
compared to *P*(more than 3 items | more than 3 dogs) = 10 + 25.45 + 53.64 = 89.09%
The number bought by those who report no pets is surprisingly large, but at least smaller than those who do say that they have more than 3 dogs.

(h) Buying eight cans puts the shopper in the fourth row of the table. Within this row, we have the joint probabilities

0	1	2	3	More than 3
0.1160	0.0469	0.0113	0.0012	0.0005

which sums to the marginal probability 0.1759. Hence, *P*(no dogs|bought 8) = *P*(no dogs *and* 7−12)/ *P* (7−12) = 0.1160/0.1759 ≈ 0.659 so that *P*(report own dogs|bought 8) = 1 − 0.659 = 0.341.
This probability suggests something odd in the data. The most likely number of dogs for someone who buys this much is none! Evidently (as suggested in part b), the number of dogs seems to be underreported.

Message

(i) The scanner data as reported here will not be very useful. Relatively few customers report owning more than three dogs. The conditional probability of buying more than 12 cans, say, is highest in this group of customers (54% versus 32% for those who claim to own no dogs), but their small number limits sales opportunities. Most customers report that they own no dogs. Nonetheless this group who claims not to own a dog is the most prevalent among shoppers who buy large quantities.

4M: Fraud Detection

(a) Mistaking honest for fraud risks annoying or embarrassing customers; missing fraud means losing merchandise.

(b) *P*(fraud|signal fraud) or *P*(honest|did not signal fraud)

(c) *P*(honest|signal fraud)=1/2.

(d) *P*(honest|signal fraud)=0.0095/0.059≈0.16

(e) If fraud is rare (say 1% or less), too likely to falsely signal as fraud (too many honest transactions among those labeled as fraud). With higher rates (5% or higher), may be adequate, depending on costs of annoying customers and size of transactions.

Chapter 9: Random Variables

Mix and Match

1. g

2. e

3. f

4. h

5. b

6. a

7. d

8. c

True/False

9. True.

10. False. Even though the probabilities sum to 1, one of them is negative.

11. False. The mean of X should be smaller than the mean of Y.

12. True. The mean always lies within the range of possible outcomes since it is a weighted average of these outcomes.

13. False. The mean is the weighted average of possible outcomes with all weights between 0 and 1; it need not be one of the outcomes.

14. False. The expected value is a parameter of the random variable and need not match the mean obtained from the data.

15. False. The variance in the shown stock examples is larger, but this need not be the case. The variance measures the spread around the mean.

16. False. Properties of the mean of a random variable are comparable to means of the data. If, for example, the probability distribution is skewed to the left, then more than half of the probability (as with histograms of data) is to the left of the mean.

17. True.

18. True. The mean is a weighted average of the outcomes with all weights between 0 and 1, so it has to lie within the range of possible outcomes. Here, that's from 0 to $500,000.

19. True, assuming all other things remain the same.

20. True. The random variable is multiplied by the constant 1.05. The SD increases by 1.05.

Think About It

21. $P(X = -2) = 0.3$.

22. $P(Y \geq 2.5) = 1/7$

23. $P(Z \leq -3) + P(Z = 3) = 0.05 + 0.05 = 0.10$.

24. $P(W > 0) = 0.17+0.29+0.27=0.73$ is larger than $P(Y > 0) = 1/7+1/7+1/7=3/7$.

25. $E(Z) > 0$ since the probabilities of the positive values are larger than the corresponding negative probabilities.

26. *W. W* has the largest mean because it has the highest probability on the largest values.

27. *Y* has the largest SD. The uniform distribution spreads the values farther from the center than the others that have some clustering.

28. 1.5. This is the best guess for how far a typical value is from the mean for this distribution.

29. 0. Notice that $P(Y \leq 0) = ½$.

30. The lower quartile of X is −2. The probability at −3 is only 0.2, but it is 0.3 at −2. In fact $P(X \leq -2) = 0.5$. The lower quartile and median of X are the same.

31. (a) $P(W=5) = p(5) = 1/10$ and $P(W = -1) = p(-1) = 9/10$. The probability distribution has two non-zero values, with the height at $W = 5$ being 9 times higher than the other.
(b) The expected value of W is $5 \times (1/10) + (-1)(9/10) = -4/10 = -0.4$. The game is not fair because the mean is not zero.

32. (a) $P(X = -1) = 999/1{,}000$ and $P(X = 499) = 1/1{,}000$. The probability distribution is a very high point at −1 and a point close to the x-axis out at 499.
(b) The mean is 499/1,000 – 999/1,000 = −500/1,000. It's not a fair game, and the state does not want it to be a fair game. They want to make a profit.
(c) On average, buying a ticket in the lottery is like paying the state half of a dollar.

33. (a) Fair. The player has half of the probability and contributes half of the pot.
(b) Better than fair to the player. The player contributes 1/100 of the pot but has a larger share of the probability of winning (1/52).

34. The total amount wagered by the two companies to contest the standard is $30 million. Since Company A has put up 1/3 of this total, the chance of its choice being accepted is 1/3. Of course, this does not imply that it actually has a 1/3 chance of winning the contest. Its advertising or promotion may be much more effective than that of its rival. Nonetheless, if the impact of advertising is comparable to money invested, then this notion of a fair game gives us a different way to think about the possibilities.

You Do It

35. The means and standard deviations are

	μ	σ
$X/3$	40	5
$2X - 100$	140	30
$X + 2$	122	15
$X - X$	0	0

36. The means and standard deviations are

	μ	σ
$2\ Y + 20$	70	20
$3\ Y$	75	30
$Y/2 + 0.25$	13	5
$6 - Y$	–19	10

37. (a) There are three possible values for the investor, depending on whether both stocks rise, both fall or one goes up while the other goes down.

Outcome	$P(X)$	X
Both increase 80% to $18,000	1/4	$36,000
One increases	1/2	$22,000
Both fall 60% to $4,000	1/4	$8,000

(b) E(X) = 36,000/4 + 22,000/2 + 8,000/4 = $22,000.
(c) Yes, the probability is symmetric around the mean gain of $2,000.

38. (a) There are four equally possible scenarios, but two have the same value at the end. A probability tree is helpful in this one to keep track of the outcomes and values.

Outcome	$P(Y)$	Y
Increase, then increase	1/4	$32,400
Increase, then decrease	1/4	$7,200
Decrease, then increase	1/4	$7,200
Decrease, then decrease	1/4	$1,600

(b) E(Y) = 32,400/4 + 7,200/2 + 1,600/4 = $12,100.
(c) Not very well because of the skewed distribution of the returns. On average, the investor makes $2,100 on the investment, but three-fourths of the time the investor loses!

39. Let the random variable X denote the earned profits. Then the probability distribution of X is $p(0) = P(X = 0) = 0.05$, $p(20{,}000) = 0.75$, $p(50{,}000) = 0.20$.
(b) The expected value of X is $(0)(0.05) + (20{,}000)(0.75) + (50{,}000)(0.20) = \$25{,}000$.
(c) The variance of X is $(0-25{,}000)^2\,(0.05) + (20{,}000-25{,}000)^2\,(0.75) + (50{,}000-25{,}000)^2\,(0.20) = 175{,}000$.
Hence $\sigma = \sqrt{175{,}000} \approx \$13{,}229$.

40. (a) Let the random variable X denote the settlement amount. Then the only way for the firm to earn a fee is for the case to go to trial and for the firm to win the trial. The probability for this is $(1/3)(1/2) = 1/6$. Hence $p(0) = P(X = 0) = 5/6$, $p(25{,}000) = 1/6$.
(b) $E(X) = 25{,}000 \times 1/6 \approx \$4{,}167$
(c) $\text{Var}(X) = (0-4{,}167)^2\,(5/6) + (25{,}000-4{,}167)^2\,(1/6) \approx 86{,}805{,}556$ so $\sigma \approx \$9{,}317$.

41. (a) $E(X) = 0\,(0.05) + 1\,(0.25) + \ldots + 5\,(0.05) = 2.25$ reams.
(b) $\text{Var}(X) = (0-2.25)^2\,(0.05) + (1-2.25)^2\,(0.25) + \ldots + (5-2.25)^2 = 1.5875$ so that $\sigma \approx 1.26$ reams.
(c) $E(20 - X) = 20 - 2.25 = 17.75$ reams.
(d) 1.26 reams (same as part b because adding or subtracting a constant has no effect).
(e) $E(500X) = 500E(X) = 1{,}125$ pages; $\text{SD}(500X) = 500\text{SD}(X) \approx 630$ pages.

42. (a) $E(Y) = 0(0.2) + 1(0.15) + \ldots + 4(0.15) = 2.05$ lights.
(b) $\text{Var}(Y) = (0 - 2.05)^2\,(0.2) + (1 - 2.05)^2\,(0.15) + \ldots + (4 - 2.05)^2\,(0.15) = 1.8475$ so $\text{SD}(Y) \approx 1.36$ lights.
(c) $E(6 - Y) = 6 - 2.05 = 3.95$ lights.
(d) $\text{SD}(6 - Y) = \text{SD}(Y) = 1.36$ lights (same as part (b)).
(e) $E(10Y) = 10(2.05) = 20.5$ minutes, $\text{SD}(10Y) = 10\text{SD}(Y) = 13.6$ minutes.

43. (a) Let R be a random variable that denotes the bolivar/dollar exchange rate in six months. The problem indicates that $P(R=2.15) = 0.6$ and $P(R=5) = 0.4$. The expected current value of the contract in dollars is then (divide by the rate to get the cost in dollars)

$$E(1{,}000{,}000/R) = 1{,}000{,}000\,E(1/R) = 1{,}000{,}000\,(\,(1/2.15)\times 0.6 + (1/5)\times 0.4)$$
$$\approx 1{,}000{,}000\,(0.359) = \$359{,}000.$$

(b) Dividing the cost in bolivars by the expected value of the exchange rate gives $1{,}000{,}000/3.29 \approx \$304{,}000$. This is substantially less than the expected cost calculated correctly in part a. The error is that $E(1/R) > 1/E(R)$.

44. (a) Let R be a random variable that denotes the peso/dollar exchange rate in a year. The expected value of the rate at the end of the year is

$$E(R) = 12 \times 0.9 + 24 \times 0.1 = 13.2 \text{ peso/dollar.}$$

The expected value of the reciprocal of this rate is

$$E(1/R) = (1/12) \times 0.9 + (1/24) \times 0.1 = 0.07917.$$

As in the previous question, $E(1/R)$ is larger than $1/E(R) = 1/13.2 = 0.07576$. This difference explains the different situations for the investors.
To get the higher expected value, an investor in the United States prefers investing in Mexico. The value of the investment in the United States is $1{,}000(1.08) \approx \$1{,}080$. It's a constant. The value of the investment in Mexico grows by 16%, but is subject to the randomness of the future exchange rate. The initial conversion of \$1,000 gives 12,000 pesos which in a year grow to $12{,}000 \times (1.16) = 13{,}920$ pesos. The expected value of these pesos in dollars at the future exchange rate is $E(13{,}920/R) = 13{,}920E(1/R) = 13{,}920 \times 0.07917 = \$1{,}102$.
(b) The Mexican investor gets a higher expected value, in comparison, by investing in the United States. The value in pesos is known: it will grow by 16% from 12,000 to $1{,}2000 \times 1.16 = 13{,}920$ pesos at the end of the

year. Converting to dollars at 12 pesos per dollar gives the investor $1,000 that earns 8% and grows to $1,080. Converting back to pesos, this is worth, on average,

$E(1080R) = 1080 \times E(R) = 1080 \times 13.2 = 14{,}256$ pesos.

Note that this conversion requires R, not $1/R$.

(c) The investor in the United States gets larger expected values in Mexico, whereas the Mexican investor gets larger expected values in the United States. This difference in effect is due to $E(1/R)>1/E(R)$, noted previously. In particular, for an investor in the United States,

$$1.08 < 1.16 \times E(1/R) \Rightarrow 1.08/1.16 < E(1/R)$$

whereas for the investor in Mexico

$$1.16 < 1.08 \times E(R) \Rightarrow 1.16/1.08 < E(R) \Rightarrow 1/E(R) < 1.08/1.16$$

It is also worth noticing that these calculations do not take account of the risks (or variances) in the values of the investments.

45. (a) Let $X = 0, 1$ denote whether a customer buys the printer when no rebate is offered (1 for yes and 0 for no). The expected value of X is $E(X) = 0 \times (1-p) + 1 \times p = p$. Similarly, let X^* denote whether a customer who is offered a rebate buys a printer; $E(X^*) = p^*$.

If the company does not offer a rebate, its profits are $60X$ with expected value

$E(60X) = 60p.$

If the company offers the rebate, the expected profits are

$E(30X^*) = 30p^*$.

For the rebate to be effective, on average, $30p^* > 60p$, so $p^*/p > 60/30 = 2$; the probability of purchase has to double.

(b) Let $Y = 0, 1$ denote whether the rebate is used. Then the expected profit when the rebate is offered (assuming X^* and Y are independent) is a random variable Z with probability distribution

$$P(Z = 0) = 1 - p^*$$
$$P(Z = 30) = 0.4p^* \text{ (purchase and use rebate)}$$
$$P(Z = 60) = 0.6p^* \text{ (purchase and do not use rebate)}$$

The expected profits are then $E(Z) = 30 \times 0.4p^* + 60 \times 0.6p^* = 48p^*$. For the rebate to be effective, $48p^* > 60p$ so $p^*/p > 60/48 = 1.25$, quite a bit smaller than in part a.

46. (a) Let the random variable X denote damage caused by lightning surges over five years. For the surge protector to have positive value (on averag(e), $E(X)$ must be larger than $50.

$E(X) = p \times 4{,}000 + (1-p)0 = 4{,}000p > 50$

or $p > 50/4000 = 0.0125$.

(b) Assuming lightning is equally likely to strike any household (which is not true, by the way) and never hits the same place twice (also not true), then the chance for a home being struck during five years is $(5 \times 10{,}000)/100{,}000{,}000 = 0.0005$, which is quite a bit smaller than the break even probability in part a. It looks like the surge protector is not a good deal.

47. (a) Let X denote the number of clients visited each day. Then the only way that he sees one client is if that client buys a policy. Hence, $P(X = 1) = 0.1$. He sees two clients if the first does not buy a policy, but the second does. Otherwise, he sees three clients.

$$P(X = 1) = 0.1$$
$$P(X = 2) = 0.09$$
$$P(X = 3) = 1 - p(1) - p(2) = 0.81.$$

(b) $E(X) = 1 \times 0.1 + 2 \times 0.09 + 3 \times 0.81 = 2.71$ clients.

(c) $2.5 \times E(X) = 6.775$ hours.

(d) He either sells nothing or one policy each day (since he stops after the first sal(e). Hence, his expected earnings are $3000 times the probability of a sale, which can be found as the sum of the probabilities of three disjoint events:

$P(\text{sal(e)}) = P(\text{first buys}) + P(\text{second buys}) + P(\text{third buys}) = 0.1 + 0.09 + 0.081 = 0.271.$

Hence, his expected earnings per day are $3000 \times 0.271 = \$813$.

48. (a) Let X denote the number of service technicians called in until the machine is either fixed or the money runs out. Then

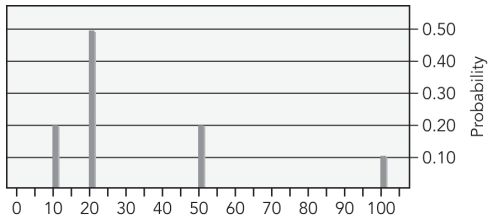
$$P(X = 1) = 0.25 = 1/4 = 16/64$$
$$P(X = 2) = 0.75 \times 0.25 = 3/16 = 12/64$$
$$P(X = 3) = 0.75 \times 0.75 \times 0.25 = 9/64$$
$$P(X = 4) = 1 - [p(1)+p(2) +p(3)] = 1- 37/64 = 27/64.$$

(b) $E(X) = 1\ P(X=1) + 2\ P(X=2) + 3\ P(X=3) + 4\ P(X=4)$
$= 1 \times (16/64) + 2 \times (12/64) + 3 \times (9/64) + 4 \times (27/64) \approx 2.734.$

(c) For the amount spent, each technician costs \$500 for the visit. Plus, you have to pay \$10,000 if none of the four fix the machine. The expected costs for the technicians is E(X) times \$500. The chance that all four of them cannot fix the machine is $(3/4)^4$ (assuming independence of the service calls). Hence, the expected total cost is

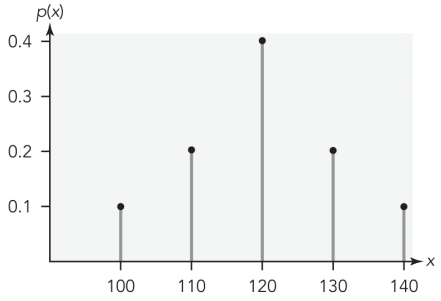
$$E(\text{total cost}) = 500\ E(X) + 10{,}000\ P(\text{all four cannot fix it})$$
$$= 500 \times 2.734 + 10{,}000 \times (0.75)^4 \approx \$4{,}531.$$

49. (a) The probability distribution of the ATM withdrawal is



(b) $p(50) + p(100) = 0.3$.
(c) $E(X) = 10{\times}0.2 + 20{\times}0.5 + 50{\times}0.2 + 100{\times}0.1 = \$32$.
(d) On average, we expect a customer to withdraw \$32. This is the average of the outcomes in the long run, not a value that occurs for any one customer.
(e) $E(X - 32)^2 = (10 - 32)^2\ 0.2 + (20 - 32)^2\ 0.5 + (50 - 32)^2\ 0.2 + (100 - 32)^2\ 0.1 = 696$
so the SD of X is $\sqrt{696} = \$26.4$.

50. (a) The probability distribution is symmetric and unimodal around 120 ¥/\$,



(b) $P(X > 120) = 0.3$.
(c) $E(X) = 120$ ¥/\$.
(d) $100{,}000 \times 120 =$ ¥12,000,000.
(e) Working directly from the probabilities, the expected value of ¥10,000 is
$10{,}000 \times [(1/100){\times}0.1 + (1/110) {\times}0.2 + (1/120) {\times}0.4 + (1/130) {\times}0.2 + (1/140) {\times}0.1]$
$\approx \$84.04 > \$10{,}000/120$.

51.
(a) Two free throws (expected value 1.72)
(b) Var(3pt)=1.664
(c) Two free throws: probability score 2 is 0.736 if independent.

52. (a) 3*89/128 ≈ 2.09
(b) 3*(111/112 + 109/123 + 22/41) ≈ 7.24. Assume fractions are probabilities.
(c) 37 yard kick: 3*(109/123) ≈ 2.66
Try for first down: 0.35*(0.5*3*118/119+0.5*7) ≈ 1.75
Kicking now has the higher expected points scored.

53. (a) Bell-shaped, several outliers, but not extreme.
(b) Mean=0.1435, s = 1.6686
(c) $S(A)$ = (0.1435 − 0.02)/1.6686 ≈ 0.074. (The amount is not relevant.)
(d) Same as (c)
(e) $S(M)$ = (0.1115 − 0.02)/1.0072 ≈ 0.091. McDonalds looks better.
(f) No. Sharpe ratios are sensitive to the frequency (daily vs. monthly).

54. (a) Bell-shaped, with one substantial negative outlier.
(b) Mean=0.0502, s = 1.7183
(c) $S(D)$ = (0.0502 − 0.02)/1.7183 ≈ 0.018
(d) No. $S(D)$ is the same regardless of the amount.
(e) $S(M)$ = (− 0.0124 − 0.02)/1.430 ≈ − 0.023; Disney is better.
(f) Future performance will resemble the past.

55. 4M Project Management

(a) The company may need to negotiate with labor unions that represent the employees. It may also need to budget for the costs of the labor force. The amount of labor that it is able to use also affects the time that the project will be completed.

(b) It is more useful to have a probability distribution that conveys that the weather conditions are not known; there's variation in the type of weather that the company can expect.

(c) For this aspect of the project, the key random variable is the total number of labor employees needed at the two sites. You could also say that the weather is the key random variable; once you know the weather, you know the labor needs.

(d) The following table shows the distribution for the number of labor employees.

	Winter Conditions			
	Mild	Typical	Cold	Severe
X = total number of labor employees	180	120	80	50
$p(x)$	0.3	0.4	0.2	0.1

(e) $E(X) = 0.3\times180 + 0.4\times120 + 0.2\times80 + 0.1\times50 = 123$

(f) $\text{Var}(X) = 0.3\times(180-123)^2 + 0.4\times(120-123)^2 + 0.2\times(80-123)^2 + 0.1\times(50-123)^2 = 1881$
$\text{SD}(X) \approx 43.37$

(g) On the basis of the projected weather conditions and estimates of labor needs, you estimate the total labor needs of the two projects to be about 123 laborers during the winter. The number could be rather different, however. There's a small chance (10%) that only 50 are needed, but a larger chance (30%) that as many as 180 will be needed.

56. 4M Credit Scores

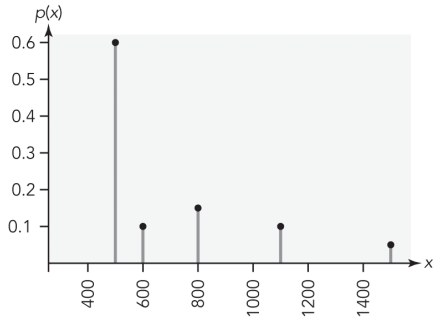
(a) He should care about the expected value of the premium for the policies that he's writing. Since he is paid by commission, selling high-value policies is worth more to him than less expensive policies. If everyone in the community had a high credit rating, he might not be able to sell enough policies to make the money he needs.

(b) He should also know the SD because it's a reminder that nothing is guaranteed. These are only probabilities, so his results will vary around the mean. The SD indicates roughly how far from the mean he could be.

(c) Let X = annual premium for a randomly selected customer. You could also work with the commission directly, but we'll handle that adjustment later.

(d) Let the random variable C denote his commission from a policy. That's determined as $C = 0.1 \times X$.

(e) This graph shows the probability distribution of X.



(f) The expected value of X is

$$500 \times 0.6 + 600 \times 0.1 + 800 \times 0.15 + 1100 \times 0.10 + 1500 \times 0.05 = \$665.$$

Using the rule for multiplying a random variable by a constant, $E(C) = 0.10 \times \$665 = \$66.5$.

(g) The variance of X is

$$(500 - 665)^2 \times 0.6 + (600 - 665)^2 \times 0.1 + (800 - 665)^2 \times 0.15 + (1100 - 665)^2 \times 0.10 + (1500 - 665)^2 \times 0.05 = 73{,}275$$

so that the SD $\approx \$271$. The SD of the commission is then $SD(C) = 0.1 \times SD(X) \approx \$27$.

(h) The salesman can expect to earn on average about \$66 for each policy that he sells. That's not locked in; there's considerable variation in the amount that he will earn for each policy. Directly from the probability distribution (you think the insurance salesman wants to hear about SDs?), he can expect 60% of the policies he sells to earn him \$50. Some pay much more. He can expect 15% of the policies he sells – to those risky customers – to pay more than \$100. These are few, so his average earnings are \$66.50.

(i) We're sure you can think of lots of things to tell the insurance company. One that is particularly interesting in this example often goes by the name principal-agent problem. An employee often has different incentives than the company has as a whole. Because he's paid by commission and hence earns the most by finding the riskiest customers, managers at the insurance company should not be surprised if they find he's writing more than 5% of his policies to the most risky group. He earns \$150 for each of these he sells, but only \$50 for the group with the best rating.