

INTRODUCTION

We learn statistics by doing statistics. Working exercises is therefore perhaps the most important aspect of studying statistics. *The Basic Practice of Statistics* (BPS) offers a large number of exercises. Most follow these principles:

- Use real settings and real data.
- Ask questions that lead to a conclusion in the real setting. A number, a graph, or “reject H_0 ” is not a full solution to an exercise in statistics.

That is, BPS attempts to illustrate how statistical methods are used in real settings, subject of course to the constraints of a text intended for beginning students.

New exercises

This supplement to BPS provides **147 additional exercises**. Of these, 109 are traditional exercises that attempt to follow the principles just stated. Students who use software can download the new data sets, as well as data sets for BPS itself, from the BPS companion Web site at www.whfreeman.com/bps.

Applets and applet exercises

Since the publication of BPS, W. H. Freeman has developed and made available online a set of interactive **statistical applets** that automate calculations and graphics in a way that can greatly assist learning. To use the applets, go to www.whfreeman.com/bps, click on “Statistical applets,” and enter the password **bpsapplets**.

The applets are excellent tools for learning. No amount of written exposition, lecturing, or blackboard sketching can demonstrate (for example) the danger of influential points in regression or the behavior of confidence intervals as clearly as interactive animation. I have included here **38 applet exercises** that guide use of the applets to gain understanding of statistical ideas. I hope that students will explore the applets on their own after completing applet exercises. Among the applets you will also find straightforward tools that for some purposes can replace calculators and statistical software. I have included notes that comment on the usefulness of applets for the relevant chapters of BPS.

An additional chapter

Finally, this supplement contains an **additional chapter** of BPS, which introduces “non-parametric” tests for use in small samples of data which clearly violate the requirements for use of common inference procedures based on normal distributions.

David S. Moore

THE BASIC PRACTICE OF STATISTICS

ADDITIONAL EXERCISES

CHAPTER 1 EXERCISES

1.1 Mutual funds. Here is a small part of a data set that describes mutual funds available to the public:

Fund	Category	Net assets (\$million)	Year to date return	Largest holding
:				
Fidelity Low-Priced Stock	Small value	6,189	4.56%	Dallas Semiconductor
Price International Stock	International stock	9,745	-0.45%	Vodafone
Vanguard 500 Index	Large blend	89,394	3.45%	General Electric
:				

- What individuals does this data set describe?
- In addition to the fund's name, how many variables does the data set contain? Which of these variables are categorical and which are quantitative?
- What are the units of measurement for each of the quantitative variables?

1.2 House prices. The National Association of Realtors reports that the “average” sales price for existing single-family homes sold in 2000 was either \$139,100 or \$177,000, depending on which “average” we use.¹ Which of these numbers is the mean price and which is the median? How do you know?

1.3 A big toe problem. Hallux abducto valgus (call it HAV) is a deformation of the big toe that is not common in youth and often requires surgery. Doctors used X-rays to measure the angle (in degrees) of deformity in 38 consecutive patients under the age of 21 who came to a medical center for surgery to correct HAV. The angle is a measure of the seriousness of the deformity. Here are the data.² (The data set is `E01-03.dat`.)

28 32 25 34 38 26 25 18 30 26 28 13 20
 21 17 16 21 23 14 32 25 21 22 20 18 26
 16 30 30 20 50 25 26 28 31 38 32 21

Make a graph and give a numerical description of this distribution. Are there any outliers? Write a brief discussion of the shape, center, and spread of the angle of deformity among young patients needing surgery for this condition.

1.4 More on a big toe problem. The HAV angle data in the previous problem contain one high outlier. Calculate the median, the mean, and the standard deviation for the full data set and also for the 37 observations remaining when you remove the outlier. How strongly does the outlier affect each of these measures?

1.5 Returns on common stocks. The total return on a stock is the change in its market price plus any dividend payments made. Total return is usually expressed as a percent of the beginning price. Figure 1 is a histogram of the distribution of the monthly returns for all stocks listed on U.S. markets for the years 1951 to 2000 (600 months).³ The low outlier is the market crash of October, 1987, when stocks lost more than 22% of their value in one month.

- Describe the overall shape of the distribution of monthly returns.
- What is the approximate center of this distribution? (For now, take the center to be the value with roughly half the months having lower returns and half having higher returns.)
- Approximately what were the smallest and largest total returns, leaving out the outlier? (This describes the spread of the distribution.)
- A return less than zero means that stocks lost value in that month. About what percent of all months had returns less than zero?

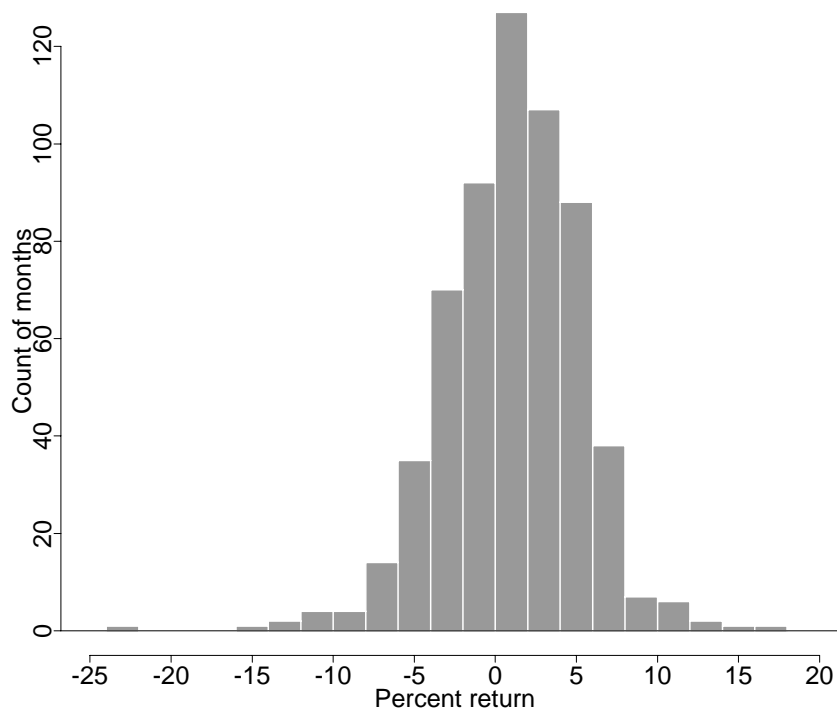


Figure 1: Monthly returns on common stocks, 1950–2000

1.6 You create the data. Create a set of 5 positive numbers (repeats allowed) that have median 10 and mean 7. What thought process did you use to create your numbers?

1.7 Where are the doctors? Table 1 gives the number of medical doctors per 100,000 people in each state. (The data set is `E01-07.dat`.)

- Why is the number of doctors per 100,000 people a better measure of the availability of health care than a simple count of the number of doctors in a state?

Table 1: Medical doctors per 100,000 population, by state (1998)

State	Doctors	State	Doctors	State	Doctors
Alabama	194	Louisiana	239	Ohio	230
Alaska	160	Maine	214	Oklahoma	166
Arizona	200	Maryland	362	Oregon	221
Arkansas	185	Massachusetts	402	Pennsylvania	282
California	244	Michigan	218	Rhode Island	324
Colorado	234	Minnesota	247	South Carolina	201
Connecticut	344	Mississippi	156	South Dakota	177
Delaware	230	Missouri	225	Tennessee	242
Florida	232	Montana	188	Texas	196
Georgia	204	Nebraska	213	Utah	197
Hawaii	252	Nevada	169	Vermont	288
Idaho	150	New Hampshire	230	Virginia	233
Illinois	253	New Jersey	287	Washington	229
Indiana	192	New Mexico	209	West Virginia	210
Iowa	171	New York	375	Wisconsin	224
Kansas	202	North Carolina	225	Wyoming	167
Kentucky	205	North Dakota	219	D.C.	702

(b) Make a graph that displays the distribution of M.D.s per 100,000 people. Write a brief description of the distribution. Are there any outliers? If so, can you explain them?

1.8 Where are the doctors, continued. Table 1 gives the number of medical doctors per 100,000 people in each state. Your graph of the distribution shows that the District of Columbia (D.C.) is a high outlier. Because D.C. is a city rather than a state, we will omit it here.

(a) Calculate both the five-number summary and \bar{x} and s for the number of doctors per 100,000 people in the 50 states. Based on your graph, which description do you prefer?

(b) What facts about the distribution can you see in the graph that the numerical summaries don't reveal? Remember that measures of center and spread are not complete descriptions of a distribution.

1.9 How much oil? How much oil wells in a given field will ultimately produce is key information in deciding whether to drill more wells. Here are the estimated total amounts of oil recovered from 64 wells in the Devonian Richmond Dolomite area of the Michigan basin, in thousands of barrels.⁴ (The data set is `E01-09.dat`.)

21.71	53.2	46.4	42.7	50.4	97.7	103.1	51.9
43.4	69.5	156.5	34.6	37.9	12.9	2.5	31.4
79.5	26.9	18.5	14.7	32.9	196	24.9	118.2
82.2	35.1	47.6	54.2	63.1	69.8	57.4	65.6
56.4	49.4	44.9	34.6	92.2	37.0	58.8	21.3
36.6	64.9	14.8	17.6	29.1	61.4	38.6	32.5
12.0	28.3	204.9	44.5	10.3	37.7	33.7	81.1
12.1	20.1	30.5	7.1	10.1	18.0	3.0	2.0

- (a) Graph the distribution and describe its main features.
- (b) Find the mean and median of the amounts recovered. Explain how the relationship between the mean and the median reflects the shape of the distribution.
- (c) Give the five-number summary and explain briefly how it reflects the shape of the distribution.

1.10 NCAA rules for athletes. The National Collegiate Athletic Association (NCAA) requires Division I athletes to score at least 820 on the combined mathematics and verbal parts of the SAT exam in order to compete in their first college year. (Higher scores are required for students with poor high school grades.) In 1999, the scores of the more than one million students taking the SATs were approximately normal with mean 1017 and standard deviation 209. What percent of all students had scores less than 820?

1.11 More NCAA rules. The NCAA considers a student a “partial qualifier” eligible to practice and receive an athletic scholarship, but not to compete, if the combined SAT score is at least 720. Use the information in the previous exercise to find the percent of all SAT scores that are less than 720.

1.12 Grading managers. Many companies “grade on a bell curve” to compare the performance of their managers and professional workers. This forces the use of some low performance ratings, so that not all workers are listed as “above average.” Ford Motor Company’s “performance management process,” for example, assigns 10% A grades, 80% B grades, and 10% C grades to the company’s 18,000 managers.⁵ It isn’t clear that companies really use normal distributions for their “bell curves.” Suppose that Ford’s evaluations range from 0 to 100 and do follow a normal distribution. What are the mean and standard deviation of this distribution?

1.13 High scores on the SAT. It is possible to score higher than 800 on the SAT, but scores above 800 are reported as 800. (That is, a student can get a reported score of 800 without a perfect paper.) In 2000, the scores of men on the math part of the SAT followed a normal distribution with mean 533 and standard deviation 115. What percent of scores were above 800 (and so reported as 800)?

One-variable calculator applet exercises

The interactive applets for *The Basic Practice of Statistics* are found on the BPS companion Web site, www.whfreeman.com/bps. You can use the one-variable statistical calculator in place of a calculator or software to do both calculations (\bar{x} and s and the five-number summary) and graphs (histograms and stemplots). The applet is more convenient than most calculators. It is less convenient than good software because, depending on your browser, it may be difficult to read in new data sets in one operation and to print output.

1.14 A big toe problem. Exercise 1.3 gives data on the angle of deformity for 38 young patients who require surgery to correct a deformity of their big toes. Enter these data into the calculator applet. Use the applet to do Exercise 1.3. (As a check on your data entry,

there should be 38 observations with mean $\bar{x} = 25.421$. You can edit entries in the data box if you mistyped an observation.)

1.15 How histograms behave. The data set menu that accompanies the applet includes the oil well production data in Exercise 1.9. Choose these data, then Click on the “Histogram” tab to see a histogram.

(a) How many classes does the applet choose to use? (You can click on the graph outside the bars to get a count of classes.)

(b) Click on the graph and drag to the left. What is the smallest number of classes you can get? What are the lower and upper bounds of each class? (Click on the bar to find out.) Make a rough sketch of this histogram.

(c) Click and drag to the right. What is the greatest number of classes you can get? How many observations does the largest class have?

(d) You see that the choice of classes changes the appearance of a histogram. Drag back and forth until you get the histogram you think best displays the distribution. How many classes did you use?

Mean and median applet exercises

1.16 Place two observations on the line by clicking below it. Why does only one arrow appear?

1.17 Place three observations on the line by clicking below it, two close together near the center of the line, and one somewhat to the right of these two.

(a) Pull the single right-most observation out to the right. (Place the cursor on the point, hold down a mouse button and drag the point.) How does the mean behave? How does the median behave? Explain briefly why each measure acts as it does.

(b) Now drag the single point to the left as far as you can. What happens to the mean? What happens to the median as you drag this point past the other two (watch carefully)?

1.18 Place 5 observations on the line by clicking below it.

(a) Add one additional observation *without changing the median*. Where is your new point?

(b) Use the applet to convince yourself that when you add yet another observation (there are now 7 in all), the median does not change no matter where you put the 7th point. Explain why this must be true.

Normal curve applet exercises

The applet allows you to do normal calculations quickly. It is somewhat limited by the number of pixels available for use, so that it can't hit every value exactly. In the exercises below, use the closest available values. In each case, *make a sketch* of the curve from the applet marked with the values you used to answer the questions asked.

1.19 The 68–95–99.7 rule for normal distributions is a useful approximation. To see how accurate the rule is, drag one flag across the other so that the applet shows the area under the curve between the two flags.

(a) Place the flags one standard deviation on either side of the mean. What is the area between these two values? What does the 68-95-99.7 rule say this area is?

(b) Repeat for locations two and three standard deviations on either side of the mean. Again compare the 68-95-99.7 rule with the area given by the applet.

1.20 How many standard deviations above and below the mean do the quartiles of any normal distribution lie? (Use the standard normal distribution to answer this question.)

1.21 In Exercise 1.12, we saw that Ford Motor Company grades its managers in such a way that the top 10% receive an A grade, the bottom 10% a C, and the middle 80% a B. Let's suppose that performance scores follow a normal distribution. How many standard deviations above and below the mean do the A/B and B/C cutoffs lie? (Use the standard normal distribution to answer this question.)

1.22 The average performance of women on the SAT, especially the math part, is lower than that of men. The reasons for this gender gap are controversial. In 2000, women's scores on the math SAT followed a normal distribution with mean 498 and standard deviation 109. The mean for men was 533. What percent of women scored higher than the male mean?

1.23 Changing the mean of a normal distribution by a moderate amount can greatly change the percent of observations in the tails. Suppose that a college is looking for applicants with SAT math scores 750 and above.

(a) In 2000, the scores of men on the math SAT followed a normal distribution with mean 533 and standard deviation 115. What percent of men scored 750 or better?

(b) Women's scores that year had a normal distribution with mean 498 and standard deviation 109. What percent of women scored 750 or better? You see that the percent of men above 750 is almost three times the percent of women with such high scores.

CHAPTER 2 EXERCISES

2.1 Age and income. How do the incomes of working-age people change with age? Because many older women have been out of the labor force for much of their lives, we look only at men between the ages of 25 and 65. Because education strongly influences income, we look only at men who have a bachelor’s degree but no higher degree. A government sample survey tells us the age and income of a random sample of 5712 such men.⁶ Figure 2 is a scatterplot of these data. Here is software output for regressing income on age. The line in the scatterplot is the least-squares regression line from this output.

	Coefficients	Standard Error	t Stat	P-value
Intercept	24874.3745	2637.4198	9.4313	5.75E-21
AGE	892.1135	61.7639	14.4439	1.79E-46

- (a) The scatterplot in Figure 2 has a distinctive form. Why do the points fall into vertical stacks?
- (b) Give some reasons why older men in this population might earn more than younger men. Give some reasons why younger men might earn more than older men. What do the data show about the relationship between age and income in the sample? Is the relationship very strong?
- (c) What is the equation of the least-squares line for predicting income from age? What specifically does the slope of this line tell us?

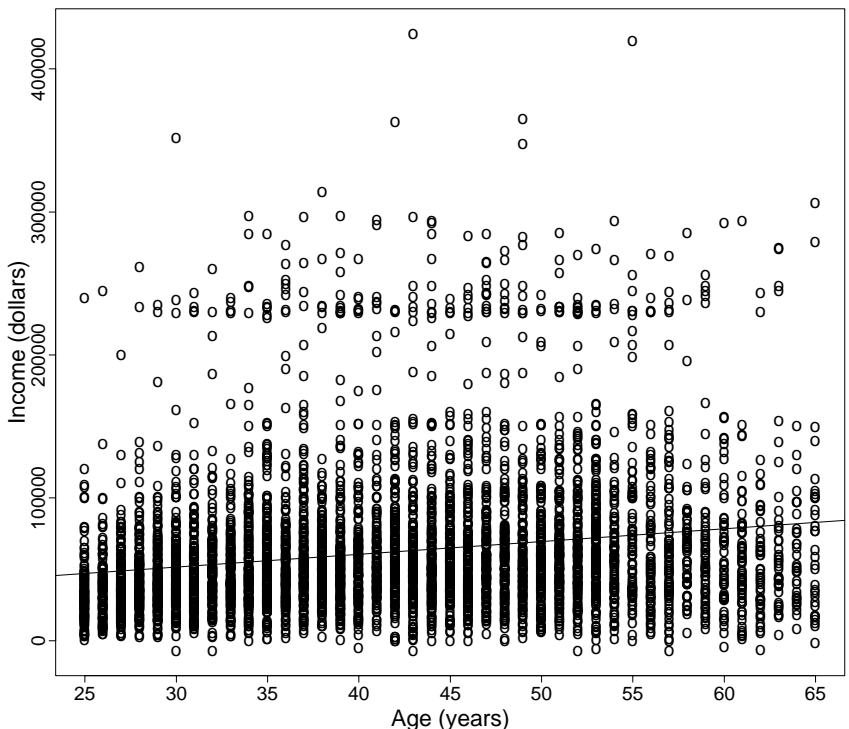


Figure 2: Age and income for 5712 men.

Table 2: Angle of deformity (degrees) for two types of foot deformity

HAV angle	MA angle	HAV angle	MA angle	HAV angle	MA angle
28	18	21	15	16	10
32	16	17	16	30	12
25	22	16	10	30	10
34	17	21	7	20	10
38	33	23	11	50	12
26	10	14	15	25	25
25	18	32	12	26	30
18	13	25	16	28	22
30	19	21	16	31	24
26	10	22	18	38	20
28	17	20	10	32	37
13	14	18	15	21	23
20	20	26	16		

2.2 Foot problems. Metatarsus adductus (call it MA) is a turning in of the front part of the foot that is common in adolescents and usually corrects itself. Hallux abducto valgus (call it HAV) is a deformation of the big toe that is not common in youth and often requires surgery. Perhaps the severity of MA can help predict the severity of HAV. Table 2 gives data on 38 consecutive patients who came to a medical center for HAV surgery.⁷ (The data set is E02-02.dat.) Using X-rays, doctors measured the angle of deformity for both MA and HAV. They speculated that there is a positive correlation—more serious MA is associated with more serious HAV.

- Make a scatterplot of the data in Table 2. (Which is the explanatory variable?)
- Describe the form, direction, and strength of the relationship between MA angle and HAV angle. Are there any clear outliers in your graph?
- Do you think the data confirm the doctors' speculation?

2.3 Foot problems, continued.

- Find the equation of the least-squares regression line for predicting HAV angle from MA angle. Add this line to the scatterplot you made in the previous problem.
- A new patient has MA angle 25 degrees. What do you predict this patient's HAV angle to be?
- Does knowing MA angle allow doctors to predict HAV angle accurately? Explain your answer from the scatterplot, then calculate a numerical measure to support your finding.

2.4 Moving in step? One reason to invest abroad is that markets in different countries don't move in step. When American stocks go down, foreign stocks may go up. So an investor who holds both bears less risk. That's the theory. Now we read that "The correlation between changes in American and European share prices has risen from 0.4 in the mid-1990s to 0.8 in 2000."⁸ Explain to an investor who knows no statistics why this fact reduces the protection provided by buying European stocks.

2.5 Interpreting correlation. The same article that claims that the correlation between changes in stock prices in Europe and the United States was 0.8 in 2000 goes on to say that “Crudely, that means that movements on Wall Street can explain 80% of price movements in Europe.” Is this true? What is the correct percent explained if $r = 0.8$?

2.6 Stocks and bonds. How is the flow of investors’ money into stock mutual funds related to the flow of money into bond mutual funds? Here are data on the net new money flowing into stock and bond mutual funds in the years 1985 to 2000, in billions of dollars.⁹ (The data set is `E02-06.dat`.) “Net” means that funds flowing out are subtracted from those flowing in. If more money leaves than arrives, the net flow will be negative. To eliminate the effect of inflation, all dollar amounts are in “real dollars” with constant buying power equal to that of a dollar in the year 2000.

Year	1985	1986	1987	1988	1989	1990	1991	1992
Stocks	12.8	34.6	28.8	-23.3	8.3	17.1	50.6	97.0
Bonds	100.8	161.8	10.6	-5.8	-1.4	9.2	74.6	87.1
Year	1993	1994	1995	1996	1997	1998	1999	2000
Stocks	151.3	133.6	140.1	238.2	243.5	165.9	194.3	309.0
Bonds	84.6	-72.0	-6.8	3.3	30.0	79.2	-6.2	-48.0

- Make a scatterplot with cash flow into stock funds as the explanatory variable.
- Find the least-squares line for predicting net bond investments from net stock investments. Add this line to your plot.
- What do the data suggest?

2.7 Illiteracy. Literacy is an important contributor to the economic and social development of a country. The data set `E02-07.dat` gives the percent of men and women at least 15 years old who were illiterate in 138 developing nations in the year 2000.¹⁰ Use software to analyze these data.

- Make a scatterplot of female illiteracy versus male illiteracy. Because schooling for women often lags behind schooling for men, we take the percent of illiterate males as our explanatory variable. Describe the form, direction, and strength of the relationship, giving a numerical measure of the strength.
- Find the equation of the least-squares regression line for predicting female illiteracy from male illiteracy and draw this line on your plot. Which countries have the largest positive residual (female illiteracy is higher than predicted) and negative residual (female illiteracy is lower than predicted)? Which countries have the highest rates of illiteracy for both men and women?

2.8 Mutual fund performance. Many mutual funds compare their performance with that of a benchmark, an index of the returns on all securities of the kind the fund buys. The Vanguard International Growth Fund, for example, takes as its benchmark the Morgan Stanley EAFE (Europe, Australasia, Far East) index of overseas stock market performance. Here are the percent returns for the fund and for the EAFE from 1982 (the first full year of the fund’s existence) to 2000.¹¹ (The data set is `E02-08.dat`.)

Year	Fund	EAFE	Year	Fund	EAFE
1982	5.27	-0.86	1992	-5.79	-11.85
1983	43.08	24.61	1993	44.74	32.94
1984	-1.02	7.86	1994	0.76	8.06
1985	56.94	56.72	1995	14.89	11.55
1986	56.71	69.94	1996	14.65	6.36
1987	12.48	24.93	1997	4.12	2.06
1988	11.61	28.59	1998	16.93	20.33
1989	24.76	10.80	1999	26.34	27.30
1990	-12.05	-23.20	2000	-8.60	-13.96
1991	4.74	12.50			

(a) Make a scatterplot suitable for predicting fund returns from EAFE returns. Is there a clear straight-line pattern? How strong is this pattern? (Give a numerical measure.) Are there any extreme outliers from the straight-line pattern?

(b) Find the equation of the least-squares regression line for predicting fund return from EAFE return. In a year when overseas markets as a group return 10% (as measured by the EAFE), what do you predict to be the return for the fund?

Two-variable calculator applet exercises

The interactive applets for *The Basic Practice of Statistics* are found on the BPS companion Web site, www.whfreeman.com/bps. You can use the two-variable statistical calculator in place of a calculator or software to do both calculations (means and standard deviations, correlation, least-squares line) and graphs (scatterplot and residual plot). The applet is more convenient than most calculators. It is less convenient than good software because, depending on your browser, it may be difficult to read in new data sets in one operation and to print output.

2.9 Illiteracy. The data set menu that accompanies the two-variable calculator includes the male and female illiteracy rates for 138 countries, described in Exercise 2.7. Choose this data set and use the calculator to do Exercise 2.7.

2.10 Mutual fund performance. Exercise 2.8 gives data on the percent return for the Vanguard International Growth Fund and its benchmark index of overseas stock market performance, the Morgan Stanley EAFE, for 19 years. Enter these data into the calculator applet and use the applet to do Exercise 2.8. (As a check on your data entry, the correlation should be $r = 0.8983$. You can edit entries in the data box.)

Correlation and regression applet exercises

This interactive applet shows how the correlation r and the least-squares regression line respond to changes in a scatterplot. No amount of talking or writing can show so clearly how these statistical measures behave. The following exercises point to some important

facts, but I hope that you will experiment a bit to gain some feeling for correlation and regression. To erase an entire scatterplot and start over, click on the trash can.

2.11 Match the correlation. You are going to make scatterplots with 10 points that have correlation close to 0.7. The lesson is that many patterns can have the same correlation. Always plot your data before you trust a correlation.

(a) Stop after adding the first two points. What is the value of the correlation? Why does it have this value?

(b) Make a lower left to upper right pattern of 10 points with correlation about $r = 0.7$. (You can drag points up or down to adjust r after you have 10 points.) Make a rough sketch of your scatterplot.

(c) Make another scatterplot with 9 points in a vertical stack at the left of the plot. Add one point far to the right and move it until the correlation is close to 0.7. Make a rough sketch of your scatterplot.

(d) Make yet another scatterplot with 10 points in a curved pattern that starts at the lower left, rises to the right, then falls again at the far right. Adjust the points up or down until you have a quite smooth curve with correlation close to 0.7. Make a rough sketch of this scatterplot also.

2.12 Is regression useful? In the previous exercise, you created three scatterplots having correlation about $r = 0.7$ between the horizontal variable x and the vertical variable y . Correlation $r = 0.7$ is considered reasonably strong in many areas of scientific work. Because there is a reasonably strong correlation, we might use a regression line to predict y from x . In which of your three scatterplots does it make sense to use a straight line for prediction?

2.13 Influence on correlation. Click on the scatterplot to create a group of 10 points in the lower left corner of the scatterplot with a strong straight-line pattern (correlation about 0.9).

(a) Add one point at the upper right that is in line with the first 10. How does the correlation change?

(b) Drag this last point down until it is opposite the group of 10 points. How small can you make the correlation? Can you make the correlation negative? You see that a single outlier can greatly strengthen or weaken a correlation. Always plot your data to check for outlying points.

2.14 Influence in regression. As in the previous exercise, create a group of 10 points in the lower left corner of the scatterplot with a strong straight-line pattern (correlation at least 0.9). Click the “Show least-squares line” box to display the regression line.

(a) Add one point at the upper right that is far from the other 10 points but exactly on the regression line. Why does this outlier have no effect on the line even though it changes the correlation?

(b) Now drag this last point down until it is opposite the group of 10 points. You see that one end of the least-squares line chases this single point, while the other end remains near the middle of the original group of 10. What about the last point makes it so influential?

2.15 Guessing a regression line. Click on the scatterplot to create a group of 15 to 20 points from lower left to upper right with a clear positive straight-line pattern (correlation

around 0.7). Click the “Draw line” button and use the mouse (right-click and drag) to draw a line through the middle of the cloud of points from lower left to upper right. Note the “thermometer” above the plot. The red portion is the sum of the squared vertical distances from the points in the plot to the least-squares line. The green portion is the “extra” sum of squares for your line—it shows by how much your line misses the smallest possible sum of squares.

(a) You drew a line by eye through the middle of the pattern. Yet the right-hand part of the bar is solid green. What does that tell you?

(b) Now click the “Show least-squares line” box. Is the slope of the least-squares line smaller (the new line is less steep) or larger (line is steeper) than that of your line? If you repeat this exercise several times, you will consistently get the same result. The least-squares line minimizes the *vertical* distances of the points from the line. It is *not* the line through the “middle” of the cloud of points. This is one reason why it is hard to draw a good regression line by eye.

CHAPTER 3 EXERCISES

3.1 Instant opinion. The Harris/Excite instant poll can be found online at news.excite.com/news/poll. The question appears on the screen, and you simply click buttons to vote “Yes,” “No,” or “Don’t Know.” On January 25, 2000, the question was “Should female athletes be paid the same as men for the work they do?” In all, 13,147 (44%) said “Yes,” another 15,182 (50%) said “No,” and the remaining 1448 said “Don’t know.”

- What is the sample size for this poll?
- That’s a much larger sample than standard sample surveys. In spite of this, we can’t trust the result to give good information about any clearly defined population. Why?
- It is still true that more men than women use the Web. How might this fact affect the poll results?

3.2 Dealing with pain. Health care providers are giving more attention to relieving the pain of cancer patients. An article in the journal *Cancer* surveyed a number of studies and concluded that controlled-release morphine tablets, which release the pain killer gradually over time, are more effective than giving standard morphine when the patient needs it.¹² The “methods” section of the article begins: “Only those published studies that were controlled (i.e., randomized, double blind, and comparative), repeated-dose studies with CR morphine tablets in cancer pain patients were considered for this review.” Explain the terms in parentheses to someone who know nothing about medical trails.

3.3 Protecting ultramarathon runners. An ultramarathon, as you might guess, is a foot race longer than the 26.2 miles of a marathon. Runners commonly develop respiratory infections after an ultramarathon. Will taking 600 milligrams of vitamin C daily reduce these infections? Researchers randomly assigned ultramarathon runners to receive either vitamin C or a placebo. Separately, they also randomly assigned these treatments in a group of non-runners the same age as the runners. All subjects were watched for 14 days after the big race to see if infections developed.¹³

- What is the name for this experimental design?
- Use a diagram to outline the design.

3.4 Exercising to lose weight. We all know that regular exercise (combined with a sensible diet) is a key to shedding those extra pounds. Experience shows that overweight people find it tough to keep exercising. Perhaps they will do better with several short sessions each day rather than one longer session. Perhaps having exercise equipment at home will help. An experiment looked at these issues. The subjects were women aged 25 to 45 whose weights were 20% to 75% higher than ideal. The study report says:¹⁴

Subjects were randomly assigned to 1 of 3 groups. All groups were prescribed a similar volume of exercise. The 3 groups differed in the way the exercise was prescribed. . . .

Long-Bout Exercise Group *Forty-nine subjects were instructed to exercise 5 d/wk; duration progressed from 20 min/d . . . to 40 min/d . . . Participants performed the exercise in one long bout.*

Short-Bout Exercise Group *Fifty-one subjects were instructed to exercise 5 d/wk . . . However, rather than exercising continuously for the prescribed duration, subjects were instructed to divide the exercise into multiple 10-minute bouts that were performed at convenient times throughout the day.*

Short-Bout Plus Exercise Equipment Group *The exercise prescription was identical to the exercise prescribed for the short-bout group . . . The 48 subjects in this group were also provided with motorized home treadmills.*

The researchers recorded weight, fitness, and whether the subject continued the exercise program.

- (a) Use a diagram to outline the design of this experiment.
- (b) How many subjects are there in all? Use Table A starting at line 114 to choose the first 10 subjects for the long-bout group.

3.5 A telephone survey. The 1998 National Gun Policy Survey, carried out by the University of Chicago's National Opinion Research Center, asked respondents' opinions about government regulation of firearms. A report from the survey says, "Participating households were identified through random-digit dialing; the respondent in each household was selected by the most-recent-birthday method."¹⁵

- (a) What is "random-digit dialing?" Why is it a practical method for obtaining (almost) an SRS of households?
- (b) The survey wants the opinion of an individual adult. Several adults may live in a household. In that case, the survey interviewed the adult with the most recent birthday. Why is this preferable to simply interviewing the person who answers the phone?

3.6 Are you sure? Late in 1996, Spain's Centro de Investigaciones Sociológicas carried out a sample survey on the attitudes of Spaniards toward private business and state intervention in the economy.¹⁶ Of the 2496 adults interviewed, 72% agreed that, "Employees with higher performance must get higher pay." On the other hand, 71% agreed that, "Everything a society produces should be distributed among its members as equally as possible and there should be no major differences." Use these conflicting results as an example in a short explanation of why opinion polls often fail to reveal public attitudes clearly.

3.7 Reducing risky sex. The National Institutes of Mental Health (NIMH) wants to know whether intense education about the risks of AIDS will help change the behavior of people who now report sexual activities that put them at risk of infection. NIMH investigators screened 38,893 people to identify 3706 suitable subjects. The subjects were assigned to a control group (1855 people) or an intervention group (1851 people). The control group attended a one-hour AIDS education session; the intervention group attended seven single-sex discussion sessions, each lasting 90 to 120 minutes. After 12 months, 64% of the intervention group and 52% of the control group said they used condoms. (None of the subjects used condoms regularly before the study began.)¹⁷

- (a) Because none of the subjects used condoms when the study started, we might just offer the intervention sessions and find that 64% used condoms 12 months after the sessions. Explain why this greatly overstates the effectiveness of the intervention.
- (b) Outline the design of this experiment.

(c) You must randomly assign 3706 subjects. How would you label them? Use line 119 of Table B to choose the first 5 subjects for the intervention group.

3.8 Growing trees faster. The concentration of carbon dioxide (CO_2) in the atmosphere is increasing rapidly due to our use of fossil fuels. Because plants use CO_2 to fuel photosynthesis, more CO_2 may cause trees and other plants to grow faster. An elaborate apparatus allows researchers to pipe extra CO_2 to a 30-meter circle of forest. We want to compare the growth in base area of trees in treated and untreated areas to see if extra CO_2 does in fact increase growth. We can afford to treat three circular areas.¹⁸

(a) Describe the design of a completely randomized experiment using 6 well-separated 30-meter circular areas in a pine forest. Sketch the circles and carry out the randomization your design calls for.

(b) Areas within the forest may differ in soil fertility. Describe a matched pairs design using three pairs of circles that will reduce the extra variation due to different fertility. Sketch the circles and carry out the randomization your design calls for.

3.9 Keeping warm during surgery (EESSE). Surgery patients are often cold because the operating room is kept cool and the body's temperature regulation is disturbed by anesthetics. Will warming patients to maintain normal body temperature reduce infections after surgery? In one experiment, patients undergoing colon/rectal surgery received intravenous fluids from a warming machine and were covered with a blanket through which air circulated. In some patients, the fluid and the air were warmed; in others, they were not. The patients received identical treatment in all other respects.¹⁹

(a) To simplify the setup of the study, we might warm the fluids and air blanket for one operating team and not for another doing the same kind of surgery. Why might this design result in bias?

(b) Outline the design of a randomized comparative experiment for this study.

(c) The operating team did not know whether fluids and air blanket were heated, nor did the doctors who followed the patients after surgery. What is this practice called? Why was it used here?

Simple random sample applet exercises

The interactive applets for *The Basic Practice of Statistics* are found on the BPS companion Web site, www.whfreeman.com/bps. The simple random sample applet can choose an SRS of any size up to $n = 40$ from a population of any size up to 500.

3.10 Sampling retail outlets. Exercise 3.9 on page 174 of BPS asks you to choose an SRS of 10 from the 440 retail outlets in New York that sell your product. Use the applet to choose this sample. Which outlets were chosen? (That was faster than using Table B.)

3.11 Testing a breakfast food. Because experimental randomization chooses SRSs of the subjects, we can use the applet here as well as for sampling problems. Example 3.12 on page 190 of BPS describes a randomized comparative experiment in which 30 rats are assigned at random to a treatment group of 15 and a control group of 15. Use the applet

to choose the 15 rats for the treatment group. Which rats did you choose? The remaining 15 rats make up the control group.

3.12 Conserving energy. The applet allows you to randomly assign subjects to more than two groups without difficulty. Example 3.13 on page 191 of BPS describes a randomized comparative experiment in which 60 houses are randomly assigned to three groups of 20.

- Use the applet to choose an SRS of 20 out of 60 houses to form the first group. Which houses are in this group?
- The “Population hopper” now contains the 40 houses that were not chosen, in scrambled order. Click “Sample” again to choose an SRS of 20 of these remaining houses to make up the second group. Which houses were chosen?
- The 20 houses remaining in the “Population hopper” form the third group. Which houses are these?

3.13 Randomization avoids bias. Suppose that the 15 even-numbered rats among the 30 rats available in the setting of Exercise 3.11 are (unknown to the experimenters) a fast-growing variety. We hope that these rats will be roughly equally distributed between the two groups. Take 10 samples of size 15 from the 30 rats. (Be sure to click “Reset” after each sample.) Record the counts of even-numbered rats in each of your 10 samples. You see that there is considerable chance variation, but no systematic bias in favor of one or the other group in assigning the fast-growing rats. Larger samples from larger population will on the average do a better job of making the two groups equivalent.

CHAPTER 4 EXERCISES

4.1 Measuring unemployment. The Bureau of Labor Statistics announces that last month it interviewed all members of the labor force in a sample of 50,000 households; **4.5%** of the people interviewed were unemployed. Is this number a parameter or a statistic? Why?

4.2 Republican voters. Voter registration records show that **68%** of all voters in Indianapolis are registered as Republicans. To test a random digit dialing device, you use the device to call 150 randomly chosen residential telephones in Indianapolis. Of the registered voters contacted, **73%** are registered Republicans. Is each of the boldface numbers a parameter or a statistic? Why?

4.3 Preparing for the GMAT. A company that offers courses to prepare would-be MBA students for the GMAT examination has the following information about its customers: 20% are currently undergraduate students in business; 15% are undergraduate students in other fields of study; 60% are college graduates who are currently employed; and 5% are college graduates who are not employed.

- (a) Is this a legitimate assignment of probabilities to customer backgrounds? Why?
- (b) What percent of customers are currently undergraduates?

4.4 Race and ethnicity. The 2000 census allowed each person to choose one or more from of a long list of races. That is, in the eyes of the Census Bureau, you belong to whatever race or races you say you belong to. “Hispanic/Latino” is a separate category; Hispanics may be of any race. If we choose a resident of the United States at random, the 2000 census gives these probabilities:

	Hispanic	Not Hispanic
Asian	0.000	0.036
Black	0.003	0.121
White	0.060	0.691
Other	0.062	0.027

- (a) Verify that this is a legitimate assignment of probabilities.
- (b) What is the probability that a randomly chosen American is Hispanic?
- (c) Non-Hispanic whites are the historical majority in the United states. What is the probability that a randomly chosen American is not a member of this group?

4.5 Tetrahedral dice. Psychologists sometimes use tetrahedral dice to study our intuition about chance behavior. A tetrahedron is a pyramid (think of Egypt) with four identical faces, each a triangle with all sides equal in length. Label the four faces of a tetrahedral die with 1, 2, 3, and 4 spots. Give a probability model for rolling such a die and recording the number of spots on the down face. Explain why you think your model is at least close to correct.

4.6 Playing “pick four.” The “pick four” games in many state lotteries announce a four-digit winning number each day. The winning number is essentially a four-digit group

from a table of random digits. You win if your choice matches the winning digits. Suppose your chosen number is 5974.

- (a) What is the probability that your number matches the winning number exactly?
- (b) What is the probability that your number matches the digits in the winning number *in any order*?

4.7 More tetrahedral dice. Tetrahedral dice are described in Exercise 4.5. Give a probability model for rolling two such dice. That is, write down all possible outcomes and give a probability to each. (Example 4.4 and Figure 4.2 in BPS may help you.) What is the probability that the sum of the down faces is 5?

4.8 Playing “pick four,” continued. The Wisconsin version of “pick four” pays out \$5000 on a \$1 bet if your number matches the winning number exactly. It pays \$200 on a \$1 bet if the digits in your number match those of the winning number in any order. You choose which of these two bets to make. On the average over many bets, your winnings will be

$$\text{mean amount won} = \text{payout amount} \times \text{probability of winning}$$

What is this mean payout for these two bets? Is one of the two bets a better choice?

4.9 An edge in “pick four.” Exercise 4.6 describes “pick four” lottery games. Some states (New Jersey, for example) use the “pari-mutual system” in which the total winnings are divided among all players who matched the winning digits. That suggests a way to get an edge. Suppose you choose to try to match the winning number exactly.

- (a) The winning number might be, for example, either 2873 or 8888. Explain why these two outcomes have exactly the same probability.
- (b) It is likely that fewer people will choose one of these numbers than the other, because it “doesn’t look random.” You prefer the less popular number because you will win more if fewer people share a winning number. Which of these two numbers do you prefer?

4.10 Polling women. Suppose that 47% of all adult women think they do not get enough time for themselves. An opinion poll interviews 1025 randomly chosen women and records the sample proportion who don’t feel they get enough time for themselves. This statistic will vary from sample to sample if the poll is repeated. The sampling distribution is approximately normal with mean 0.47 and standard deviation about 0.016. Sketch this normal curve and use it to answer the following questions.

- (a) The truth about the population is 0.47. In what range will the middle 95% of all sample results fall?
- (b) What is the probability that the poll gets a sample in which fewer than 45% say they do not get enough time for themselves?

4.11 Will you have an accident? The probability that a randomly chosen driver will be involved in an accident in the next year is about 0.2. This is based on the proportion of millions of drivers who have accidents. “Accident” includes things like crumpling a fender in your own driveway, not just highway accidents.

- (a) What do you think is your own probability of being in an accident in the next year? This is a *personal probability* that rests on your own assessment of chance, not on many

repeated trials.

- (b) Give some reasons why your personal probability might be a more accurate prediction of your “true chance” of having an accident than the probability for a random driver.
- (c) Almost everyone says their personal probability is lower than the random driver probability. Why do you think this is true?

4.12 What probability doesn’t say. The probability of a head in tossing a coin is $1/2$. This means that as we make more tosses, the *proportion* of heads will eventually get close to 0.5 . It does not mean that the *count* of heads will get close to $1/2$ the number of tosses. To see why, imagine that the proportion of heads is 0.51 in 100 tosses, 1000 tosses, 10,000 tosses, and 100,000 tosses of a coin. How many heads came up in each set of tosses? How close is the number of heads to half the number of tosses?

4.13 A sampling distribution. We want to know what percent of American adults approve of legal gambling. This population proportion p is a parameter. To estimate p , take an SRS and find the proportion \hat{p} in the sample who approve of gambling. If we take many SRSs of the same size, the proportion \hat{p} will vary from sample to sample. The distribution of its values in all SRSs is the sampling distribution of this statistic.

Figure 3 on the following page is a small population. Each circle represents an adult. The circles containing dots are people who disapprove of legal gambling, and the empty circles are people who approve. You can check that 60 of the 100 circles are empty, so in this population the proportion who approve of gambling is $p = 60/100 = 0.6$.

- (a) Label the population 00 to 99 left-to-right across the rows, starting at the top left. Use line 101 of Table B to draw an SRS of size 5. What is the proportion \hat{p} of the people in your sample who approve of gambling?
- (b) Take 9 more SRSs of size 5 (10 in all), using lines 102 to 110 of Table B, a different line for each sample. You now have 10 values of the sample proportion \hat{p} . What are they?
- (c) Because your samples have only 5 people, the only values \hat{p} can take are $0/5$, $1/5$, $2/5$, $3/5$, $4/5$, and $5/5$. That is, \hat{p} is always 0, 0.2, 0.4, 0.6, 0.8, or 1. Mark these numbers on a line and make a histogram of your 10 results by putting a bar above each number to show how many samples had that outcome. (You have begun to construct the sampling distribution of \hat{p} , though of course 10 samples is a small start.)
- (d) Taking samples of size 5 from a population of size 100 is not a practical setting, but let’s look at your results anyway. How many of your 10 samples estimated the population proportion $p = 0.6$ exactly correctly? Is the true value 0.6 roughly in the center of your sample values? Explain why 0.6 would be in the center of the sample values if you took a large number of samples.

4.14 Insurance. The idea of insurance is that we all face risks that are unlikely but carry high cost. Think of a fire destroying your home. So we form a group to share the risk: we all pay a small amount, and the insurance policy pays a large amount to those few of us whose homes burn down. An insurance company looks at the records for millions of homeowners and sees that the mean loss from fire in a year is $\mu = \$250$ per person. (Most of us have no loss, but a few lose their homes. The $\$250$ is the average loss.) The company plans to sell fire insurance for $\$250$ plus enough to cover its costs and profit. Explain clearly why it

would be stupid to sell only 12 policies. Then explain why selling thousands of such policies is a safe business.

4.15 More on insurance. In fact, the insurance company sees that in the entire population of homeowners, the mean loss from fire is $\mu = \$250$ and the standard deviation of the loss is $\sigma = \$300$. The distribution of losses is strongly right-skewed: many policies have \$0 loss, but a few have large losses. If the company sells 10,000 policies, what is the approximate probability that the average loss will be greater than \$260?

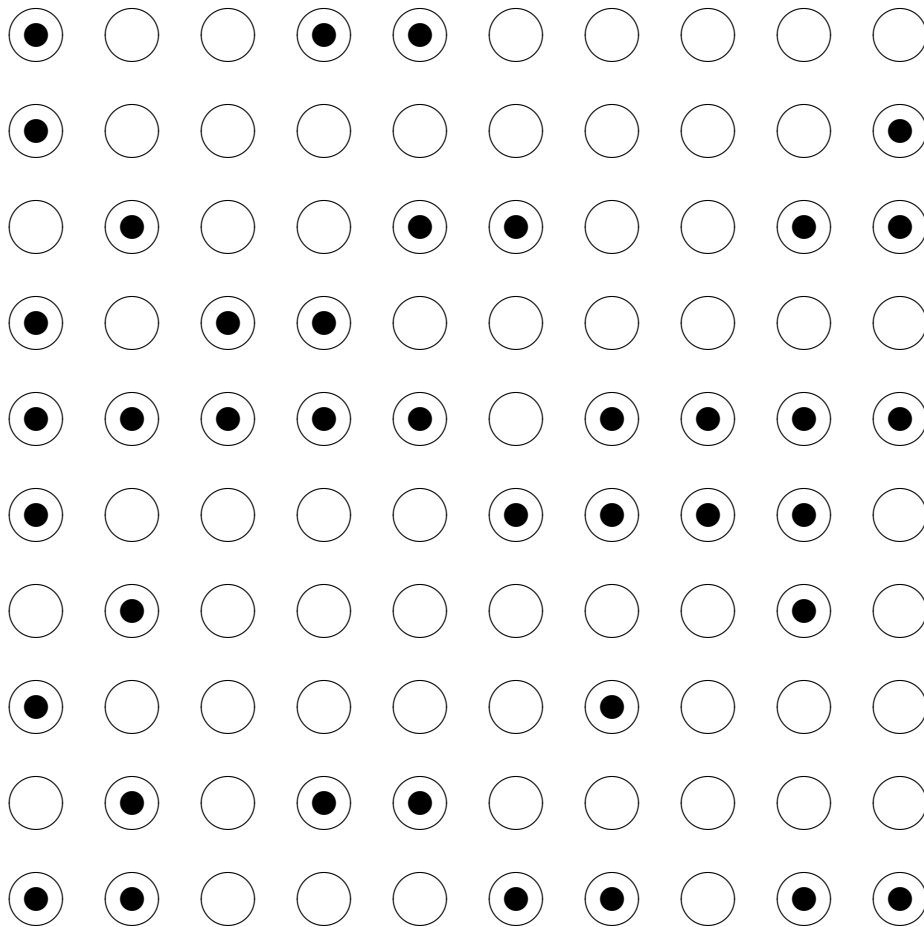


Figure 3: A population in which 60% approve of legal gambling.

Probability applet exercises

The interactive applets for *The Basic Practice of Statistics* are found on the BPS companion Web site, www.whfreeman.com/bps. The coin-tossing in the probability applet is a model for any setting with repeated independent success-or-failure trials, each of which has the same probability of a success. You can simulate up to 40 trials at once.

4.16 The nature of probability. Suppose that you toss a balanced coin very many times. To simulate this, set the “Probability of a head” in the applet to 0.5. The applet allows no more than 40 tosses at once, but you can add 40 more by clicking “Toss” again. Check the “Show true probability” box to display the probability 0.5 on the graph.

(a) Simulate 200 tosses of a coin by clicking “Toss” five times. The graph shows how the proportion of heads changes as you make more tosses. What was this proportion after 200 tosses? Make a rough sketch of the graph that displays how the proportion eventually gets close to the probability 0.5.

(b) Click “Reset” and do another 200 tosses. What was the proportion of heads in these 200 tosses? Sketch the graph again. The two graphs, representing two sets of 200 tosses, often look very different. What they have in common is that the proportion of heads eventually gets close to the probability 0.5.

4.17 What probability doesn’t say. The idea of probability is that the *proportion* of heads in many tosses of a balanced coin eventually gets close to 0.5. But does the actual *count* of heads get close to one-half the number of tosses? Let’s find out. Set the “Probability of heads” in the applet to 0.5 and the number of tosses to 40. You can extend the number of tosses by clicking “Toss” again to get 40 more. Don’t click “Reset” during this exercise.

(a) After 40 tosses, what is the proportion of heads? What is the count of heads? What is the difference between the count of heads and 20 (one-half the number of tosses)?

(b) Keep going to 120 tosses. Again record the proportion and count of heads and the difference between the count and 60 (half the number of tosses).

(c) Keep going. Stop at 240 tosses and again at 480 tosses to record the same facts. Although it may take a long time, the laws of probability say that the proportion of heads will always get close to 0.5 and also that the difference between the count of heads and half the number of tosses will always grow without limit.

4.18 Not a great bet. In Exercise 4.6 you found the probability that the winning number in a “pick-four” lottery matches the digits in your number in any order. Enter this probability of winning as the “Probability of heads” in the applet. Enter 31 as the number of tosses. This represents a bet every day for a month. Simulate a month’s play. Keep playing every day, a month at a time (click “Reset” to start a new month) until you win. You will often wait a long time to win!

4.19 A sampling distribution. You can use the probability applet to speed up and improve Exercise 4.13. You have a population in which 60% of the individuals approve of legal gambling. You want to take many small samples from this population to observe how the sample proportion who approve of gambling varies from sample to sample. Set the “Probability of heads” in the applet to 0.6 and the number of tosses to 5. This simulates

an SRS of size 5 from a very large population, not just 100 individuals as in Exercise 4.13. By alternating between “Toss” and “Reset” you can take many samples quickly. (a) Take 50 samples, recording the number of heads (that is, the number in the sample who approve of gambling) in each sample. Make a histogram of the 50 sample proportions.

(b) Another population contains only 20% who approve of legal gambling. Take 50 samples of size 5 from this population, record the number in each sample who approve, and make a histogram of the 50 sample proportions. How do the centers of your two histograms reflect the differing truths about the two populations?

Expected value applet exercise

4.20 The law of large numbers. Suppose that you roll two balanced dice and look at the spots on the up faces. There are 36 possible outcomes, displayed in Figure 4.2 on page 221 of BPS. Because the dice are balanced, all 36 outcomes are equally likely. Add the spots on the up faces. The average of the 36 totals is 7. This is the population mean μ for the idealized population that contains the results of rolling two dice forever. (The mean is also called the “expected value,” which explains the name of the applet. We do not expect to get the value μ on one roll, so the term is a bit misleading.) The law of large numbers says that the average \bar{x} from a finite number of rolls gets closer and closer to 7 as we do more and more rolls.

(a) Click “More dice” in the expected value applet once to get two dice. Click “Show mean” to see the mean 7 on the graph. Leaving the number of rolls at 1, click “Roll dice” three times. Note the count of spots for each roll (what were they?) and the average for the three rolls. You see that the graph displays at each point the average number of spots for all rolls up to the last one. Now you understand the display.

(b) Set the number of rolls to 100 and click “Roll dice.” The applet rolls the two dice 100 times. The graph shows how the average count of spots changes as we make more rolls. That is, the graph shows \bar{x} as we continue to roll the dice. Make a rough sketch of the final graph.

(c) Repeat your work from (b). Click “Reset” to start over, then roll two dice 100 times. Make a sketch of the final graph of the mean \bar{x} against the number of rolls. Your two graphs will often look very different. What they have in common is that the average eventually gets close to the population mean $\mu = 7$. The law of large numbers says that this will *always* happen if you keep on rolling the dice.

4.21 What’s the mean? Suppose that you roll three balanced dice. We wonder what the mean number of spots on the up faces of the three dice is. The law of large numbers says that we can find out by experience: roll three dice many times, and the actual average number of spots will eventually approach the mean. Set up the applet to roll three dice. Don’t click “Show mean” yet. Roll the dice until you are confident you know the mean quite closely, then click “Show mean” to verify your discovery. What is the mean? Make a rough sketch of the path the averages \bar{x} followed as you kept adding more rolls.

Simple random sample applet exercise

4.22 A sampling distribution. We can use the simple random sample applet to help grasp the idea of a sampling distribution. Form a population labeled 1 to 100. We will choose an SRS of 10 of these numbers. That is, in this exercise, the numbers themselves are the population, not just labels for 100 individuals. The mean of the whole numbers 1 to 100 is $\mu = 50.5$. This is the population mean.

(a) Use the applet to choose an SRS of size 10. Which 10 numbers were chosen? What is their mean? This is the sample mean \bar{x} .

(b) Although the population and its mean $\mu = 50.5$ remain fixed, the sample mean changes as we take more samples. Take another SRS of size 10. (Use the “Reset” button to return to the original population before taking the second sample.) What are the 10 numbers in your sample? What is their mean? This is another value of \bar{x} .

(c) Take 8 more SRSs from this same population and record their means. You now have 10 values of the sample mean \bar{x} from 10 SRSs of the same size from the same population. Make a histogram of the 10 values and mark the population mean $\mu = 50.5$ on the horizontal axis. Are your 10 sample values roughly centered at the population value μ ? (If you kept going forever, your \bar{x} -values would form the sampling distribution of the sample mean; the population mean μ would indeed be the center of this distribution.)

CHAPTER 5 EXERCISES

5.1 Race and ethnicity. The 2000 census allowed each person to choose one or more from of a long list of races. That is, in the eyes of the Census Bureau, you belong to whatever race or races you say you belong to. “Hispanic/Latino” is a separate category; Hispanics may be of any race. If we choose a resident of the United States at random, the 2000 census gives these probabilities:

	Hispanic	Not Hispanic
Asian	0.000	0.036
Black	0.003	0.121
White	0.060	0.691
Other	0.062	0.027

- (a) What is the probability that a randomly chosen person is white?
- (b) You know that the person chosen is Hispanic. What is the conditional probability that this person is white?

5.2 More on race and ethnicity.

- (a) What is the probability that a randomly chosen American is Hispanic?
- (b) You know that the person chosen is black. What is the conditional probability that this person is Hispanic?

5.3 At the gym. Many conditional probability calculations are just common sense made automatic. For example, 10% of adults belong to health clubs, and 40% of these health club members go to the club at least twice a week. What percent of all adults go to a health club at least twice a week? Write the information given in terms of probabilities and use the general multiplication rule.

5.4 A hot stock. You purchase a hot stock for \$1000. The stock either gains 30% or loses 25% each day, and its behaviors on consecutive days are independent of each other. You plan to sell the stock after two days. What are the possible values of the stock after two days, and what is the probability for each value? (*Hint:* Remember that the value is multiplied by 1.30 on day of 30% increase and multiplied by 0.75 on a day of 25% loss.)

5.5 Should you invest? Consider the hot stock of Exercise 5.4.

- (a) What is the probability that the stock is worth more after two days than the \$1000 you paid for it? You see that you will usually lose money if you pay \$1000 for this stock.
- (b) The *mean value* of the stock after two days turns out to be about \$1050. The law of large numbers says that on the average over many such \$1000 investments you will come out about \$50 ahead. Make a probability histogram of the distribution of possible values from Exercise 5.4. Use what you know about the behavior of means to explain briefly why the mean is much larger than most of the possible values.

5.6 Teen-age drivers. An insurance company has the following information about drivers aged 16 to 18 years: 20% are involved in accidents each year; 10% in this age group are A students; among those involved in an accident, 5% are A students.

(a) Let A be the event that a young driver is an A student and C the event that a young driver is involved in an accident this year. State the information given in terms of probabilities and conditional probabilities for the events A and C .

(b) What is the probability that a randomly chosen young driver is an A student and is involved in an accident?

5.7 More on teen-age drivers. Use your work from Exercise 5.6 to find the percent of A students who are involved in accidents. (Start by expressing this as a conditional probability.)

5.8 Preparing for the GMAT. A company that offers courses to prepare would-be MBA students for the GMAT examination finds that 40% of its customers are currently undergraduate students and 60% are college graduates. After completing the course, 50% of the undergraduates and 70% of the graduates achieve scores of at least 600 on the GMAT.

(a) What percent of customers are undergraduates *and* score at least 600? What percent of customers are graduates *and* score at least 600?

(b) What percent of all customers score at least 600 on the GMAT?

5.9 Screening job applicants. A company retains a psychologist to assess whether job applicants are suited for assembly-line work. The psychologist classifies applicants as A (well suited), B (marginal), or C (not suited). The company is concerned about event D: an employee leaves the company within a year of being hired. Data on all people hired in the past five years gives these probabilities:

$$\begin{array}{lll} P(A) = 0.4 & P(B) = 0.3 & P(C) = 0.3 \\ P(A \text{ and } D) = 0.1 & P(B \text{ and } D) = 0.1 & P(C \text{ and } D) = 0.2 \end{array}$$

Sketch a Venn diagram of the events A, B, C, and D and mark on your diagram the probabilities of all combinations of psychological assessment and leaving (or not) within a year. What is $P(D)$, the probability that an employee leaves within a year?

Probability applet exercise

The interactive applets for *The Basic Practice of Statistics* are found on the BPS companion Web site, www.whfreeman.com/bps. The coin-tossing in the probability applet simulates independent success/failure outcomes. The random count of heads (or tails) in the specified number of tosses therefore follows a binomial distribution.

5.10 Inspecting switches. Example 5.10 on page 273 of BPS concerns the count of bad switches in inspection samples of size 10. The count has the binomial distribution with $n = 10$ and $p = 0.1$. Set these values for the number of tosses and probability of heads in the probability applet. The example calculates that the probability of getting a sample with exactly 1 bad switch is 0.3874. Of course, when we inspect only a few lots, the proportion of samples with exactly 1 bad switch will differ from this probability. Click “Toss” and “Reset” repeatedly to simulate inspecting 20 lots. Record the number of bad switches (the count of heads) in each of the 20 samples. What proportion of the 20 lots had exactly 1 bad switch? Remember that probability tells us only what happens in the long run.