Unit 12: Correlation

PREREQUISITES

Unit 10: Scatterplots is a prerequisite for this unit. However, this unit does not discuss the close relationship between correlation and regression, so Unit 11: Regression, is not a prerequisite. This also allows flexibility in presentation order between Units 11 and 12. The formula for computing the correlation coefficient makes use of both the mean and the standard deviation of the *x* and *y* values. Therefore, students should be familiar with those measures, which were covered in Unit 4: Measures of Center, and Unit 6: Standard Deviation. In addition, students should have some familiarity with summation notation.

ADDITIONAL TOPIC COVERAGE

Additional coverage on correlation can be found in *The Basic Practice of Statistics*, Chapter 4, Scatterplots and Correlation.

ACTIVITY DESCRIPTION

In the unit activity on correlation, students observe how the scatter of points about a line affects the value of *r*. Of particular importance is the discovery that even a single outlier can have a huge effect on correlation, even as extreme as turning an otherwise strong positive correlation into a negative correlation. Students should use technology – graphing calculators, spreadsheet or statistical computing software – for computation of correlation. The data sets in this activity are small. So, students can either draw the scatterplots by hand or use technology.

THE VIDEO SOLUTIONS

1. It would form a straight line. In fact, the points would all fall on the line $y = x$.
In this case, $r = 1$.

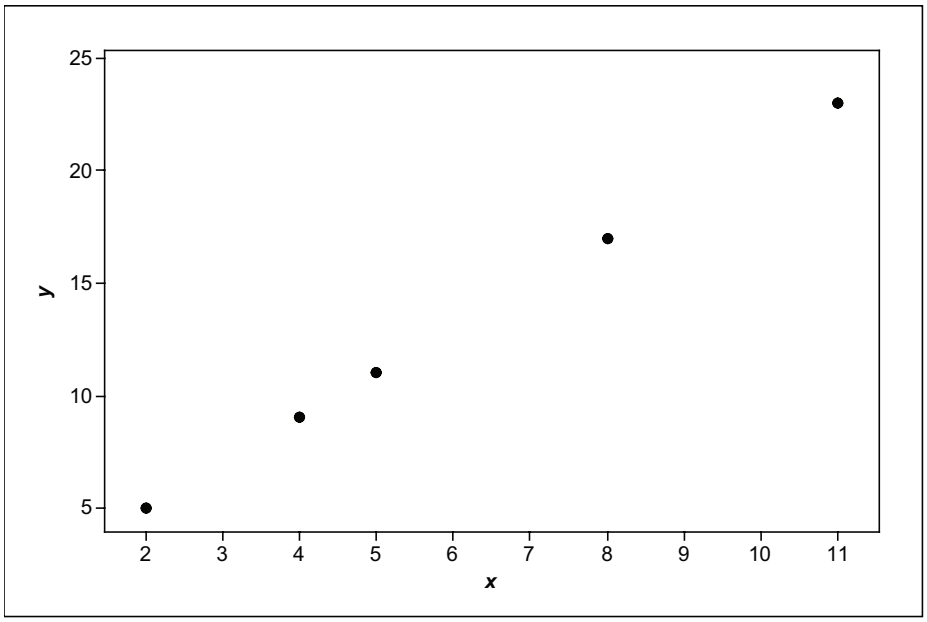
2. $-1 \leq r \leq 1$

3. The correlation between twins raised apart should be close to the correlation between twins raised together.

4. No – scaling can make a fairly strong correlation appear to be fairly weak and a weak correlation appear relatively strong. If two scatterplots are drawn on the same scale, then the data which produced the plot that is more scattered will have the lower correlation.

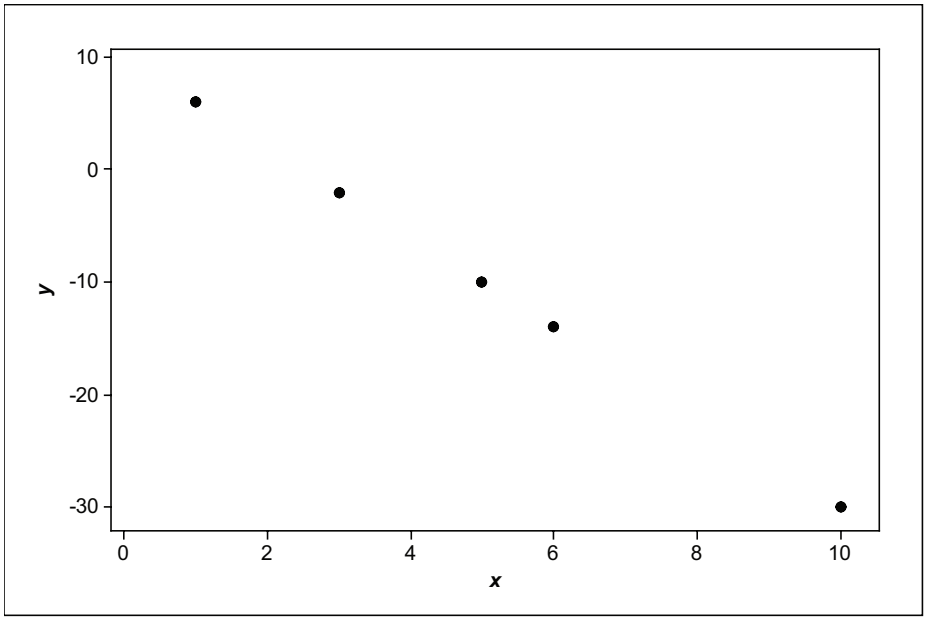
UNIT ACTIVITY SOLUTIONS

1. a.



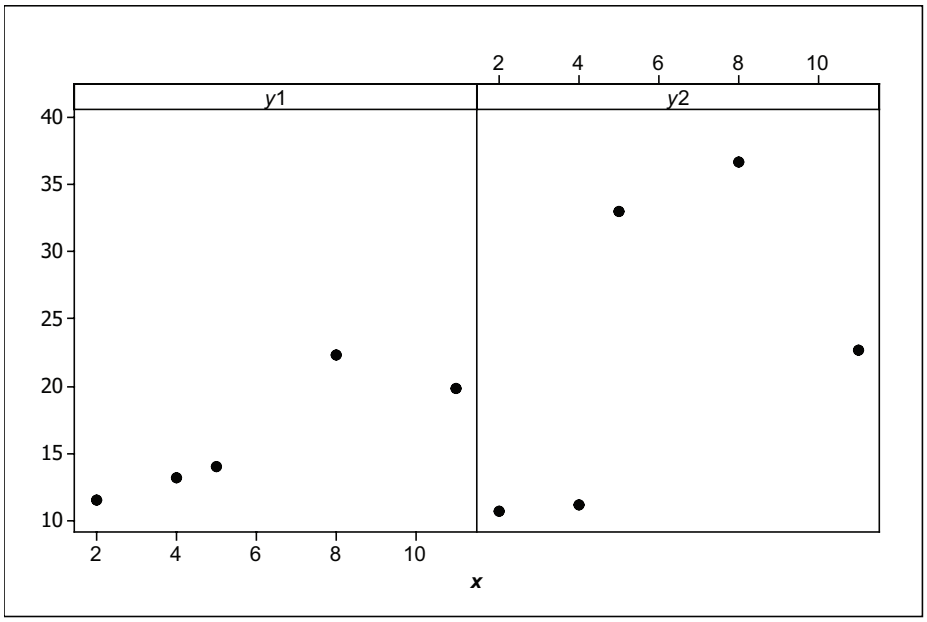
b. $r = 1$; the dots in the scatterplot fall exactly on a line with positive slope.

2. a.



b. The correlation coefficient: $r = -1$. The dots in the scatterplot fall exactly on a line with negative slope.

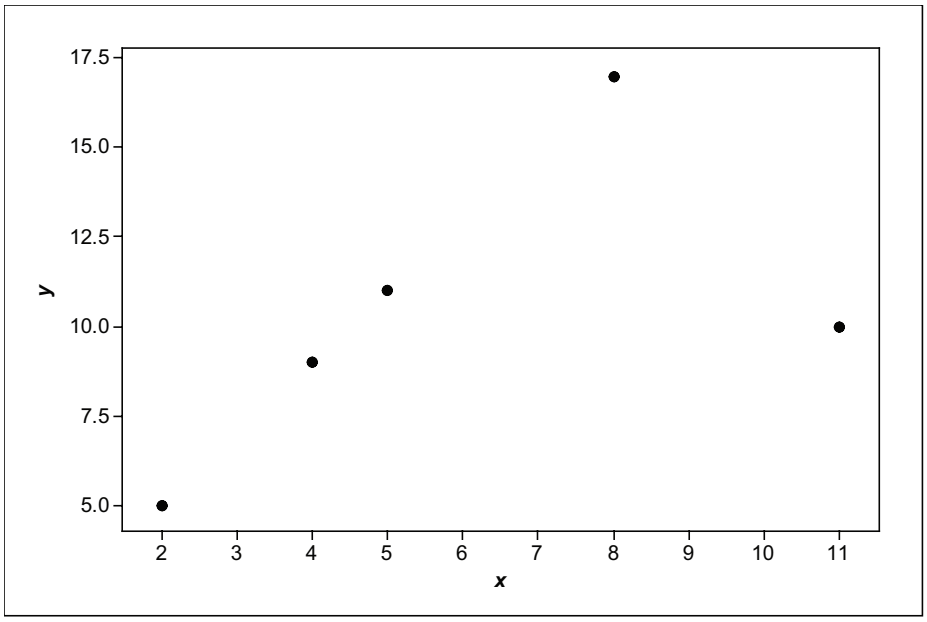
3. a.



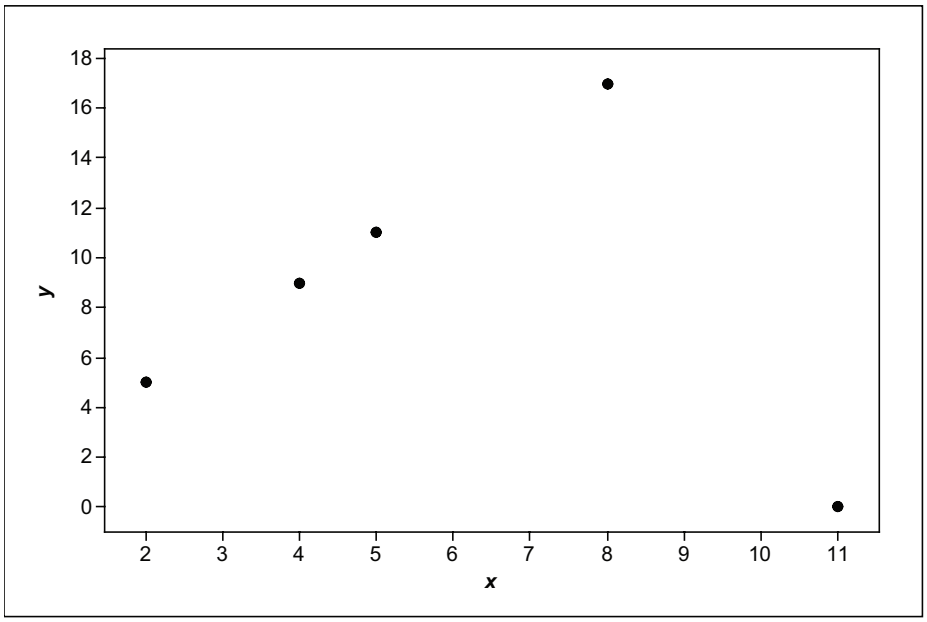
b. The plot of y_1 versus x (or the first plot) appears to show a stronger relationship.

c. For x and y_1: $r = 0.872$; strong relationship.
For x and y_2: $r = 0.522$; moderate relationship.

4. a. $r = 0.571$



b. r = - 0.166



5. Sample answer: When data points fall exactly on a line, r = ±1; if the slope of the line is positive, then r = +1 but if the slope of the line is negative, r = -1. The more scattered the data points are about a line, the closer r is to 0. A single outlier can have a huge effect on correlation. Even when the remaining data fall perfectly on a line with positive slope, a single outlier can dramatically reduce the value of r or even change it from positive to negative.

EXERCISE SOLUTIONS

1. a. Data from Marilyn A. Houck, et al. 1990. Allometric scaling in the earliest fossil bird, *Archaeopteryx lithographica. Science* 247: 195 – 198. The authors conclude from a variety of evidence that all specimens represent the same species.

There is no explanatory/response relation, so either variable can serve as *x*. In the sample solution, we have used femur length as *x* and humerus length as *y*.

The scatterplot shows a strong positive linear relationship between femur length and humerus length.

b. Femur length: $\bar{x} = 58.2$ cm and $s_x = 13.198$ cm

Humerus length: $\bar{y} = 66.0$ cm and $s_y = 15.890$ cm

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

$$r = \frac{1}{4}\left[\left(\frac{38-58.2}{13.198}\right)\left(\frac{41-66}{15.890}\right)+\left(\frac{56-58.2}{13.198}\right)\left(\frac{63-66}{15.890}\right)+\left(\frac{59-58.2}{13.198}\right)\left(\frac{70-66}{15.890}\right)+\left(\frac{64-58.2}{13.198}\right)\left(\frac{72-66}{15.890}\right)+\left(\frac{74-58.2}{13.198}\right)\left(\frac{84-66}{15.890}\right)\right]$$

$$r \approx 0.994$$

Note: Expect some round off error in student answers, especially if students chose to round the standard deviations to fewer than three decimals.

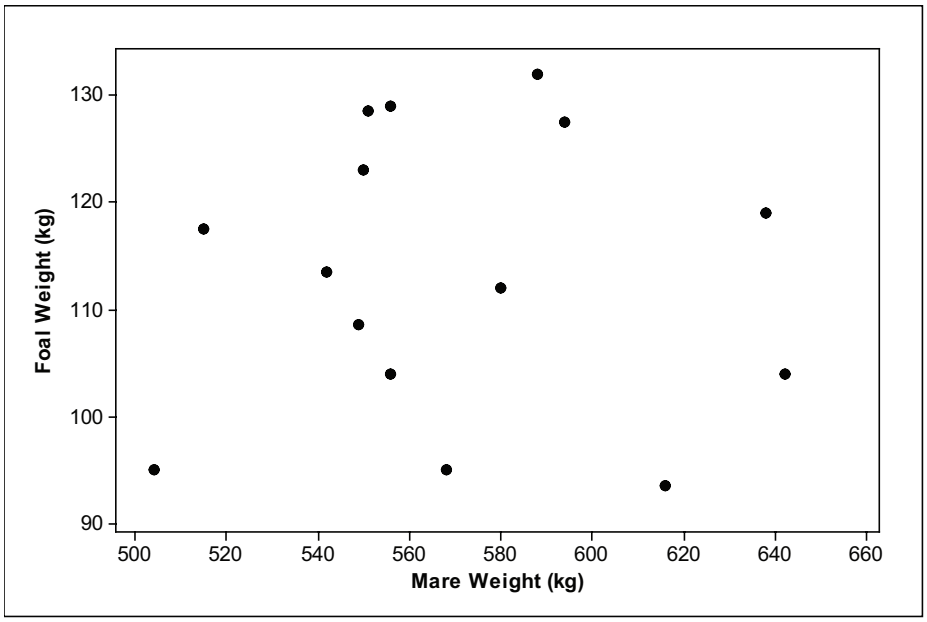
c. The high positive correlation supports the hypothesis that all five specimens appear to belong to the same species.

2. a. Gender is not a quantitative variable. You cannot calculate a correlation between gender (male or female) and anything. Correlation only makes sense for quantitative variables. (Note: “Correlation” is sometimes used to mean any kind of association, including an association between two categorical variables. However, it is better to restrict its usage to quantitative variables and *r*.)

b. Since $-1 \le r \le 1$, it is impossible for $r = 1.09 > 1$

c. r has no units – it's just a number between -1 and 1. So, r can't be measured in bushels.

3. a.



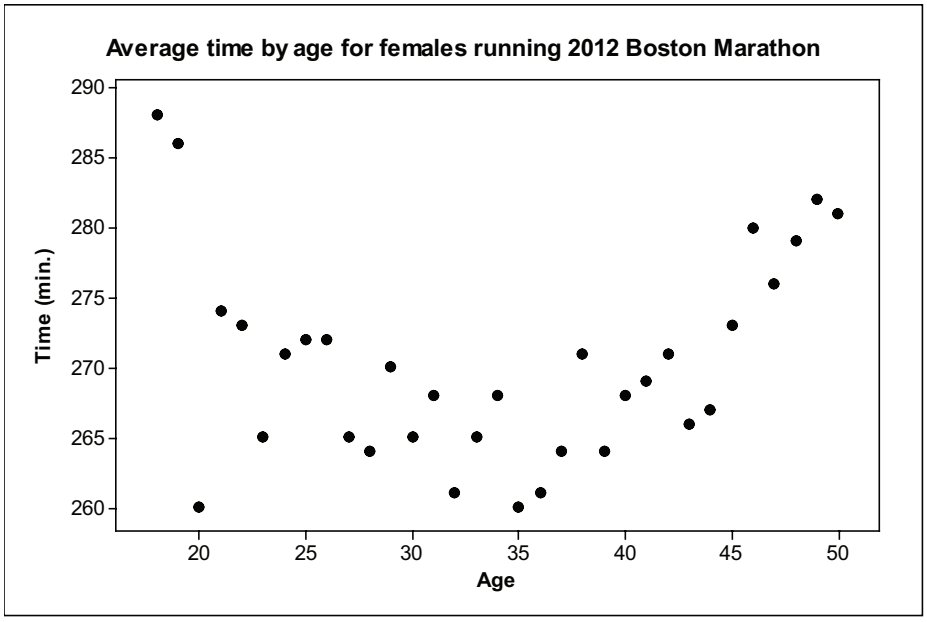
b. There doesn't appear to be any association between foal weight and mare weight. Some of the heaviest mares had the lightest foals while one of the lightest mares had a fairly heavy foal.

c. Expect the correlation to be closer to 0 than to 1 or -1.

d. $r = 0.001$. This confirms the answer to (c).

4. a. $r = 0.097$. If the relationship between average time and age were linear, it would mean there was a very weak, or practically no, relationship between these two variables.

b.

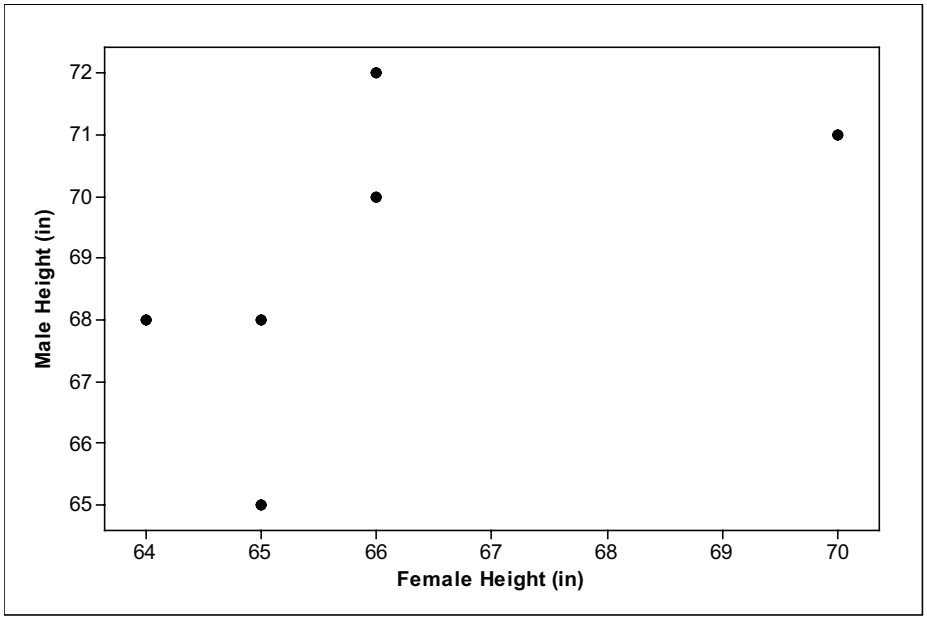


The relationship shows a curved pattern. Younger runners tend to have higher times but older runners also tend to have higher times. The best runners appear to be between 30 and 40.

c. Sample answer: The correlation coefficient measures the strength of a *linear* relationship. In part (a), the correlation was close to 0, indicating little relationship between the two variables. However, in (b) a curved relationship was noted – so there was a relationship between these two variables, but it didn't happen to be linear. Correlation should not be used to measure the strength of a relationship unless a scatterplot indicates that the form of that relationship is linear.

REVIEW QUESTIONS SOLUTIONS

1. a. In the scatterplot below, female height is on the horizontal axis since we expect that the height of the woman may influence whom she is willing to date. From the scatterplot the association between female and male heights appears to be positive. Hence, the correlation coefficient *r* should be positive. The data appear to be fairly spread out, so although the value of *r* should be positive, it should be less than 1 and not terribly close to 1.



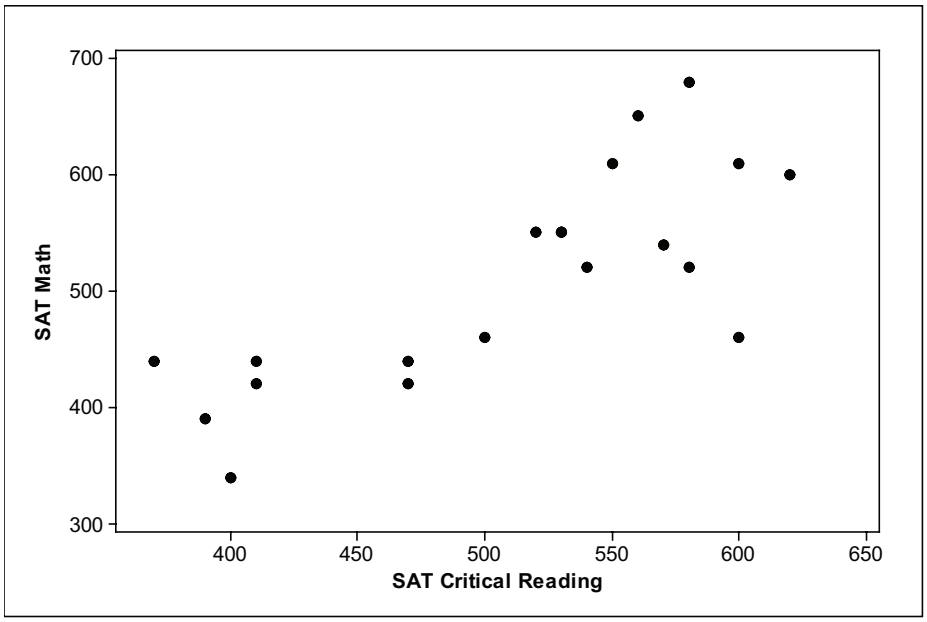
b. $r = 0.565$. The strength of the relationship between female heights and the heights of their male dates is in the low end of the moderate range.

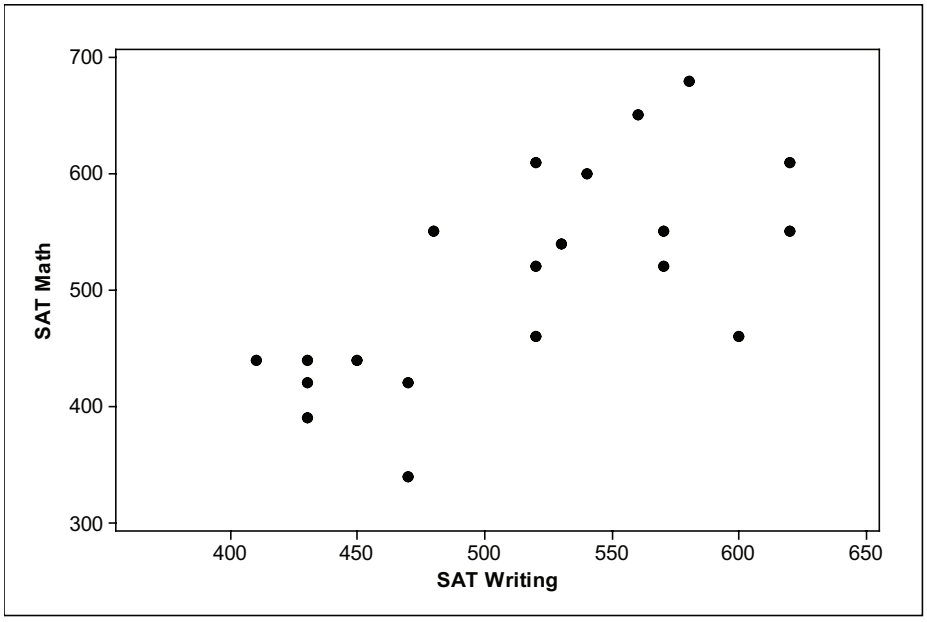
2. a. It would be $r = 1$. In this case the points fall exactly on the line $y = x + 3$, so there is a perfect positive linear association between the variables.

b. Changing all male heights by 6 inches does not change the correlation *r*. A positive correlation indicates that taller women tend to date taller men than shorter women date. It doesn't say anything about whether the women are dating men that are taller than they are.

c. Changing the height units from inches to centimeters did not affect the correlation; $r = 0.565$.

3. a.





The SAT Critical Reading exam appears to be more highly correlated with the SAT Math exam than the SAT Writing exam. The dots in the scatterplot of SAT Math versus SAT Critical Reading appear less spread out than the dots in the scatterplot of SAT Math versus SAT Writing.

b. The correlation between SAT Math and SAT Critical Reading is $r = 0.784$; the correlation between SAT Math and SAT Writing is $r = 0.680$. So the correlation between the SAT Math exam and the SAT Critical Reading exam is higher. The strength of that relationship should be classified as moderate (but it's at the upper end of moderate).

Unit 13: Two-Way Tables

PREREQUISITES

Students should know how to compute percentages.

ADDITIONAL TOPIC COVERAGE

Additional coverage on two-way tables can be found in *The Basic Practice of Statistics*, Chapter 6, Two-Way Tables.

ACTIVITY DESCRIPTION

Following the activity description, there is a questionnaire with three questions. Feel free to use this questionnaire or to use a modified version. If your class consists entirely of students of one class (freshman, sophomore, junior, senior), you might replace question 1 with the following question on gender:

What is your gender? Male Female.

If you decide to add questions, give students an opportunity to create the questions. Questions 2 and 3 are directly related to the video. In these questions students are asked to rate the physical beauty of their campus (or school) and their current level of happiness. Regardless of how you modify the questionnaire, be sure to include these two questions.

After the class data have been collected, they should be entered into an Excel or statistical software spreadsheet and distributed to the class. Sample data are available if you decide not to collect data from your class. The sample answers to the activity are based on these data. Once students have collected the data, they can complete the activity. This activity can be done either individually or in groups.

The categorical data collected from the Happiness Survey is ordinal data. In other words, there is an inherent order in the categories. When making tables or graphic displays using software, students will need to impose that order (unless the categories' alphabetical order is the same as the inherent order). So, for all tables and graphic displays that involve the variable Physical Beauty, check that the categories Bad, OK, and Good appear in that order (or the reverse order) and not in alphabetical order Bad, Good, and OK.

HAPPINESS SURVEY

Circle your answers to the following questions:

What is your class year?

Fr So Jr Sr

Rate the physical beauty of your campus (or school):

Bad OK Good

Rate your level of happiness today:

Unhappy So-so Happy

Happiness	Physical Beauty	Class
Happy	Good	Jr
Happy	Good	Jr
Unhappy	Good	Jr
Unhappy	Good	Sr
Happy	OK	Jr
Happy	Good	Sr
Happy	Good	Sr
Unhappy	Bad	Sr
Happy	Good	Jr
Happy	Good	Jr
So So	Good	Sr
Happy	OK	Jr
Happy	OK	Sr
Happy	OK	Sr
Happy	Good	Sr
So So	Good	Sr
Happy	OK	Sr
Happy	OK	Sr
Unhappy	Good	Sr
Happy	OK	Sr
Happy	OK	Jr
Happy	Bad	Sr
Happy	Good	Sr

Table T1. Sample Data collected from a college introductory statistics class.

THE VIDEO SOLUTIONS

1. Sample answers: race, gender, car color.

2. Somerville included a Happiness Survey.

3. The row variable was Happiness – the values of happiness were used to label the rows of the table. The column variable was Physical Beauty and its values were used to label the columns of the table.

4. The percentage of respondents rating Somerville's physical beauty as Bad went down as the level of happiness went up.

UNIT ACTIVITY SOLUTIONS

1. Sample answer (based on sample data):

		Physical Beauty			Total
		Bad	OK	Good	
Happiness	Happy	1	8	8	17
	So So	0	0	2	2
	Unhappy	1	0	3	4
Total		2	8	13	23

2. Sample answer: 17/23 × 100% = 73.9%

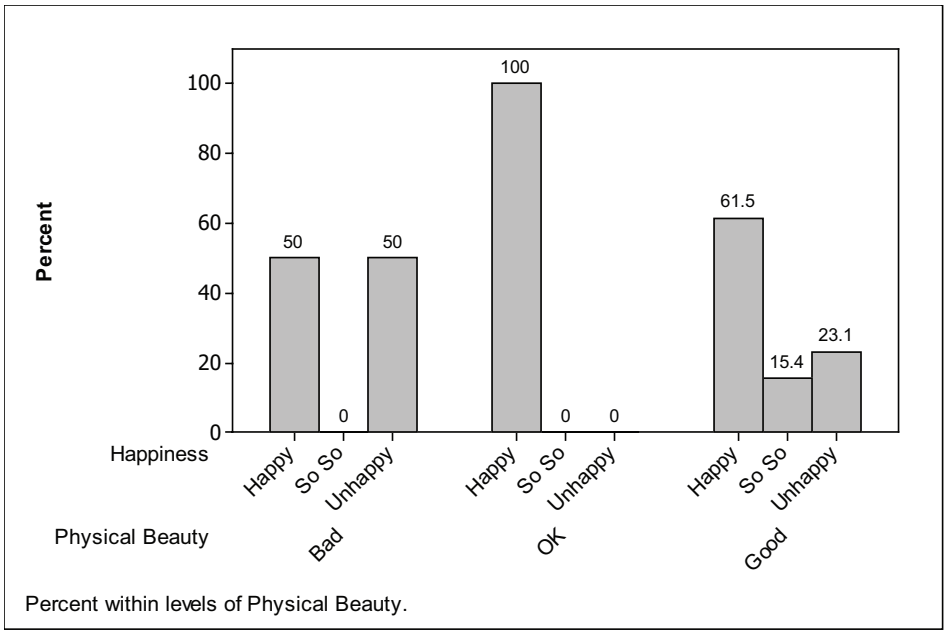
3. Sample answer: 13/23 × 100% = 56.5%

4. a.

% within Happiness		Physical Beauty			Total
		Bad	OK	Good	
Happiness	Happy	5.90%	47.10%	47.10%	100%
	So So	0.0%	0.0%	100.00%	100%
	Unhappy	25.00%	0.0%	75.00%	100%

b. Sample answer: Percent of happy students who rated campus physical beauty as Good: 8/17 × 100% = 47.1%. Percent of unhappy students who rated campus physical beauty as Good: 3/4 × 100% = 75.0%. A higher percentage of the unhappy students rated campus beauty as Good. However, there were only four students who rated themselves as unhappy.

5. a. Sample answer: Students' bar charts should show three distributions, one for each level of physical beauty. (*See chart on next page...*)



b. Sample answer: Fifty percent of the students who rated the physical beauty of the campus as bad were happy and fifty percent were unhappy. A higher percentage of students who rated campus physical beauty as Good were happy (61.5%) compared to those who rated campus physical beauty as bad. All eight of the students who rated campus physical beauty as OK were happy.

6. Sample answer. The class consisted only of juniors and seniors; 65.2% were seniors and 34.8% were juniors. The two-way tables below show the breakdown of Class with Happiness and Physical Beauty.

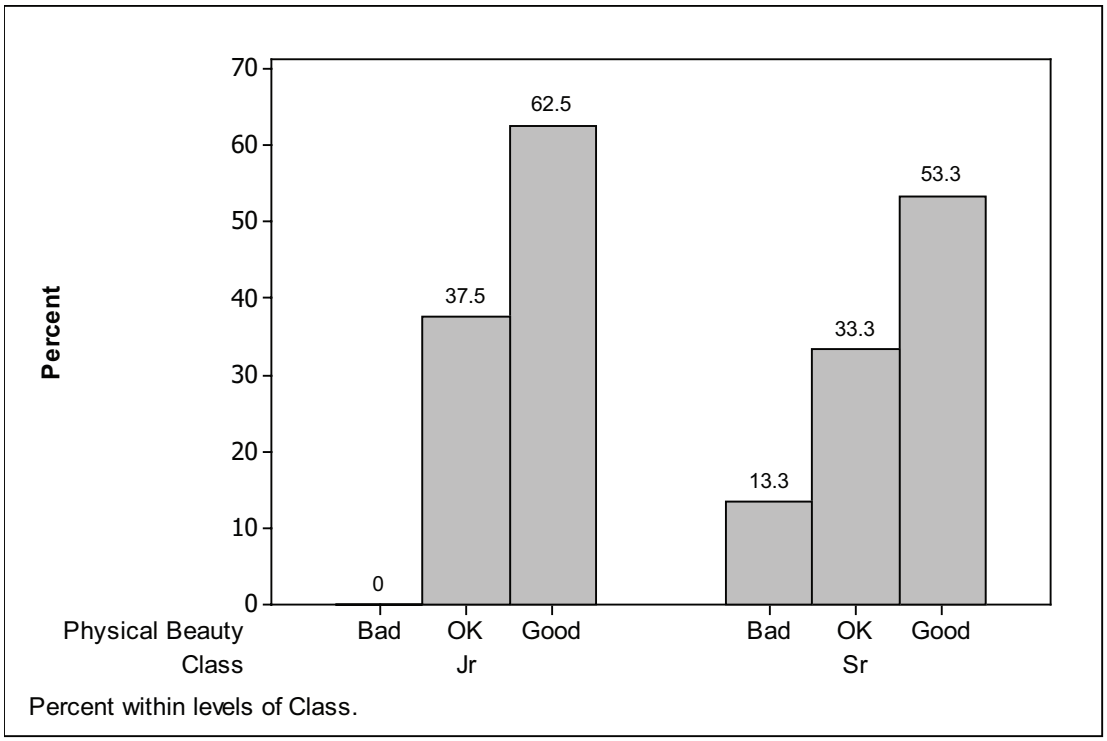
Count		Happiness			Total
		Happy	So So	Unhappy	
Class	Junior	7	0	1	8
	Senior	10	2	3	15
Total		17	2	4	23

Count		Physical Beauty			Total
		Bad	OK	Good	
Class	Junior	0	3	5	8
	Senior	2	5	8	15
Total		2	8	13	23

The juniors were happier than the seniors; 87.5% of the juniors rated themselves as happy compared to only 66.7% of the seniors. Below is a two-way table showing the distribution of Happiness conditioned on Class.

% within Class		Happiness			Total
		Happy	So So	Unhappy	
Class	Junior	87.50%	0.0%	12.50%	100.00%
	Senior	66.70%	13.30%	20.00%	100.00%

The graphic display below shows the distribution of Physical Beauty for each level of Class.



Notice that a higher percentage of juniors rated the campus physical beauty as Good compared to seniors; 62.5% of the juniors rated campus physical beauty as Good compared to 53.3% of the seniors. Notice also that none of the juniors rated campus physical beauty as Bad compared to 13.3% of the seniors. So, it appears that juniors have a “rosier” outlook than the seniors, both in terms of their happiness but also in how they see the beauty of their surroundings.

EXERCISE SOLUTIONS

1. a.

		Intelligence			Total
		Below Average	Average	Above Average	
Gender	Female	437	2243	4072	6752
	Male	456	1643	4593	6692
Total		893	3886	8665	13444

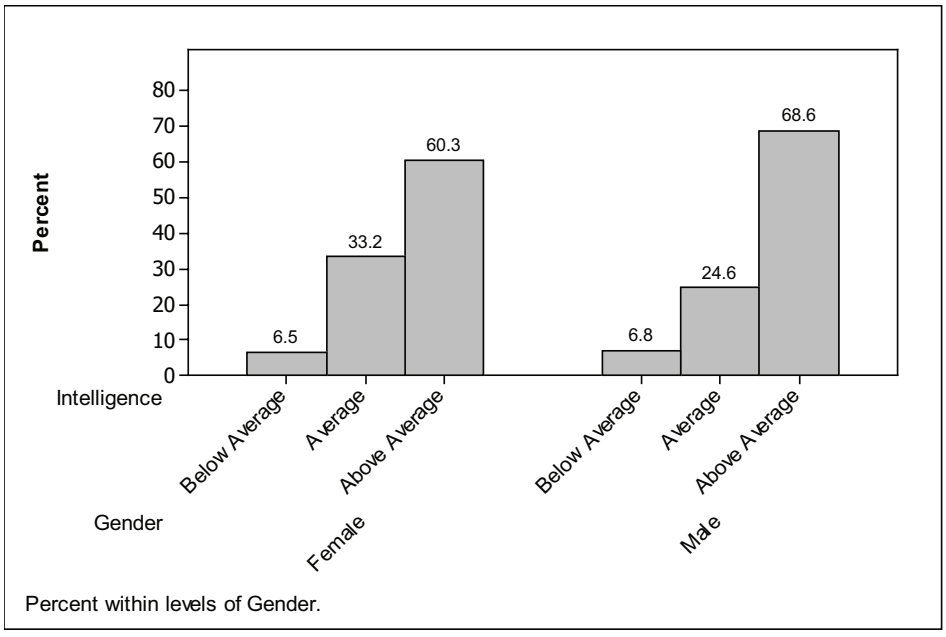
b. Male: 6692/13444 × 100% = 49.8%; Female: 6752/13444 × 100% = 50.2%

c. Above Average: 8665/13444 × 100% = 64.5%. A sizable majority felt that their intelligence was above average compared to others their age.

2. a.

		Intelligence			Total
		Below Average	Average	Above Average	
Gender	Female	437/6752 × 100% 6.47%	2243/6752 × 100% 0.3322	4072/6752 × 100% 60.31%	100%
	Male	456/6692 × 100% 6.81%	1643/6692 × 100% 24.60%	4593/6692 × 100% 68.60%	100%

b.



c. Sample answer: The percentages of male and female students who rated their intelligence as below average compared to others their age were about the same, 6.5% for females and 6.8% for males. A higher percentage of female students rated their intelligence as average compared to males, 33.2% for females and only 24.6% for males. A higher percentage of male students rated themselves as having above average intelligence than female students, 68.6% for males compared to only 60.3% for females. However, it should be noted a majority of both male and female 12th graders responded that they had above average intelligence compared to others their age.

3. a. Notice that the total in the table below is 100.2%. This is due to rounding the percentages.

		Political Preference					
		Rep	Ind	Dem	Oth	No Pref /Hvnt Decid	Total
Gender	Female	9.6	5.5	12.4	0.7	22.1	50.3
	Male	12.3	6.6	10.1	1.4	19.5	49.9
Total		21.9	12.1	22.5	2.1	41.6	100.2

b.

		Political Preference				
		Rep	Ind	Dem	Oth	No Pref/ Hvnt Decid
Gender	Female	19.2	10.9	24.6	1.3	44
	Male	24.6	13.2	20.2	2.8	39.2

c.

		Political Preference				
		Rep	Ind	Dem	Oth	No Pref/ Hvnt Decid
Gender	Female	44	45.4	55.1	32.7	53.1
	Male	56	54.6	44.9	67.3	46.1

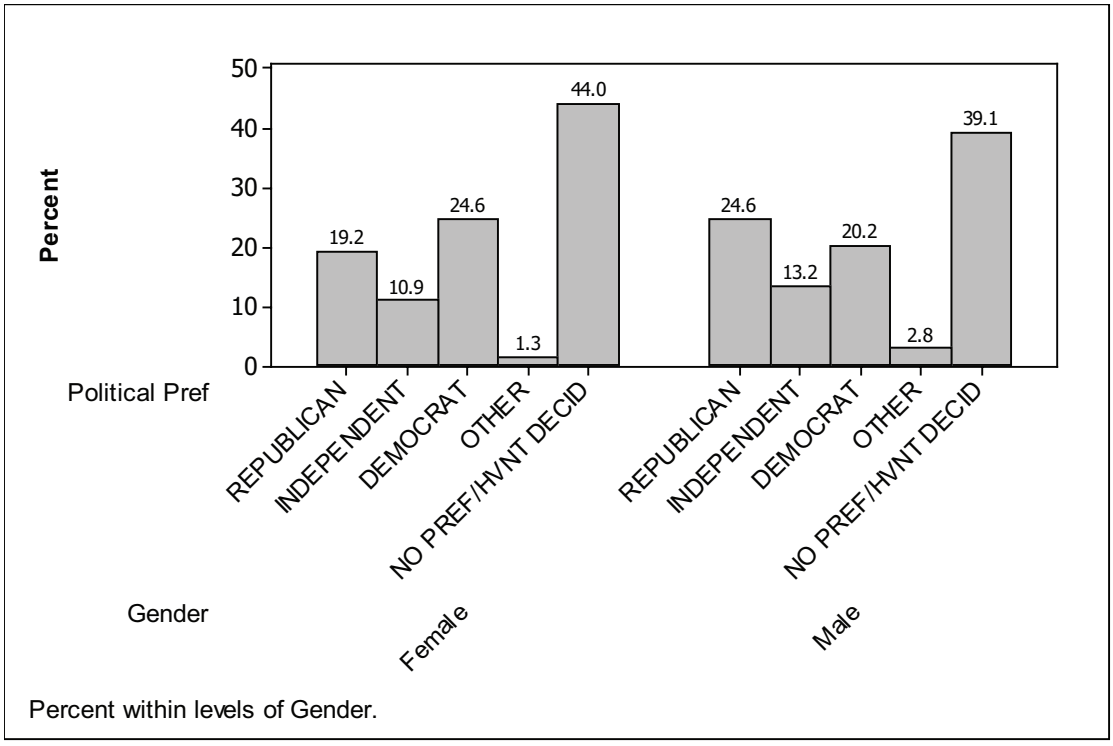
4. a. Percentage of respondents who were female and Democrats: 12.4%
Percentage of respondents who were males and Independents: 6.6%

b. Males were more likely to identify themselves as Republican than females; 24.6% of the males responded Republican compared to only 19.2% of the females.

c. Republicans were more likely to be male; 56% of the students who responded they were Republicans were male and only 44.0% were female.

In this question, the condition is Republican; there were 2895 students in this category. Then these 2895 students were broken down by gender. In (b) students are first broken down by gender; there were 6637 females and 6583 males. Then we see how many of each gender were Republican.

d. Sample answer: A higher percentage of male students (24.6%) responded Republican compared to females (19.2%). A higher percentage of female students (24.6%) responded Democrat compared to males. It is interesting to note that the same percentage of females responded Democrat as males who responded Republican. A considerably higher percentage of females (44.0%) reported having no preference or were undecided compared to males (39.1%). A very low percentage of either gender responded “Other” for political preference.



REVIEW QUESTIONS SOLUTIONS

1. a.

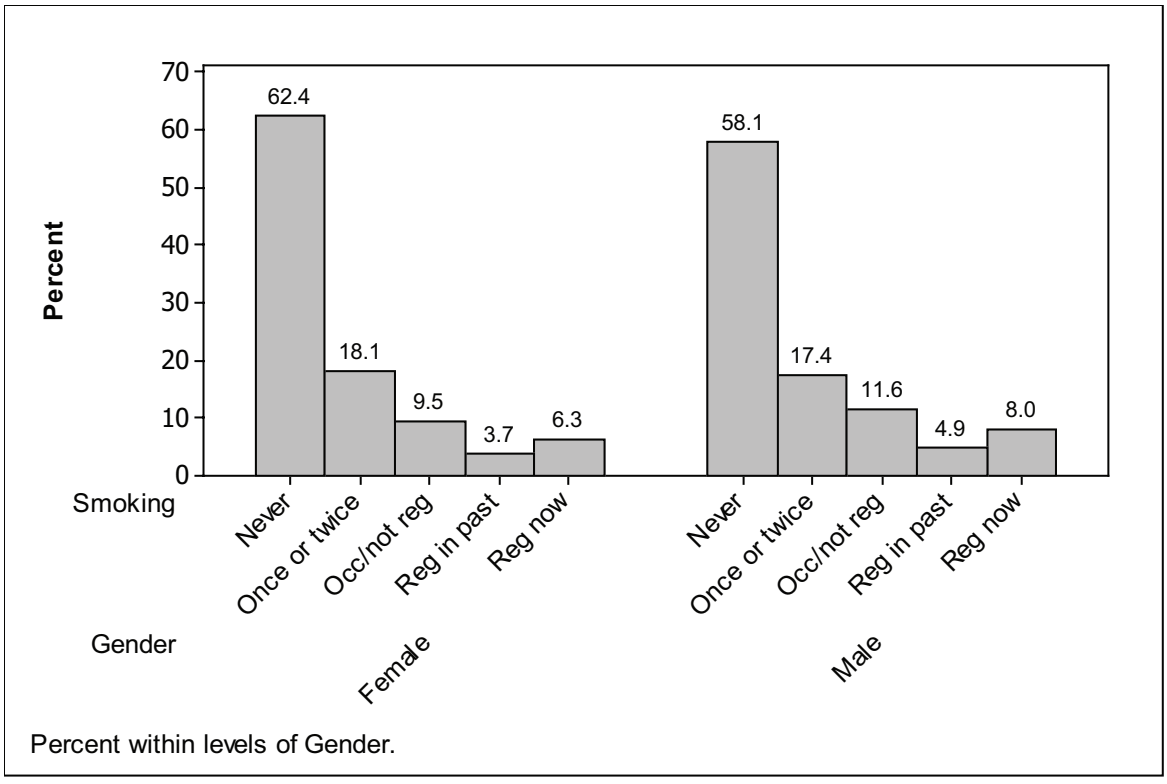
		Smoking					Total
		Never	Once or twice	Occ/ not reg	Reg in past	Reg now	
Gender	Female	4244	1228	649	253	428	6802
	Male	3957	1182	793	337	543	6812
Total		8201	2410	1442	590	971	13614

b. Females who never smoked: 4244/6802 × 100% = 62.4%

Males who never smoked: 3957/6812 × 100% = 58.1%

A higher percentage of the female respondents (62.4%) did not smoke compared to the male respondents (58.1%).

c.



2. a. 13,627

b. Never smoked: 8229/13627 × 100% = 60.4%
Smoked at least once: 100% - 60.4% = 39.6%

c. 3465/13627 × 100% = 25.4%

3. a. 3465/4751 × 100% = 72.9%
b. (926 + 46)/2221 × 100% = 43.8%

4. a. (489 + 168)/967 × 100% = 67.9%

b. (330 + 126)/591 × 100% = 77.2%

c. (3792 + 3465)/8229 × 100% = 88.2%

5. a. Grade D: sum = 99%; Grade C-, C, or C+ : sum = 100%;

Grade B-, B, or B+: sum = 100%; Grade A- or A: sum = 99%. The failure to sum to 100% is due to round-off error.

b. Sample answer: Here are the most striking patterns showing a relationship between grades and alcohol consumption. As grades increase, the percentage of students who had not consumed alcohol increases. As grades increase, the percentage of students who had consumed alcohol 6 or more times in the last 30 days decreases.

Unit 14: The Question of Causation

PREREQUISITES

Even though there are no specific mathematical or statistical prerequisites, the ideas presented are too sophisticated for most pre-secondary students. Unit 12, Correlation, is helpful for full appreciation of this unit, but it is not an essential prerequisite.

This video can also be shown in connection with science, because the issue of how to get evidence for cause-and-effect arises there.

ACTIVITY DESCRIPTION

In this activity, students are asked to use the Internet to find retrospective and prospective studies that have been conducted. To help narrow student searches, physical activity is specified as a suspected risk factor or protective factor for the retrospective study. For the prospective study, the outcome is specified as some disease or disorder. Question 3 is open and students are free to find any retrospective or prospective study. Give students an opportunity to share some of the studies they have found with the class.

THE VIDEO SOLUTIONS

1. People who have a higher standard of living are more likely to own multiple cars. Also, people who have more money tend to live longer, probably because they can afford excellent medical care and good quality food. The variable responsible for the car-lifespan association is wealth.

2. There could be a lurking variable, perhaps some gene, that made some people more apt to smoke but that also caused lung cancer. Hence, the lurking variable was the cause for the cancer, not the smoking.

3. A retrospective study takes a group of people – say a group with lung cancer – and looks back to see what characteristics members of the group have in common. With the lung cancer group, smoking stood out. A prospective study takes a group of people – say a group that contains both smokers and nonsmokers but otherwise have many similar characteristics – and then follows the group members over time. For example, years later the cancer rates in the smoking and nonsmoking group could be compared.

4. The smokers chose to smoke. Smoking was not imposed as a treatment on these group members.

5. The animal studies showed that smoking caused cancer in animals – hence, it was carcinogenic.

UNIT ACTIVITY SOLUTIONS

1. Sample study: Study involved 21 Type 2 diabetics; found increased physical activity associated with decreased healthcare costs. Data were gathered from participants' previous medical records. This study does not prove that increased physical activity causes a decrease in healthcare costs. More evidence is needed.

2. Sample study: A prospective study involved 3,875 young adults (age 20 – 32) who were not diabetic at the time of the study. The study was conducted to assess whether there was an association between mercury levels (determined from toenail clippings) and incidence of diabetes. These young adults were followed for 18 years. The study found that there was a significantly higher risk of diabetes associated with greater levels of mercury exposure. Because this is not an experiment, the study does not prove that mercury exposure causes diabetes.

3. Sample study: A prospective study to see if emotions of younger adults were more affected by physical activity and sleep than older adults. Participants reported in diaries their current day's mood and physical activity, as well as how many hours of sleep they got the night before. They recorded this information for four weeks. Findings of the study indicate that emotions of younger adults were more affected by physical activity and sleep than emotions of older adults. Prospective studies do not prove cause-and-effect relationships.

EXERCISE SOLUTIONS

1. The children range from 6 to 11 years in age. Because older children have had more time to develop their reading skills, generally they will be better readers than the younger children. In addition, older children tend to be bigger than younger children and should have bigger feet. In this situation, age is a lurking variable that explains the observed association between foot length and reading level.

2. Students who elect to take at least two years of a foreign language are, on average, both better students and more interested in language than those who take no foreign language. They would do better in English even if they did not study a foreign language. These characteristics of the students are lurking variables.

3. a. Sample answer: A retrospective study starts with the outcomes and looks back. Select two groups, one a group of patients with heart disease and the other a group of people of similar age and background who do not have heart disease. Look for differences between the groups – in particular, do the heart disease patients tend to be more overweight?

In the discussion, ask about matching the people in the two groups, which is an important part of comparative studies. Also ask about problems with such a study. For example, perhaps people in the heart disease group lost weight after becoming ill, so they do not appear overweight at the time of the study but might have been overweight before the diagnosis. Perhaps the heart problems affected memory and the people in the heart disease group don't remember past habits accurately.

b. Sample answer: A prospective study starts with the suspected cause and looks forward to the outcome. Select two groups of people, one with people who are overweight and another group of similar people (in age, sex, type of occupation, etc.) who are not overweight. Follow all the subjects for many years and observe the heart disease rates in both groups.

In the discussion, note that directly comparing overweight and normal weight people is more convincing than a retrospective study. Also note that the prospective study is expensive and takes many years.

4. a. Study 2 is the retrospective study. We start with the outcomes – workers who filed lost-time claims due to back injuries. The data from those workers were then retrieved from past

records that were available. Study 1 is the prospective study. In that study, the group was all workers. Researchers had to wait until a worker was injured before following up to see if the worker had physical therapy or returned to work.

b. Study 1 would be more costly. Researchers have to work over a period of years to both collect and analyze the data. In addition, the outcome of missed work due to a low-back injury is probably somewhat rare. So, the group to be followed would have to be pretty large.

c. Sample answer: Even though prospective studies are generally preferred, in this case, the outcome, low-back pain due to injury, is probably somewhat uncommon. So, the cost and time required to conduct a prospective study makes it not feasible in this situation. Sources of the data for the retrospective study have already been identified. So, that study would be more cost-effective and faster to conduct.

REVIEW QUESTIONS SOLUTIONS

1. Economic conditions may have improved over time, especially if the program was started during an economic downturn. Perhaps a major employer moved into the region during the four years. (Any change over the four years that helps employment is confounded with the training program.)

2. Positive association does not imply causation. Education, mental health, and home environment are other variables that might affect both marijuana usage and relationship problems for teens a number of years later.

3. This is not an experiment because nothing is actually done to the subjects. They are just observed over time. In particular, they choose whether to exercise. An experiment would assign subjects to different amounts of exercise. Instead, this is a prospective study. The two groups are chosen because they differ in the suspected cause (regular exercise) of reduced risk for heart disease. We follow them forward in time to observe the outcome of interest, whether they have heart attacks.

4. a. Clearly, Study 1 is a retrospective study. The men are dead – so, it would not be possible to do a prospective study.

b. Study 2 is a prospective study. The women joined the study and then the study followed them over time.

c. Breast cancer rates are relatively low (even though it is a common cancer), so the group size being studied would need to be large. If not sufficiently large, there would be a risk that none of the participants in the study develop breast cancer. In a large group, it is likely that there will be sufficient numbers of women who develop breast cancer. In Study 1, since it is a retrospective study, we already have a sizable group of white men who had heart disease. We don't have to wait to see if the disease develops in a retrospective study.

Unit 15: Designing Experiments

PREREQUISITES

Unit 14, The Question of Causation, demonstrates the difficulty of establishing a cause-and-effect relationship between an explanatory variable and a response variable when an experiment is not possible. Therefore, Unit 14 provides a good motivation for this unit. Units 15 through 17 form a sequence on statistical methods for producing data. This unit has no statistical or mathematical prerequisites beyond basic arithmetic.

The video also would be appropriate to show to science classes (especially biology classes because of the choice of examples) as well as to mathematics and statistics classes.

ACTIVITY DESCRIPTION

In the activity, students are asked to collect news items that describe medical experiments. Such items appear quite often in newspapers and magazines. Have the class discuss what information appears in the articles about the design of the experiment. Comparison (treatment and control groups) will often be mentioned, but the fact that the subjects were assigned at random will often not appear. Some news items may contain observational studies as well – more grist for discussion.

THE VIDEO SOLUTIONS

1. The study did not impose human populations on the various coral reefs. Instead, scientists simply observed the health of the coral reefs in four areas where human interaction with the areas was varied from no humans living in the area to a sizable population of humans currently living in the area.

2. The subjects were patients suffering with osteoarthritis of the knee. Researchers wanted to compare the effects on joint pain of the dietary supplements of Glucosamine or Chondroitin compared to a prescription medication or a placebo.

3. Randomization produces groups of subjects that should be similar in all respects before the treatments are applied. It allows us to equalize the effect from unknown or uncontrollable sources of variation.

4. Sample answer: His sample size was extremely small (the last two he called 7 and 8, so there were 8 subjects total). He treated the two subjects differently – one was allowed to sit and the other had to stand for over an hour. This difference in treatment would certainly affect subjects' moods. He didn't randomly assign the medications. He interacted with the patients sympathizing with their responses. He didn't record exactly what one of his patients said and instead recorded only the higher ranking of mood.

UNIT ACTIVITY SOLUTIONS

Sample answer: A portion of a news article appears below.

Study: Mediterranean diet may not protect aging brain
MNS News, January 25, 2013
Andrew Seaman of Reuters

http://news.msn.com/science-technology/study-mediterranean-diet-may-not-protect-aging-brain

Hopes that a Mediterranean diet would be as good for the head as it is for the heart may have been dampened by a French study that found little benefit for aging brains from the diet rich in fruit, vegetables, whole grains, nuts, wine and olive oil.

The study, published in the *American Journal of Clinical Nutrition*, looked at the participants' dietary patterns in middle age and measured their cognitive performance at around age 65, but found no connection between Mediterranean eating and mental performance.

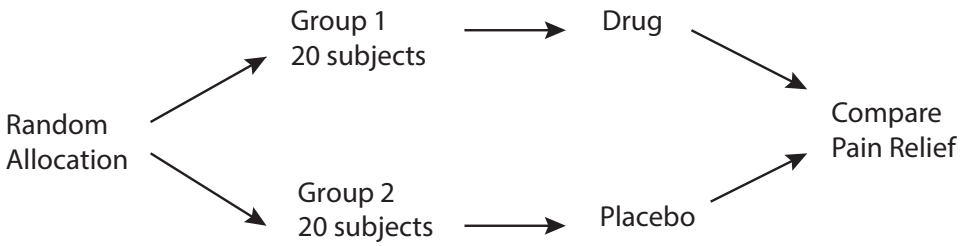
The study looked at patients' dietary patterns in middle age. Hence, this appears to be an observational study. There is no mention that a particular diet was imposed on the participants. The response variable is cognitive performance at age 65.

EXERCISE SOLUTIONS

1. No, this is not an experiment. The political scientist just gathers information from the subjects without imposing any treatment that could change their behavior. This is an observational study.

2. Yes, this is an experiment. The tasters are asked to react to specific treatments imposed on them by the experimenter, in this case to eat and compare the taste of two muffins.

3. a. The design is best outlined by a diagram. Students should give the size of the two groups (there are several reasons to use equal-sized groups in most cases), and to specify the treatments and the specific response they will look for.



b. Sample answers:

Using Table B from *The Basic Practice of Statistics*

Label the 40 subjects 01 to 40 in alphabetical order. Reading line 131 (this is a bit tedious) we get the 20 subjects in the drug group to be those with the following labels:

05	32	19	04	25	29	20	16	37	39
31	18	07	13	33	02	36	23	27	35

Subjects' names have been bolded in the table below.

01 Abrams	09 Daniels	17 Halsey	**25 Lippman**	**33 Rosen**
02 Adamson	10 Durr	**18 Howard**	26 Martinez	34 Solomon
03 Afifi	11 Edwards	**19 Hwang**	**27 McNeill**	**35 Thompson**
04 Brown	12 Fluharty	**20 Iselin**	28 Morse	**36 Travers**
05 Cansico	**13 Garcia**	21 Janle	**29 Ng**	**37 Turner**
06 Chen	14 Gerson	22 Kaplan	30 Obramowitz	38 Ullman
07 Cranston	15 Green	**23 Krushchev**	**31 Rivera**	**39 Williams**
08 Curzakis	**16 Gutierrez**	24 Lattimore	**32 Roberts**	40 Wong

Using Excel's random number generator Rand()

The names were entered into a column. In a second column 40 random numbers were generated. Then the names were sorted by the value of their associated random number. The bolded names in the table will be selected to receive the drug. The remaining 20 subjects will receive the placebo.

Janle	**Afifi**	**Travers**	**Kaplan**	**Roberts**	**Solomon**
Abrams	**Daniels**	**McNeil**	**Cansico**	**Rivera**	**Obramowitz**
Ng	**Durr**	**Garcia**	**Brown**	**Gutierrez**	**Lattimore**
Turner	**Curzakis**	Morse	Ullmann	Thompson	Krushchev
Hwang	Gerson	Adamson	Cranston	Edwards	Iselin
Williams	Lippman	Martinez	Howard	Fluharty	Chen
Rosen	Halsey	Green	Wong		

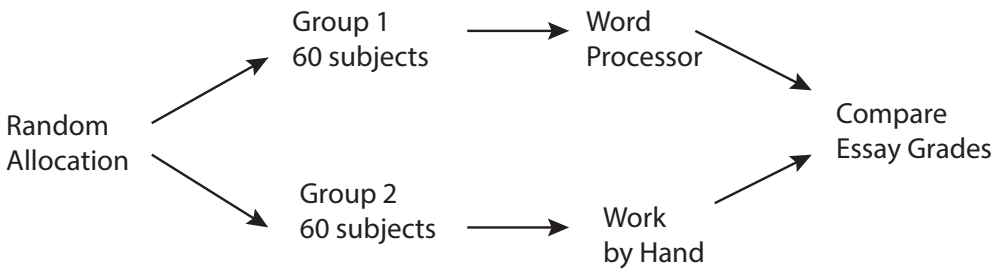
4. a. This is a double-blind experiment. Neither Dr. Colman, who is conducting the experiment, nor his patients know whether they are getting the remedy or the placebo.

b. In this case, both the person who conducted the experiment and the participants know which type of soda they are drinking. So this is not an example of either a single-blind or a double-blind experiment.

c. Since Janet labeled the cakes, she knows which is which. Her friends do not know which cake is made using cocoa and which using baking chocolate. Hence, this is a single-blind experiment.

REVIEW QUESTIONS SOLUTIONS

1. a. The response variable is not given explicitly, so students must specify what response they will measure. The grade on the essay is a reasonable choice, but there may be other good choices. Here is the outline of the design.



b. Sample answer: The essays should be read and graded by the same person (or two people – and scores averaged). More importantly, the experiment should be double-blind: the person reading the essays must not know which were word-processed. This means that all essays have to be retyped in the same form before being graded, so that only the quality of the essay influences the grade.

2. Sample answer: This is a very poorly designed experiment. First, it involved only two classes, which were at different schools and taught by different teachers (one more experienced than the other). The treatments, using the animated lessons compared to using handouts/ discussions, were not randomly assigned. Instead, only one of the two schools had sufficient numbers of computers to allow implementation of the animated science curriculum. Miss Earls' school was probably in a more affluent area than Mrs. Morrow's school – this conclusion is based on students' access to computers and the lower class size. So we don't even know if students in the two classes were similar in terms of their academic preparedness. Furthermore, Miss Earls designed the test, which may be biased toward the animated science lessons.

Miss Earls' school should not have purchased the animation science curriculum based solely on the outcome of this experiment.

3. a. This is an observational study – a prospective study. It takes a group of people, both smokers and nonsmokers, and observes them over a nine-year period. The response variable is whether or not the subject gets diabetes. The purpose of the study is to describe the

response variable (diabetic/not diabetic) for those who were smokers versus nonsmokers at the start of the study as well as those who were smokers and later quit smoking.

Background: In the article “Smoking, Smoking Cessation, and Risk for Type 2 Diabetes Mellitus” the design of the study was listed as a prospective cohort study.

b. You cannot conclude that quitting smoking causes diabetes. Most people who quit smoking also gain weight. Weight increases are also associated with diabetes. So, it would be impossible to tell whether the diabetes was caused by the cessation of smoking or the weight gain.

4. Sample answer: Randomly select the stores that consumer pairs will enter. Consumers should dress similarly – casual clothes suitable for visiting a mall (not too shabby but not too upscale either). The consumer pairs should be randomly assigned. Each pair should be randomly assigned to a store, complete their task and then be randomly assigned to the next store. There should be recorders with stopwatches to record the time it takes for a clerk to respond to a consumer pair.

Unit 16: Census and Sampling

PREREQUISITES

Using a random number table or a computer random number generator to select simple random samples was covered in Unit 15, Designing Experiments, and is needed in this section.

The description of the U.S. Census and the discussion of sampling ideas have much to say about where social and economic data come from. This unit therefore forms a bridge between mathematics/statistics and social science.

ACTIVITY DESCRIPTION

Give students an opportunity to explore the 2010 (or most current) Census home page. They can use Google or some other search engine to find the url, which changes periodically. (You can try http://www.census.gov.) This activity is designed to make students aware that they can access information from the U.S. Census.

In the activity, students are asked to find the current U.S. population at the time they logged in to the U.S. Census homepage. If students log in at different times, this number will be different. However, just being aware of the size of the current U.S. population will help students appreciate the enormity of the task involved in conducting a census every ten years. Next, students can focus on their own state – and use U.S. Census data to find its 2010 population and some basic demographic information. Question 3 is more open and asks students to compare demographic information on their state with a neighboring state.

THE VIDEO SOLUTIONS

1. The overall accuracy has improved over the years.

2. Representation to Congress for specific regions of the country is apportioned based on U.S. Census information as are federal funds. So an undercount in a certain area of the country means reduced representation and fewer federal dollars compared to what the area should receive.

3. A sample is chosen in such a way that each individual in the population has an equal chance of being selected for the sample.

4. A sample of 150 pounds of potatoes is taken by selecting 5 buckets of potatoes from various locations in the truck. From this sample, a smaller sample of 40 pounds of potatoes is selected for the cooking test (a hole is punched in these potatoes). Remaining potatoes are inspected for defects. There are many other samples along the way to test for: correct thickness, golden color, proper salt content, and satisfactory bag weight.

UNIT ACTIVITY SOLUTIONS

1. Sample answer: On 1/29/13 at 12:40 p.m. EST the population was 315,248,529. (By 1:00 pm the population was 315,248,599.)

2. a. Sample answer: For Massachusetts the 2010 population was 6,547,629.

b. Sample answer: Percentage of males: 3,166,628/6,547,629 × 100% ≈ 48.4%; Percentage of females: 3,381,001/6,547,629 × 100% ≈ 51.6%.

c. Percentage under 18: 1418923/6547629 × 100% ≈ 21.7%; Percentage of 65 & over: 902,724/6547629 × 100% ≈ 13.8%. A higher percentage of the population was under 18 than was 65 or over.

3. Sample answer comparing Massachusetts and Connecticut: Connecticut's population was only 3,574,097 compared to 6,547,629 for Massachusetts. The percentage of males in CT was slightly higher than in MA, 48.7% compared to 48.4%, respectively. The under 18 population in CT was 22.9%, slightly higher than in MA, which was 21.7%. The 65 & over population is also higher in CT at 14.2% compared to only 13.8% in MA. (Students could also compare housing, race, and more on age.)

EXERCISE SOLUTIONS

1. a. The population is not at all clear. Reasonable populations are all residents in the station's viewing area or all viewers of the 6 o'clock news. However, certain viewers who feel strongly about this question could call their friends and ask them to vote.

b. This is a voluntary response poll. The self-chosen respondents have stronger feelings on the issue than the population as a whole. In the case of gun control, the strong feelings of those opposed to gun control are well known. The poll results will almost certainly overstate the percentage of the general public who oppose the ordinance. In addition, different news stations attract different types of viewers. So this poll reflects the opinions of the viewers of this station and not necessarily the residents of the viewing area. Furthermore, many people do not get their news from television. Non-viewers of the 6 o'clock news are not included.

2. Selecting the sample using Table B from *The Basic Practice of Statistics*

01 Agarwal	**08 Dewald**	15 Hixson	22 Puri
02 Anderson	**09 Fernandez**	16 Klassen	23 Rodriguez
03 Baxter	10 Frank	17 Mihalko	**24 Rubin**
04 Bowman	11 Fuhrmann	18 Moser	25 Santiago
05 Bruvold	12 Goel	19 Naber	26 Shen
06 Casella	13 Gupta	**20 Petrucelli**	27 Shyr
07 Cordero	**14 Hicks**	21 Pliego	28 Sundheim

There are 28 students. Label them 01 to 28 in alphabetical order.

Line 136 of Table B is:

08421 44753 77377 28744 75592 08563 79140 92454

Reading two-digit groups and skipping those not used as labels, our sample contains the students labeled 08 14 20 09 24. These names have been bolded in the list above.

Selecting the sample using Excel's Rand ()

Step 1: Enter the names into column A of an Excel spreadsheet.

Step 2: In column B use Rand () to generate a column of 28 random numbers.

Step 3. Use Data>Sort to order the names in column A by their corresponding random number in column B.

Step 4. Select the first 5 names from the sorted list from Step 3.

Name	Rand
Agarwal	0.47616
Anderson	0.42692
Baxter	0.44405
Bowman	0.27579
Bruvold	0.12247
Casella	0.82995
Cordero	0.12318
Dewald	0.57609
Fernandez	0.31248
Frank	0.62985
Fuhrmann	0.93206
Goel	0.80434
Gupta	0.44848
Hicks	0.77278
Hixson	0.64893
Klassen	0.84144
Mihalko	0.19905
Naber	0.06491
Petrucelli	0.3356
Pliego	0.43135
Puri	0.42294
Rodriguez	0.963
Rubin	0.3613
Santiago	0.45452
Shen	0.13584
Shyr	0.54541
Sundheim	0.03402

Names ordered by Rand
Sundheim
Naber
Bruvold
Cordero
Shen
Mihalko
Bowman
Fernandez
Petrucelli
Rubin
Puri
Anderson
Pliego
Baxter
Gupta
Santiago
Agarwal
Shyr
Dewald
Frank
Hixson
Hicks
Goel
Casella
Klassen
Fuhrmann
Rodriguez

Sample answer: Sundheim, Naber, Bruvold, Cordero, Shen

3. a. The population would be all the Hudson Valley Patch Facebook readers or it could be all residents of the Hudson Valley region in New York state. (If the latter, the sample in (b) will miss all of the non-Facebook users in Hudson Valley.)

b. The sample would be the readers who voluntarily voted for the worst Valentine's Day gift.

c. No. First, not all Hudson Valley residents are on Facebook and connected to Hudson Valley Patch. In particular, the votes do not represent the opinions of non-Facebook users.

4. a. Population: all home sales in Worcester County, Massachusetts; sample: 50 home sales.

b. Population: all veterans who served in combat; sample: the 25 veterans examined by the psychologist.

c. Population: all seniors attending Eastern Connecticut State University; sample: 20 seniors questioned by the educator.

Some students may decide that the population is all students attending Eastern. However, then the educator has selected an unrepresentative sample because it only contains seniors.

REVIEW QUESTIONS SOLUTIONS

1. a. This question can be done using Table B from *The Basic Practice of Statistics*, but it is very tedious to do so. The sample answer relies on use of Minitab's Uniform random number generator.

Sample answer: We used Minitab's Uniform random number generator to assign a random number between 0 and 1 to each of the 48 students. Then we sorted the students by arranging their assigned random numbers from smallest to largest. The first 8 students in the sorted list were selected for the first group. (See sorted list in solutions to (b).)

b. Sample answer: The first 8 students in the ordered list are assigned to Section 1, the second set of 8 students are assigned to Section 2, and so forth. A complete listing of students and their sections appears below.

Names	Uniform	Section	Names	Uniform	Section
Juarez	0.0026	1	Elsevier	0.5289	4
Swokowski	0.0083	1	Stevenson	0.5366	4
Scott	0.0762	1	Fernandez	0.5616	4
Burns	0.1651	1	Barrett	0.5665	4
Schiller	0.1810	1	Poe	0.6165	4
Erskine	0.1882	1	Beerbohm	0.6184	4
Hyde	0.2005	1	Garcia	0.6562	4
Flury	0.2192	1	Rodriguez	0.6752	4
Taylor	0.2227	2	Orsini	0.6785	5
Arnold	0.2618	2	Putnam	0.6896	5
Jones	0.2908	2	Rowley	0.6905	5
Perlman	0.3187	2	Deneuve	0.7393	5
Nguyen	0.3546	2	Neale	0.7754	5
Kempthorne	0.3601	2	Moore	0.7772	5
Chang	0.3770	2	Campbell	0.7804	5
Quincy	0.4251	2	Dodington	0.7896	5
Prizzi	0.4559	3	Drummond	0.7964	6
Smith	0.4751	3	Oakley	0.7978	6
Ward	0.4786	3	Levine	0.8051	6
Hardy	0.4828	3	Ashford	0.8823	6
Colon	0.4888	3	Bartkowski	0.8961	6
Munroe	0.4966	3	Martinez	0.8965	6
Holmes	0.5011	3	Randall	0.9285	6
Vuong	0.5087	3	Rostenkowski	0.9408	6

2. a. The population is the set of students entering a college. The sample is the group of students questioned by this professor during their orientation.

b. The population consists of patients suffering from arthritic knees. The sample consists of 10 of the physical therapist's patients who had arthritic knees.

3. a. The population consists of all students who graduated from this university at least five years ago. (That way you can determine what they were doing 5 years after graduation.) You may want to narrow the population to students who graduated between 5 and 8 years ago or narrow even further to students who graduated exactly 5 years ago.

b. The cost to conduct a census would be too high and it would take too long to gather the results. If a survey is mailed to the graduates, you would have to track down those who didn't respond and try to get their information with one-on-one phone calls or home visits. So, this would greatly add to the time required to gather this information. Furthermore, no matter how hard you tried, it would be impossible to track down every graduate who graduated 5 or more years ago. Some will have left the area (or even the country) without providing forwarding addresses. Given a complete list of graduates in the population of interest, you could focus on a sample. Since the size is small, you could contact each person in the sample.

4. a. Sample answer:

Pros of conducting a census:

- If it is possible to contact everyone in the population, you get a true measure of the proportion of the population that supports the measure.
- Not only do you know the overall population proportion supporting the measure, you can also determine if there are specific subgroups of the population (even if the subgroup is a low percentage of the population) that oppose the measure.

Cons of conducting a census:

- It may not be possible to contact everyone in the population in one month. Some people may be away that month. Others may not want to be contacted at all and hence those people's views will not be represented in the census.
- If it is a large population, you may not have the manpower to contact everyone in one month.
- It will cost more to conduct a census than to take a sample.

b. Sample answer:

Pros of taking a sample:

- Costs would generally be lower than for a census.
- If good sampling techniques are used, the results collected from the sample should be representative of the views of the population.
- It may take less time to gather and analyze the data.

Cons of taking a sample:

- Data may not be representative of the population. This is particularly true if the sample size is small or if an inadequate sampling plan (such as voluntary sampling) is used.
- There is variability due to sampling. Different samples could lead to different results.
- Since you are working with a sample, you may not be able to get detailed information about certain subgroups within the population who oppose the measure, particularly if those subgroups are small in comparison to the population. (This problem may be fixed by revising the sampling plan.)

Unit 17: Samples and Surveys

PREREQUISITES

This unit continues the discussion of sampling begun in Unit 16, Census and Sampling, which is a prerequisite.

ADDITIONAL TOPIC COVERAGE

Additional coverage on sampling designs can be found in *The Basic Practice of Statistics*, Chapter 8, Producing Data: Sampling.

ACTIVITY DESCRIPTION

In this activity, students design a short survey questionnaire. Then they develop a sampling design for selecting a sample of students to complete the questionnaire. If possible, they go out and collect the data. Once they have collected the data, encourage students to analyze the responses using techniques from Unit 13, Two-Way Tables or Unit 29, Inference for Two-Way Tables. Some other units, such as Unit 28, Inference for Proportions, may also be useful.

Students should work in small groups on this activity. In question 1 they are asked to create questions for a survey questionnaire. Each group should come up with several questions. After groups have completed drafting their questions, the class should compose a 4 – 7 question survey using a subset of the questions written by the groups. To make the questionnaire interesting, encourage students to select questions on a variety of topics. (Or, to save time, you can select the questions for the class survey questionnaire yourself.)

In question 2, students are asked to describe in detail their sampling design for selecting a sample of at least 100 students from their school or campus. (If your school/campus is small, you may want to involve another school in this activity.) The goal is to select a representative sample. After groups have completed their sampling plans, allow them to share them with the class. Then select a design for the survey.

Among instructors who have tried this activity, there is a difference of opinion of the value of spending a lot of time on the careful collection of the survey data. Much depends upon the seriousness of the questions in the survey. If the conclusions have the potential of affecting life

at your school or campus, then there is some intrinsic value in spending the time apart from experiencing the difficulties in getting reliable information through opinion polls.

THE VIDEO SOLUTIONS

1. The sample was drawn from telephone lists and lists of people who owned cars. In 1936 only people who had considerable wealth had phones and bought cars. So, no data were collected from people who were not wealthy – an overwhelming majority of those people happened to be Roosevelt supporters.

2. A simple random sample might miss counties that are rural, suburban or urban, particularly if one of these types of counties is sparse among the various types of counties in a state.

3. Sometimes the entire population is divided into groups with similar characteristics. For example, the population could be divided into males and females. These non-overlapping groups are called strata.

4. The interviewer should not react any differently to those responses than he/she does to “normal” responses. In other words, there should be no reaction on the part of the interviewer.

UNIT ACTIVITY SOLUTIONS

1. Sample questionnaire from Pattonville High School:

Q1. Students at Pattonville High School are required to do 50 community service hours to graduate. Do you agree or disagree that this is a reasonable requirement?

Agree Disagree

Q2. Do you think that there is mutual respect between the students and teachers at our high school?

Yes No Not Sure

Q3. Some public schools begin the school year late in August. Other schools do not open until after Labor Day. Do you think that public schools should begin their year after Labor Day?

Yes No No Opinion

Q4. During this year have you participated in an organized sport?

Yes No

2. a. Sample answer: The population is all students at the school, college or university.

Students could decide that they only wanted to get information from seniors or that they were only interested in students majoring in STEM fields, or that they only wanted information from students in a certain dorm. So, the target population can vary.

b. The designs could be highly variable.

Sample answer from high school: We decided to use a multistage sampling plan. We started by taking a random sample of 12 homerooms and asked for a list of students in each of these homerooms. Then we took a random sample of 10 students in each homeroom. This gave us a sample of 120 students. We asked the homeroom teachers to administer the surveys to the students selected from their homeroom. We selected a total of 120 students in case some students were out of school on the day the surveys were administered.

Sample answer from college/university: We decided to use a multistage sampling plan. We randomly selected 6 dorms and then randomly selected two floors from each dorm. Then we got a list of the room numbers and randomly selected 10 rooms. We got the names of

the students in these rooms and plan to e-mail them the questionnaire. We will work with the RAs in each dorm to get the word out that we need this for our class and that students should respond. For students who fail to respond via e-mail, members of the class will go to their dorm rooms and try to track them down.

EXERCISE SOLUTIONS

1. Many African-American respondents would be unwilling to make negative comments about the police department to a police officer. To get trustworthy information, the poll should be taken by an independent agency that can preserve the confidentiality of the answers. Because the survey specifically seeks the opinions of African Americans, it would be a good idea to employ African-American interviewers.

2. a. The alternatives offered are slanted. We are asked to choose between a strong statement using the word “confiscation” versus a phrase from the Constitution. Forced to make this choice, many respondents will choose (ii) even though they might be sympathetic to a more reasonable gun control measure in light of concern over gun violence. One neutral question is “Some people have proposed greater restrictions on owning firearms. Do you agree or disagree that greater restrictions are needed?” However, a great variety of rewordings are possible and students may suggest even better alternatives.

b. The wording is impossibly complex. A translation into simpler English might be slanted because the question suggests reasons why recycling is desirable. Here’s one possible rewording: “Would you be willing to pay more for the products you buy if the extra cost were used to save resources by encouraging recycling?”

3. Sample answer: Because we want to ensure that both men and women are part of the sample (and we don’t know the percentage of women employees), it is best to use a stratified sampling plan. Particularly if there are fewer women than men, stratifying can ensure that we can draw conclusions about the women separately. Get a list of all employees separated by gender. Draw a random sample from the men and another random sample from the women.

Some students may explore multistage sampling. Perhaps the company has employees in many geographic locations, for example. We might sample locations, and then sample individuals at each location. Either or both stages could employ stratified samples.

4. a. Sample answer: The sample is biased because it will not include students who never eat breakfast or students who were out late Thursday night and didn’t show for Friday breakfast.

Alternative sampling plan: On Tuesday and Friday hand out questionnaires during breakfast, lunch, and dinner. Make sure to ask students whether they had already filled out the questionnaire so that they don't fill it out twice. This way the sample will include students from each mealtime and also students who show up on a Monday/Wednesday/Friday schedule or on a Tuesday/Thursday schedule.

b. Sample answer: The sample is biased because it does not include voters from any of the other 49 states. (Also the east and west coast tend to be more liberal than the interior states.)

One possibility is to use a multistage sampling plan in which a random sample of states are selected, then a random sample of counties within those states, and then a random sample of voters from each of the selected counties.

REVIEW QUESTIONS SOLUTIONS

1. The question is slanted in favor of a freeze by suggesting desirable outcomes.

2. The question has no single correct answer, but the problem almost demands a stratified sampling of colleges to ensure representation of each type of institution.

Sample answer: For the first stage, we need to pick some colleges and universities. We would start by looking at the number of colleges in the state – if there are only a few Class I universities, all of them large, we would include all of them in our sample of colleges and universities. Then we would sample a few from each of the other classes. We plan to select 10 institutions in the first sampling stage making sure, if possible, to have at least two from each class. For the second stage, we would take a simple random sample of 20 faculty members from each of the selected institutions.

Some students may decide to stratify the faculty by discipline (business, engineering, science, etc.), though that will quickly use up the 200 available places. However, they could also stratify faculty by rank or stratify disciplines by STEM and non-STEM fields.

3. a. Multistage sampling design.

b. Stratified random sample.

c. This could be a convenience sample. The resident assistants could select friends, or students who just walked through the door at a certain time – whatever was easiest.

d. Voluntary sampling.

4. Sample answer: The Ann Landers' poll did not ask the question in a neutral fashion. The letter that prompted the poll included the concern that the couple had friends who appeared to resent having had children. Including that statement may make it more likely for others who felt the same way to volunteer a response to the poll – thus, resulting in an unusually low estimate for the "Yes" response. The sample of those who responded to Lander's poll was clearly not representative of the population of parents in the U.S. The *Good Housekeeping* poll was also a voluntary poll. However, the sidebar introducing the question was quite neutral

and their response rate was much closer to the response rates of the *Kansas City Star* poll and *Newsday* poll. Even though the *Kansas City Star* poll randomly selected their participants, the sample was only from one state and shouldn't be used to make a statement about U.S. parents in general. Since the *Newsday* poll used random selection techniques to select a sample from the nation, the results from that poll are the most trustworthy.

Unit 18: Introduction to Probability

PREREQUISITES

Students need to know how to compute proportions (or relative frequencies) and percentages. They should be comfortable with converting proportions to percentages and percentages to proportions.

ADDITIONAL TOPIC COVERAGE

Additional coverage of probability can be found in *The Basic Practice of Statistics*, Chapter 10, Introducing Probability. Unit 19, Probability Models, continues the topic of probability.

ACTIVITY DESCRIPTION

In this activity, students will be flipping coins and tossing thumbtacks repeatedly a large number of times. This activity helps students better understand random phenomena.

MATERIALS

Coins (of the same denomination) for each student or each pair of students; thumbtacks (See Figure 18.4.).

Students can work individually or in pairs. If working in pairs, one student should flip the coin (or toss the tack) while the other records the outcomes. Then they can switch.

This activity provides plenty of opportunity to discuss random phenomena. For example, when flipping a coin, students might be quite surprised at the length of the longest run. Assuming the coin is a fair coin, many students may expect the outcomes to closely alternate between heads and tails. After students have gathered the data from 100 flips, determined the longest run, and calculated the proportion in 100 flips, have them share their results with the class.

Many students will think that 100 flips is “over the long run” but it is not. There will still be a fair amount of variability in the proportion of heads. So, at least for the coin flipping activity, members of the class should combine their results to get a better estimate of the probability of heads. Students may choose to do the same for tossing the thumbtack in Part II, even though the activity does not explicitly ask them to do so.

In Part II, students toss a tack and record whether it lands point down or point up.
When tossing a tack, students will get better results if they shake the tack in closed cupped hands before releasing the tack.

THE VIDEO SOLUTIONS

1. A random phenomenon is an event in which individual outcomes are uncertain but which has a regular pattern if repeated many times.

2. For any particular instance in the future, weather is not predictable with perfect accuracy. However, over time weather exhibits patterns.

3. It can mean that he is sure that 70% of the viewing area will get rain (coverage). It can also be used as the likelihood of any rain at all (level of confidence).

4. In the short run, the proportion of heads can be quite variable. In the long run, the proportion gets close to 0.5 and stays close to 0.5 as we continue flipping the coin.

5. The event whose probability is close to one is more likely to occur than the event whose probability is close to zero.

6. Essentially zero.

UNIT ACTIVITY:
OBSERVING RANDOM PHENOMENA SOLUTIONS

1. a. Sample answers: It means that getting heads is just as likely as getting tails.
If you flip the coin many, many times, approximately 50% of the flips should be heads.
The probability of getting heads is ½ .

b. Student answers will vary. Some students might expect to see patterns close to alternating heads and tails – so, their estimate might be quite small, 2 or 3. They will probably be quite surprised by the length of the longest run when they actually flip a coin repeatedly 100 or more times.

2. a. Sample data:

H T T H H T T T T H H H H T H T T H T H
T H T T H T H T T T H H T H H T T H H H
H H T H T T T H H H H H H T T H T H H T
T T T T H H T T T T H T H T H T T T H H
T H H T T H H T T H T T T T H H H T H T

b. Sample answer: Based on sample data in 2(a): There is a run of 6 heads.

c. Sample answer: 10 flips – 0.40; 20 flips – 0.50; 50 flips – 0.52; 100 flips – 0.48.

d. Sample answer: It is difficult to tell. There is quite a lot of variability in the proportions. While the proportion in 100 flips is less than 0.50, the proportion after 50 flips was the same amount above 0.50.

3. a. Sample answer: Together individuals in the class flipped the coin 2000 times. The total number of heads observed was 994. The proportion of heads was 0.497.

b. Sample answer: The proportion of 0.497 is close to 0.50. If we kept flipping the coin, the results would probably get closer to 0.50.

4. a. Sample data from 100 tosses:

D	U	D	U	U	U	U	D	D	U	U	U	D	D	U	U	D	D	D	D
U	U	D	U	D	D	D	D	U	D	D	D	D	U	D	U	D	U	D	U
D	D	U	D	D	D	D	U	U	U	U	U	D	U	D	U	D	U	D	D
U	D	D	U	U	D	D	U	D	U	U	D	U	D	D	U	D	U	D	D
D	D	U	U	D	D	U	U	U	D	D	U	U	D	D	U	U	D	D	U

Of the 100 tosses, the tack landed point up 46 times and down 54 times.

b. Probability of point up = 0.46 and probability of point down = 0.54.

c. 0.54 + 0.46 = 1.

EXERCISE SOLUTIONS

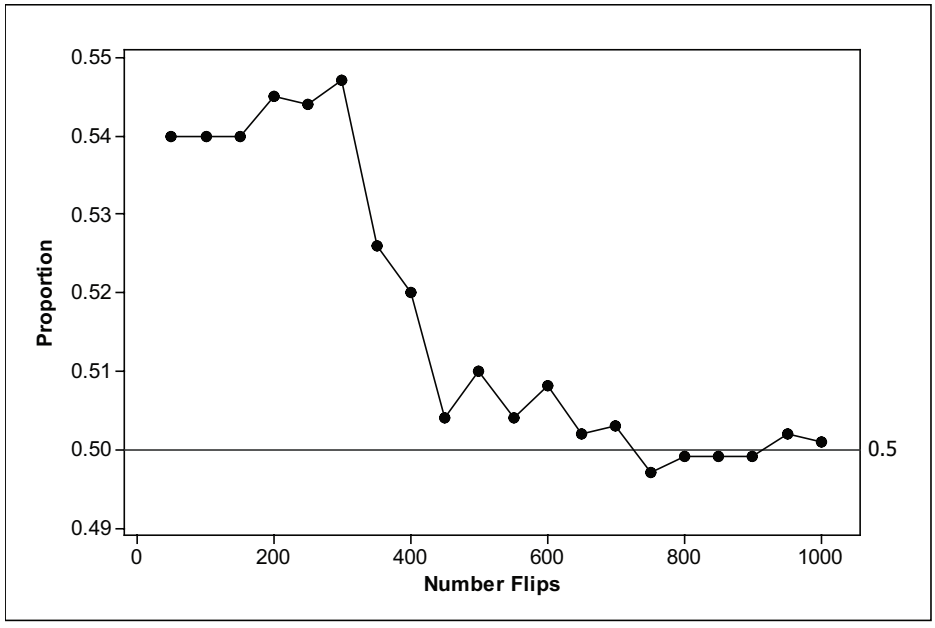
1. Sample answers: Time it takes to get to school each day; time it takes to complete a test or quiz; whether it rains or not; the temperature; length of the line when paying for lunch.

2. a. Data set (b) appears to be random. There does not appear to be a short-term predictable pattern. Data set (a) is all T's and hence totally predictable. In data set (c), the pattern alternates between H and T, and hence is totally predictable in the short term. For data set (d) the pattern is THHT which repeats and hence is totally predictable.

3. a.

Number Flips	Number Heads	Proportion	Number Flips	Number Heads	Proportion
50	27	0.540	550	277	0.504
100	54	0.540	600	305	0.508
150	81	0.540	650	326	0.502
200	109	0.545	700	352	0.503
250	136	0.544	750	373	0.497
300	164	0.547	800	399	0.499
350	184	0.526	850	424	0.499
400	208	0.520	900	449	0.499
450	227	0.504	950	477	0.502
500	255	0.510	1000	501	0.501

b.



c. Yes. Although the proportion started on the high side, after many flips, the proportions appear to be oscillating about 0.5.

4. a. Alex's chances are low – 3 out of 20 would be below 1/3.

b. Sample answer: When it is humid and cloudy and the temperature drops, very often rain results. Cooler air can't hold as much moisture as warm air. So, the chance for rain is high.

c. The chances are 4 out of 52, which is low.

d. Your chances are moderate. Half the deck are red cards. So, even with one card out of the deck, there would be either 25 or 26 red cards left. So, there are still about half as many red cards as there are cards left in the deck.

REVIEW QUESTIONS SOLUTIONS

1. a. 0

b. 0.5

c. 1

d. 0.0002

2. a. Sample answer: Just because a probability is high that some event will occur, it does not mean that the event is going to occur. In this case, there is a 30% chance that the event will not occur – which is still a sizable chance.

b. Amanda is not correct. The 70% refers to the percentage of occurrence over the long run. In the short term, random phenomena such as rain are unpredictable. You won't know for certain the outcome, until you observe what happens. A 70% chance of rain does not mean that it will rain exactly 7 times in 10 days.

3. a. The spinner is most likely to land on sector 5 because it is the largest sector on the spinner.

b. Sector 4 appears to be 1/4 of the area of the entire spinner. Hence, I would expect 1/4 of the 1000 spins or 250 spins to stop on sector 4. Since a probability of 1/4 refers to the proportion over the long run – over more than 1000 spins – it is likely that the actual outcome will be slightly more or less than 250.

c. Since sector 3 appears to be twice as large as sector 2, it is twice as likely that the spinner lands on 3 than on 2.

d. Approximately six sectors the size of the 3-sector would fit in the lower half of the spinner and six in the upper half. Five of the 3-sectors would cover the combined area of the 4-sector and 2-sector. Hence, the probability of landing on an even numbered sector is around 5/12.

4. a. 14,278

b.

Response	Frequency	Probability
No	1845	0.1292
Yes/Some	2637	0.1847
Yes/Most	2648	0.1855
Yes/Nrly All	7148	0.5006

c. The sum of the probabilities is 1.

d. There were 12,433 responses that were not ‘No’ out of a total of 14,278 responses. The estimated probability is the proportion 12,433/14,278 ≈ 0.8708.

Unit 19: Probability Models

PREREQUISITES

Students should be introduced to the concept of probability before working through this unit. Unit 18, Introduction to Probability, will provide that needed background.

ADDITIONAL TOPIC COVERAGE

Additional coverage of probability models can be found in *The Basic Practice of Statistics*, Chapter 10, Introducing Probability.

ACTIVITY DESCRIPTION

In this activity, students collect data from rolling dice to see how closely the probability models in Tables 19.1 and 19.2 capture the patterns of real data. Students can work individually or in pairs. Collecting the data – outcomes from 100 rolls of one die and sums from rolling two dice – could be done as a homework assignment to be completed before beginning the activity.

MATERIALS

Two dice for each student (or each pair).

The activity is in two parts. In Part I, students roll a single (fair) die 100 times and organize their results into a relative frequency table. Then they compare the relative frequencies from their data with the probability model in Table 19.2. They represent their results with a histogram and compare it to Figure 19.2, the probability histogram for rolling a single die. Students may find that their results do not match the probability model as closely as they expect. This is where you can remind them, for example, that the probability of rolling a 3 is the relative frequency (or proportion) of 3's over many, many rolls – 100 rolls is not enough.

To get more data, individuals or pairs combine their results to form the class data. The relative frequency tables and histograms for the class data should come closer to the probability model and probability histogram for rolling a single die.

Part II is a repeat of Part I, only this time students roll two dice and collect data on the sums.

THE VIDEO SOLUTIONS

1. A probability model is the set of all possible outcomes together with the probabilities associated with those outcomes.

2. $S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

3. $P(7) = 6/36 = 1/6$

4. $P(\text{not } A) = 1 - P(A)$

5. If events *A* and *B* are mutually exclusive, you can use the Addition Rule to calculate $P(A \text{ or } B)$, the probability that either *A* or *B* occurs.

6. If events *A* and *B* are independent, you can use the Multiplication Rule to calculate $P(A \text{ and } B)$, the probability that both *A* and *B* occur.

UNIT ACTIVITY:
PROBABILITY MODELS AND DATA SOLUTIONS

1.a. Sample answer: Here are the 100 outcomes when we rolled the die. (See solution to (b) for the frequencies.)

2 6 3 3 5 5 6 4 3 6 2 1 1 6 4 1 3 6 4 1

3 3 5 3 3 4 5 6 3 5 5 6 3 3 1 2 2 4 1 4

2 4 4 6 3 4 3 1 3 6 5 5 2 3 4 4 5 4 6 3

1 2 1 2 5 5 6 6 6 5 4 4 4 4 3 2 2 5 1 4

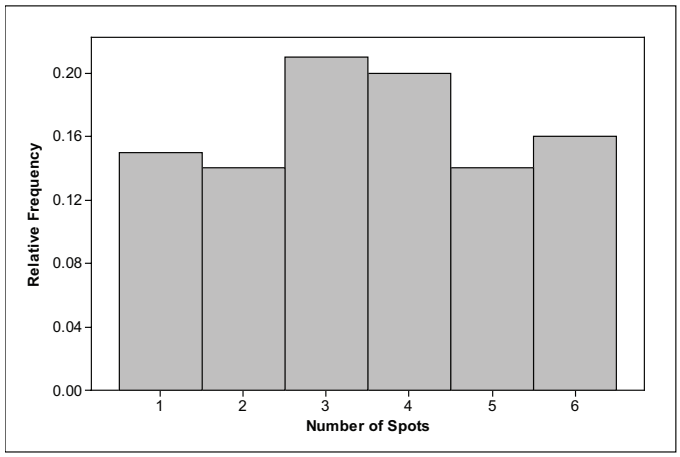
2 3 4 1 3 4 2 4 6 6 1 1 6 2 1 2 1 3 5 3

b. Sample answer (based on data from 1(a)):

Number of Spots	Frequency	Relative Frequency
1	15	0.15
2	14	0.14
3	21	0.21
4	20	0.20
5	14	0.14
6	16	0.16

The relative frequencies range in value from 0.14 to 0.21, so they are somewhat close to $1/6 \approx 0.167$ but they are certainly not equal.

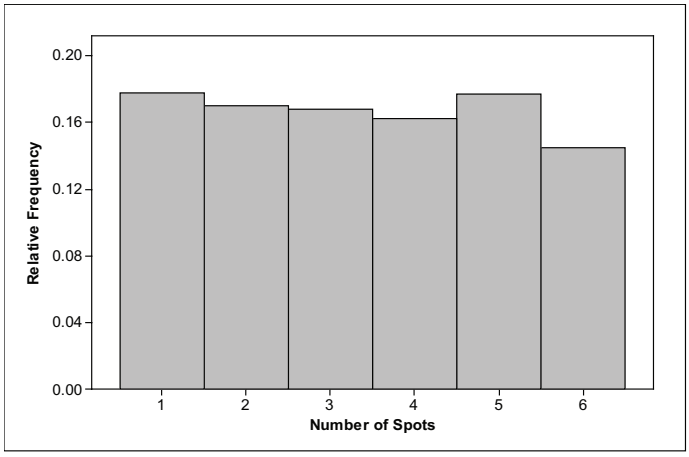
c. Sample answer: Unlike Figure 19.2, the bars are uneven in height.



2. Sample answer:

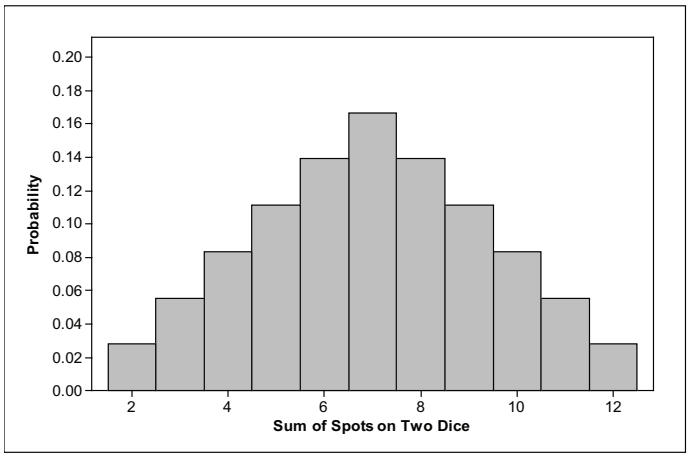
Number of Spots	Frequency	Relative Frequency
1	178	0.178
2	170	0.170
3	168	0.168
4	162	0.162
5	177	0.177
6	145	0.145

For the combined class results the relative frequencies are closer to $1/6^{th}$ than when the die was rolled 100 times. The members of the class rolled the die 1000 times.



The histogram based on the sample data of 1000 rolls has a closer resemblance to the probability histogram than the histogram based on sample data of 100 rolls.

3.

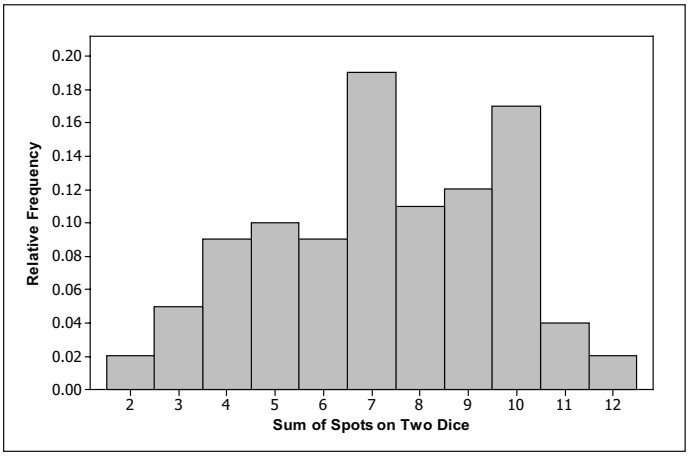


4. a. See solution to (b).

b. Sample answer: Most of the relative frequencies are close to the probabilities. The biggest difference is for the sum of 10; the relative frequency was 0.17 compared to a probability of around 0.083, a difference of 0.087 (more than double the actual probability).

Sum of Spots	Frequency	Relative Frequency	Probability
2	2	0.02	0.028
3	5	0.05	0.056
4	9	0.09	0.083
5	10	0.10	0.111
6	9	0.09	0.139
7	19	0.19	0.167
8	11	0.11	0.139
9	12	0.12	0.111
10	17	0.17	0.083
11	4	0.04	0.056
12	2	0.02	0.028

c. Sample answer (see next page): The left side of the histogram resembles the probability histogram more than the right side. (Expect a fair amount of variability in the shapes of students' histograms.)

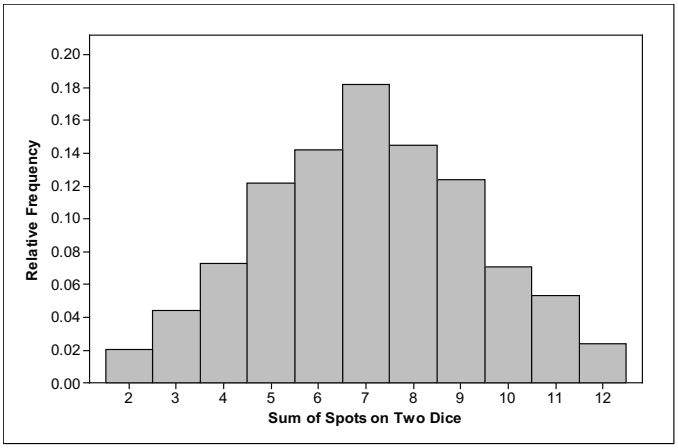


5. Sample answer: The relative frequencies are closer for the 1000 rolls than they were for the 100 rolls. The largest discrepancy was for the sum of seven where the relative frequency was off by 0.015.

Sum of Spots	Frequency	Relative Frequency	Probability
2	20	0.020	0.028
3	44	0.044	0.056
4	73	0.073	0.083
5	122	0.122	0.111
6	142	0.142	0.139
7	182	0.182	0.167
8	145	0.145	0.139
9	124	0.124	0.111
10	71	0.071	0.083
11	53	0.053	0.056
12	24	0.024	0.028

The histogram from 1000 rolls strongly resembles the probability histogram. So, the probability model appears to match what happens with the data over many rolls of the dice (certainly more than 100).

See next page...


Relative Frequency
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2
3
4
5
6
7
8
9
10
11
12
Sum of Spots on Two Dice

EXERCISE SOLUTIONS

1. a. 0.01

b. P(not Public Transportation) = 1 – P(Public Transportation) = 1 – 0.05 = 0.95

c. P(Drives) = P(Drives Alone) + P(Carpool) = 0.76 + 0.12 = 0.88

d. P(not Drive) = 1 – P(Drive) = 1 – 0.88 = 0.12

2. a. From 1(c) P(Drives) = 0.88. Since the workers were chosen randomly, the fact that one drives to work does not affect that probability that the other also drives to work. So, we can use the Multiplication Rule.

P(Both drive) = P(Worker 1 drives and Worker 2 drives)

= P(Worker 1 drives)P(Worker 2 drives)

= (0.88)(0.88) = 0.7744

b. From 1(d) P(not Drives) = 0.12.

P(Neither drives) = P(Worker 1 does not drive and Worker 2 does not drive)

= P(Worker 1 does not drive)P(Worker 2 does not drive)

= (0.12)(0.12) = 0.0144

P(at least one drives) = 1 – P(neither drives) = 1 – 0.0144 = 0.9856

3. a. It would mean that the probability of having that blood type is 1/2 or 50%; it means that a person is just a likely to have that blood type as to have a different blood type.

b. P(Rh+) = P(A+, B+, AB+, O+)

= P(A+) + P(B+) + P(AB+) + P(O+)

= 0.357 + 0.085 +0.034 + 0.374 = 0.850

The chance that a person has Rh-positive blood is higher than 50-50.

c. 85%

d. Sample answer: This is not a valid reason. First, P(O+ or A+) = 0.374 + 0.357 = 0.731. Assuming that the chances of needing a transfusion do not depend on blood type, you would expect about 73% of the people who need blood transfusions to be either type O+ or A+. In order to meet the higher need, the blood banks would want approximately 73% of blood donations to be from people who are type O+ or A+. In addition, it is particularly important for people with type O+ blood to donate because their blood type could be used in 85% of all blood transfusions – for all people who have Rh-positive blood.

4. a. First, we determine the probability of having type O blood:

P(O+ or O-) = P(O+) + P(O-) = 0.374 + 0.066 = 0.440

In a random sample of size 2, the fact that Person 1 is type O does not affect the probability that Person 2 is type O. So, we can use the Multiplication Rule:

P(Person 1 is O and Person 2 is O) = P(Person 1 is O)P(Person 2 is O) = (0.440)(0.440) ≈ 0.1936

b. First, we determine the probability of not having type O blood:

P(not O) = 1 – P(O) = 1 – 0.440 = 0.560

P(exactly one is O) =

P(Person 1 is O and Person 2 is not O) + P(Person 1 is not O and Person 2 is O) =

P(Person 1 is O)P(Person 2 is not O) + P(Person 1 is not O)P(Person 2 is O) =

(0.440)(0.560) + (0.560)(0.440) = 0.4928

c. P(Person 1 not O and Person 2 not O) = P(Person 1 not O)P(Person 2 not O) = (0.560)(0.560) ≈ 0.3136

REVIEW QUESTIONS SOLUTIONS

1. a. $P(C) = P(\text{Straight or Right}) = P(\text{Straight}) + P(\text{Right}) = 0.6 + 0.25 = 0.85$

b. $P(\text{not Straight}) = 1 - P(\text{Straight}) = 1 - 0.6 = 0.4$ (Complement Rule).

Let D be the event that neither vehicle goes straight.

$P(D) = P(\text{Vehicle 1 not Straight and Vehicle 2 not Straight}) =$

$P(\text{Vehicle 1 not Straight})P(\text{Vehicle 2 not Straight}) =$

$(0.4)(0.4) = 0.16$. (Multiplication Rule)

c. If D is the event that neither vehicle goes straight, then not D is the event that at least one of the vehicles goes straight. $P(\text{not } D) = 1 - P(D) = 1 - 0.16 = 0.84$. (Complement Rule)

2. a. $S = \{GGG, GGN, GNG, GNN, NGG, NGN, NNG, NNN\}$

b. $A = \{GNN, NGN, NNG\}$, $B = \{GGN, GNG, NGG\}$, $C = \{GGG\}$, and

$D = \{GNN, NGN, NNG, GGN, GNG, NGG, GGG\}$

c. A and B are mutually exclusive; A and C are mutually exclusive; B and C are mutually exclusive.

d. $P(C) = P(G)P(G)P(G) = (0.51)(0.51)(0.51) = 0.132651 \approx 0.133$

$P(D) = 1 - P(NNN) = 1 - P(N)P(N)P(N) = 1 - (0.49)(0.49)(0.49) = 0.882351 \approx 0.882$

To find $P(A)$ we first find:

$P(GNN) = P(G)P(N)P(N) = (0.51)(0.49)(0.49) = 0.122451$

Note that $P(NGN) = P(NNG) = P(GNN)$

$P(A) = (3)(0.122451) = 0.367353 \approx 0.367$

To find $P(B)$ we first find:

$P(GGN) = (0.51)(0.51)(0.49) = 0.127449$

Since $P(GGN) = P(GNG) = P(NGG)$, $P(B) = (3)(0.127449) = 0.382347 \approx 0.382$

3. a. All the probabilities are between 0 and 1; the sum of the probabilities equals 1: 0.223 + 0.188 + 0.138 + 0.179 + 0.272 = 1

b. P(Less than \$100,000) = 1 – P(\$100,000 or over) = 1 – 0.272 = 0.728

c. P(\$75,000 or over) = P(\$75,000 to \$99,000 or \$100,000 or over)

= P(\$75,000 to \$99,000) + P(\$100,000 or over)

= 0.179 + 0.272 = 0.451

d. P(Below \$75,000) = 1 – P(\$75,000 or over) = 1 – 0.451 = 0.549; 54.9% of households will have total incomes below \$75,000.

4. a. Since the three households were chosen randomly, the fact that one of the households has a total income of under \$25,000 does not affect the chances that either of the other two households has a total income of under \$25,000. We need this for independence.

P(all three households have total incomes under \$25,000) =
P(House 1 under \$25,000)P(House 2 under \$25,000)P(House 3 under \$25,000) =
(0.223)(0.223)(0.223) = 0.0111

b. P(\$25,000 or over) = 1 – 0.223 = 0.777
P(all three households have total incomes of \$25,000 or more) =
P(House 1 \$25,000 or more)P(House 2 \$25,000 or more)P(House 3 \$25,000 or more) =
(0.777)(0.777)(0.777) = 0.469097 ≈ 0.469

c. P(at least one household \$25,000 or over) = 1 – P(no household \$25,000 or over) =
1 – P(all households under \$25,000) = 1 – 0.0111 = 0.9889

Unit 20: Random Variables

PREREQUISITES

Students should be familiar with the background on probability covered in Unit 18, Introduction to Probability, and Unit 19, Probability Models. They should also be familiar with the material on the normal distribution contained in Units 7 and 8, Normal Curves and Normal Calculations, respectively.

ADDITIONAL TOPIC COVERAGE

Additional coverage of probability can be found in *The Basic Practice of Statistics*, Chapter 10, Introducing Probability. Unit 21, Binomial Distributions, continues the discussion on distributions of discrete random variables.

ACTIVITY DESCRIPTION

This activity is based on data from the March Supplement Survey 2012, part of the Current Population Survey (CPS), which is sponsored jointly by the U.S. Census Bureau and by the U.S. Bureau of Labor Statistics. The series of questions that serve as the basis for this activity focuses on households. Keep in mind that a household can consist of more than one family. The nested sequence of survey questions is as follows:

Any children in household?
- If yes, any eat school lunch?
  - If yes, any get lunch free?
    - If yes, how many get lunch free?

The answer to each question can be viewed as a random phenomenon. Hence, in the activity, each question generates a random variable.

The first question is a yes or no question and hence, the random variable associated with it has two possible values. We assume that the random variable is a model for what is true in the

population of all U.S. households. Therefore, the mean and standard deviation of the random variable are considered to be population characteristics. In question 1, students use a column approach to calculate the mean and variance. This column approach is particularly useful in question 4 when students have to find the mean and variance of a random variable that has 8 possible values. In addition, this column approach to finding the mean and variance can be easily adapted to spreadsheets.

Question 2 provides the raw data from the 2012 March Supplement Survey and asks students to use those data to estimate the probabilities. In 2(b) students are asked to create an area model for the first two random variables. That model allows them to view the nesting of the questions. Check that students understand that to find a proportion of a proportion, you multiply proportions.

THE VIDEO SOLUTIONS

1. Random variable x can take on the values 0, 1, 2, 3, or 4.

2. The probabilities add up to 1.

3. The most likely number of heads is 2.

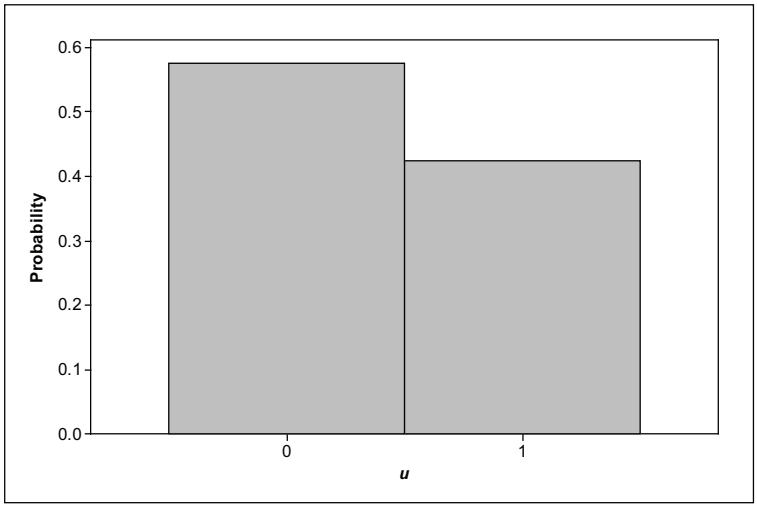
4. Failure of at least one of the O-rings.

5. Multiplication Rule.

6. P(at least one field joint failed) = 1 – p(0).

UNIT ACTIVITY SOLUTIONS

1. a.



b. It is less than 0.5. The balance point would be at 0.5 if both bars in the probability histogram were equal in height. Since the bar over $u = 0$ is taller than the bar over $u = 1$, the balance point must be less than where the two bars meet, at 0.5.

c.

u	$p(u)$	$(u)(p(u))$
0	0.586	0
1	0.432	0.432
	Sum =	0.432

d.

u	$p(u)$	$(u - 0.432)^2$	$((u-0.432)^2)(p(u))$
0	0.586	0.1866	0.1094
1	0.432	0.3226	0.1394
		Sum =	0.2487

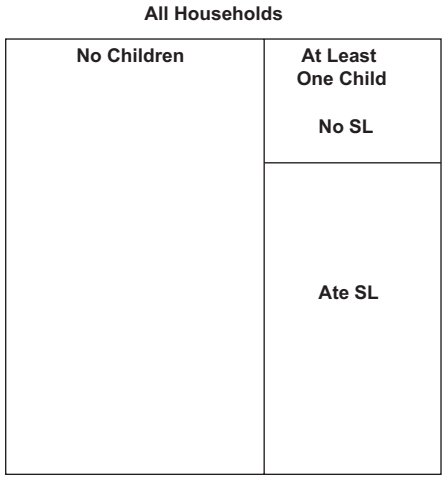
$\sigma_u^2 \approx 0.2487$; $\sigma_u \approx 0.4987$

2. a.

v	Frequency	Probability
0	32,491	0.303
1	74,690	0.697
Total	107,181	1

b. (0.432)(0.697) ≈ 0.301

Sample diagram:

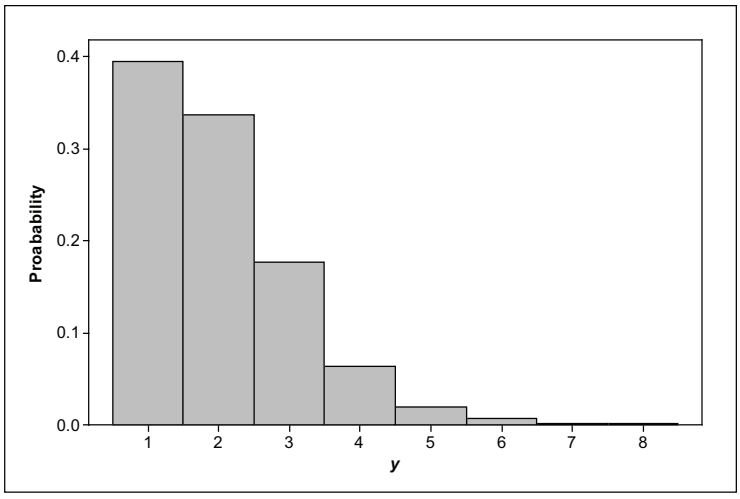


3. We would expect the proportion:

(number households participating in free lunch program) / 74,690

to be approximately 0.568. Hence, we get (0.568)(74,690) = 42,424 households where at least one child participates in the free lunch program.

4. a.



b. Sample answer: Around 2.

c. $\mu_y \approx 2.007$

y	p(y)	(y)(p(y))
1	0.394	0.394
2	0.337	0.674
3	0.176	0.528
4	0.063	0.252
5	0.019	0.095
6	0.007	0.042
7	0.002	0.014
8	0.001	0.008
	Sum =	2.007

d.

y	*p*(*y*)	(*y* - 2.007)^2	((*y* - 2.007)^2)(*p*(*y*))
1	0.394	1.0140	0.3995
2	0.337	0.0000	0.0000
3	0.176	0.9860	0.1735
4	0.063	3.9720	0.2502
5	0.019	8.9580	0.1702
6	0.007	15.9440	0.1116
7	0.002	24.9300	0.0499
8	0.001	35.9160	0.0359
		Sum =	1.1909

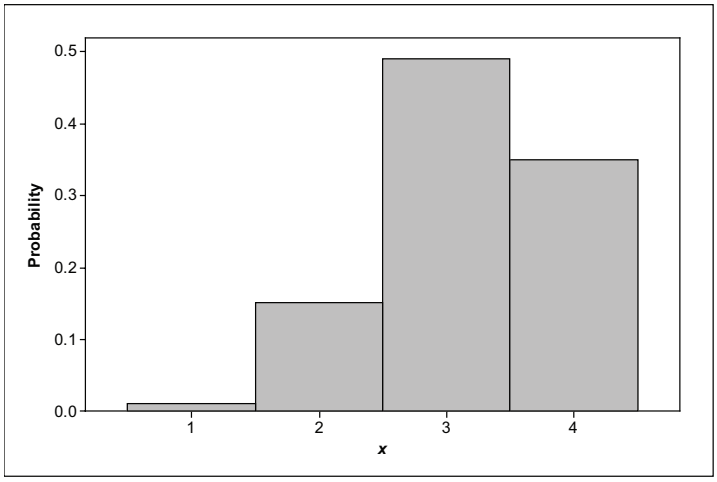
$\sigma_y^2 \approx 1.1909$; $\sigma_y \approx 1.091$

EXERCISE SOLUTIONS

1. a. $P(x \geq 3) = P(x = 3 \text{ or } x = 4) = p(3) + p(4) = 0.49 + 0.35 = 0.84$.

b. $P(x < 3) = 1 - P(x \geq 3) = 1 - 0.84 = 0.16$ The events “below B” and “B or above” are complementary events. Therefore the two events sum to 1.

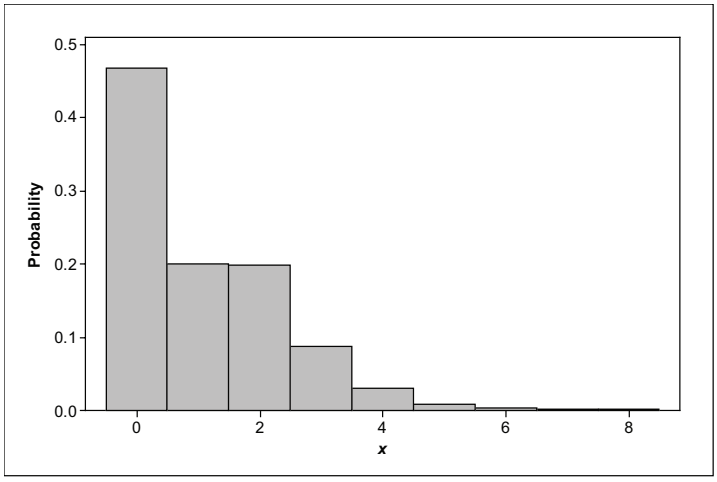
c. The histogram below shows that if a 12th grade student is randomly selected, the most likely outcome is x = 3, and the next most likely is x = 4. Not surprisingly, the least likely outcome is x = 1; students with D averages probably drop out of high school before completing the 12th grade.



2.a. $P(x \geq 1) = 1 - P(x < 1) = 1 - p(0) = 0.532$

b. $P(2 \leq x \leq 4) = P(x = 2 \text{ or } x = 3 \text{ or } x = 4) = p(2) + p(3) + p(4)$
= 0.199 + 0.087 + 0.031 = 0.317

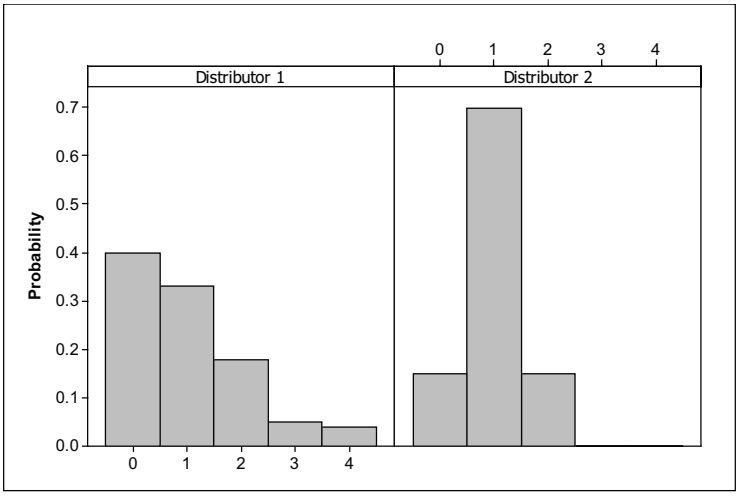
c. The histogram is skewed to the right.



d. μ = 1.068 children under 15 per household. The calculations follow:

0 × 0.468 + 1 × 0.200 + 2 × 0.199 + 3 × 0.087 + 4 × 0.031 + 5 × 0.009 + 6 × 0.003 + 7 × 0.002 + 8 × 0.001 = 1.068 child per household.

3. a. The histogram for Distributor 1 is skewed to the right and the histogram for Distributor 2 is symmetric.



b. Using the approach outlined in the unit activity, we get $\mu_x = 1$ and $\mu_y = 1$.

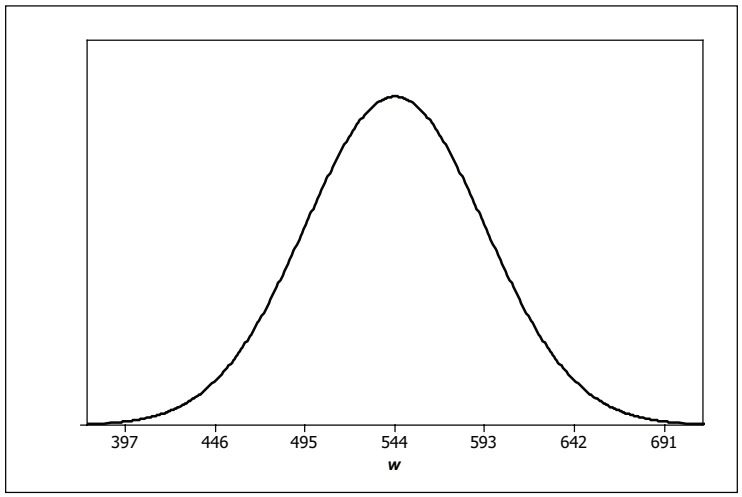
x or *y*	*p*(*x*)	*p*(*y*)	(*x*)(*p*(*x*))	(*y*)(*p*(*y*))
0	0.40	0.15	0	0
1	0.33	0.70	0.33	0.7
2	0.18	0.15	0.36	0.3
3	0.05	0	0.15	0
4	0.04	0	0.16	0
		Sum =	1	1

c. Using the approach outlined in the unit activity, we get $\sigma_x^2 = 1.14$ and $\sigma_y^2 = 0.3$. Therefore, $\sigma_x \approx 1.07$ and $\sigma_y \approx 0.55$. (Calculations follow.)

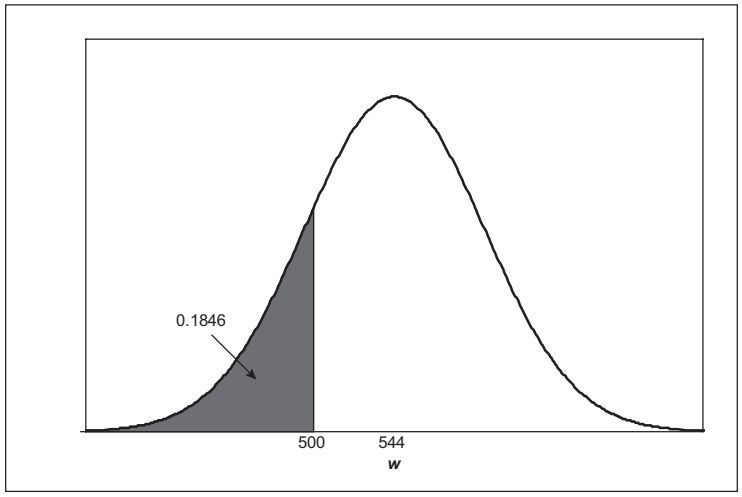
x or *y*	*p*(*x*)	*p*(*y*)	((*x*-1)^2)(*p*(x))	((*y*-1)^2)(*p*(*y*))
0	0.40	0.15	0.40	0.15
1	0.33	0.70	0	0
2	0.18	0.15	0.18	0.15
3	0.05	0	0.20	0
4	0.04	0	0.36	0
		Sum =	1.14	0.3

d. For both distributors, the mean number of defects in lots of four is 1. Hence, on average, over many, many lots, there will be one defect out of four, regardless of which distributor is used. However, the standard deviations show that for Distributor 2 the number of defects is more concentrated about the mean than it is for Distributor 1. Hence, when purchasing from Distributor 1, there will be more variability in a long sequence of lots than when purchasing from Distributor 2. Given the expected number of defects out of four is the same for both distributors, it makes sense to go with the distributor that has the more consistent lots.

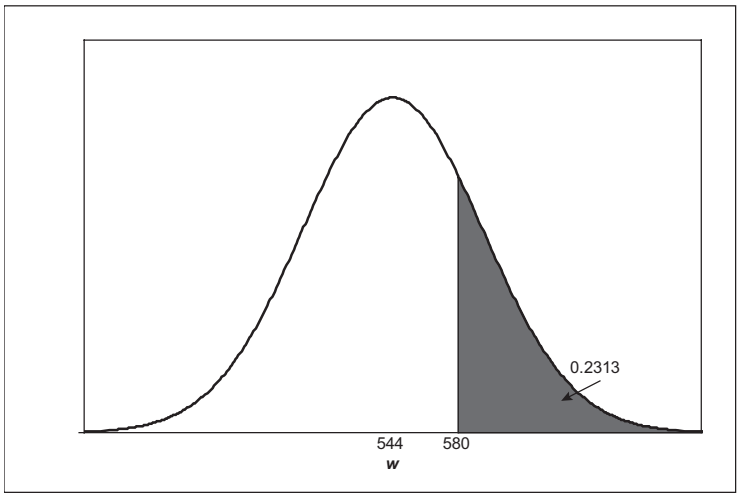
4. a



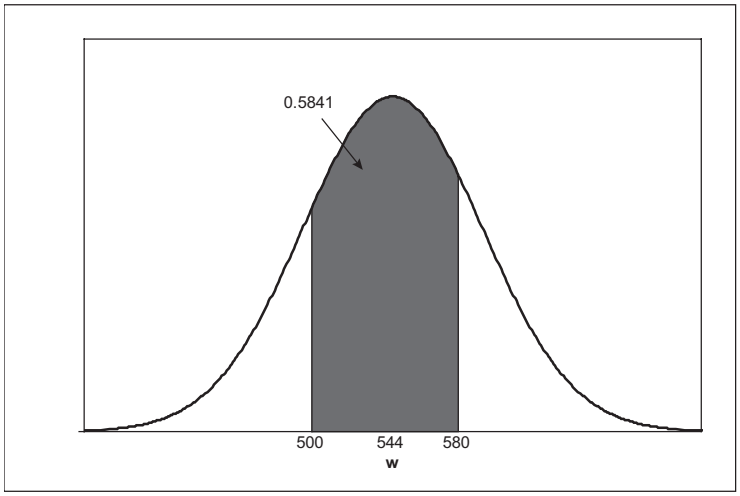
b. $P(w < 500) = 0.1846$



c. $P(w \geq 580) = 0.2313$



d. $P(500 \leq w \leq 580) = 1 - (0.2313 + 0.1846) = 0.5841$



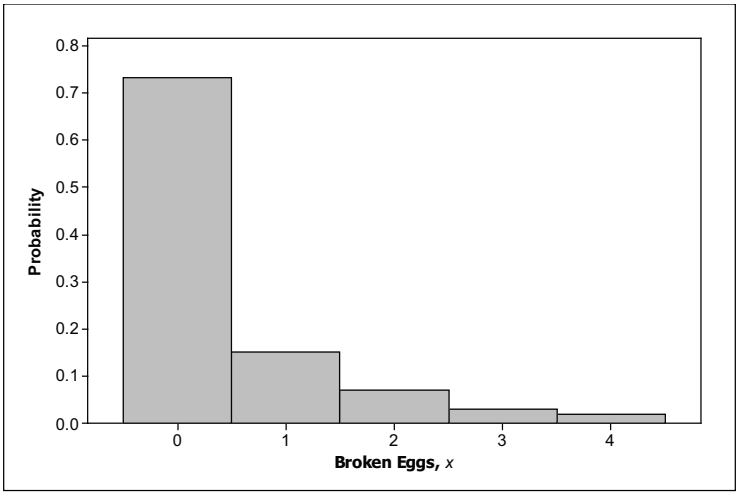
REVIEW QUESTIONS SOLUTIONS

1. a. $P(x < 4) = p(0) + p(1) + p(2) + p(3)$ = 0.73 + 0.15 + 0.07 + 0.03 = 0.98.

$p(4)$ = 1 – 0.98 = 0.02. This means that out of many, many egg cartons, roughly 2% will have four broken eggs.

b. $P(x \geq 2) = p(2) + p(3) = p(4)$ = 0.07 +0.03 + 0.02 = 0.12

c.



d. μ = (0)(0.73) + (1)(0.15) + (2)(0.07) + (3)(0.03) + (4)(0.02) = 0.46. This means that in the long run, after opening many, many cartons, the average number of broken eggs is 0.46. It is also the balance point for the probability histogram.

2. a. This is an example of a discrete random variable. There will be a finite number of Cheerios in a box. The weight limit of the box adds a cap to the number of Cheerios that will fit in a box.

b. This is an example of a continuous random variable. Time to finish takes values in an interval.

c. This is discrete. Possible amounts are separated by pennies. You could put the possible amounts in a list.

Comment: In some situations, money is treated as a continuous random variable. For example, this is true for determining the amount of money in a savings account that is accruing interest. Of course, the bank always rounds down to the nearest penny when paying out money.

d. Length is an example of a continuous random variable that can take values in an interval.

3. a. S = {HHH, HHT, HTH, THH, TTH, THT, HTT, TTT}

b.

x	0	1	2	3
$p(x)$	0.125	0.375	0.375	0.125

$\mu = (0)(0.125) + (1)(0.375) + (2)(0.375) + (3)(0.125) = 1.5$

$\sigma^2 = (0-1.5)^2(0.125) + (1-1.5)^2(0.375) + (2-1.5)^2(0.375) + (3-1.5)^2(0.125) = 0.75$

$\sigma = \sqrt{0.75} \approx 0.866$

c.

y	1	3
$p(y)$	0.75	0.25

$\mu = (1)(0.75) + (3)(0.25) = 1.5$

$\sigma^2 = (1-1.5)^2(0.75) + (3-1.5)^2(0.25) = 0.75$; $\sigma = \sqrt{0.75} \approx 0.866$

d. In this case, the value of w is always 3.

w	3
$p(w)$	1

$\mu = (3)(1) = 3$

$\sigma^2 = (3-3)^2(1) = 0$; $\sigma = 0$

4. a. $P(w < 68) \approx 0.2525$

b. $P(w \geq 75) \approx 0.0478$

c. $P(68 < w < 75) \approx 0.6997$

Unit 21: Binomial Distributions

PREREQUISITES

Students should have a background in basic probability and random variables. These topics are covered in Units 18 – 21.

ADDITIONAL TOPIC COVERAGE

Additional information on this topic can be found in *The Basic Practice of Statistics*, Chapter 13, Binomial Distributions.

ACTIVITY DESCRIPTION

Students can work on this activity either individually or in small groups. For this activity, students revisit the context of inheritance of a trait from parents. Instead of the context of sickle cell disease used in the video, the context switches to eye color. Students simulate the outcomes for 30 families with four children in which both parents have brown eyes but carry a recessive gene for blue eyes. Success is a child with blue eyes.

Two methods for simulating the data are provided in the activity. However, you may decide to ask students to use another method. Here are some additional methods:

- Students could have containers with 100 slips of paper, 25 marked success and 75 marked failure.
- Students could use rand on their graphing calculators.
- Students could use statistical software to generate random data from the Bernoulli distribution with $p = 0.25$.

MATERIALS

Two coins of different denominations per student or group of students or access to Excel. (Optional, depending on choice of method for collecting data)

After students have completed the data collection process and have compiled their data in question 2(a), they should hand in a copy of their first three columns of Table 21.2. Ask them to write their names on their tables. That way you can return them for use in Unit 28, Inference for Proportions. Once you have the individual/group data, combine the results to get a table of class data. Students will need the class data for question 4.

Question 2(b) asks students to find the theoretical probabilities for a binomial distribution with $n = 4$ and $p = 0.25$. This provides an opportunity for students to learn how to use technology to find these probabilities. Statistical software and Excel (BINOM.DIST) have built-in binomial distribution functions. In addition, the TI-84 has a built-in binomial function (binompdf). An alternative is to have students do an online search for binomial calculators. You could also instruct them to use the formulas in the Content Overview to calculate the probabilities and then to check their results using technology.

Students should save their individual/group data and class data from this activity for use in the activity in Unit 28, Inference for Proportions. (Save copies in case students lose theirs.)

THE VIDEO SOLUTIONS

1. Success and failure.

2. Free throws are always shot from the same distance. There is no defensive pressure during play. The probability of making the shot is based on individual player's shooting skills.

3. The probability is 0.25, or 25%. Parents' genetic makeup never changes from child to child so the probability is the same for each child born to these parents.

4. $\mu = np$.

5. The four conditions are listed below.

    1. There are a fixed number of n trials or observations.
    2. The trials are independent.
    3. The trials end in one of two possible outcomes: Success (S) or Failure (F).
    4. The probability of success, p, is the same for all trials.

UNIT ACTIVITY SOLUTIONS

1. Sample answer (student data will differ):

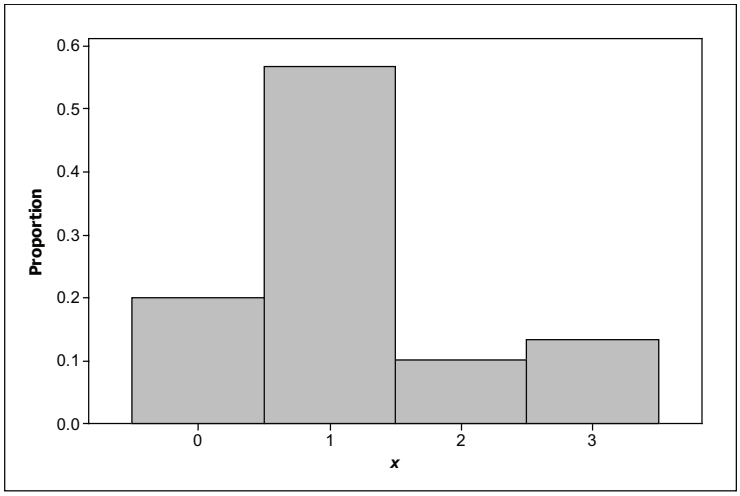
	Trial #				
Sample	1	2	3	4	*x*
1	F	F	F	S	1
2	F	F	F	F	0
3	F	S	S	F	2
4	F	F	S	F	1
5	F	F	F	F	0
6	S	F	F	F	1
7	S	F	F	F	1
8	S	S	S	F	3
9	F	F	F	S	1
10	F	S	F	F	1
11	S	F	F	F	1
12	S	F	S	S	3
13	F	F	F	F	0
14	S	S	S	F	3
15	F	S	F	F	1
16	F	F	F	S	1
17	F	S	F	S	2
18	F	F	S	F	1
19	F	S	F	S	2
20	F	F	S	F	1
21	F	S	F	F	1
22	F	F	F	S	1
23	F	F	S	F	1
24	S	F	S	F	2
25	F	F	F	S	1
26	F	F	F	F	0
27	F	S	S	S	3
28	F	F	F	F	0
29	F	F	F	F	0
30	F	S	F	F	1

2. a. See solution to (b).

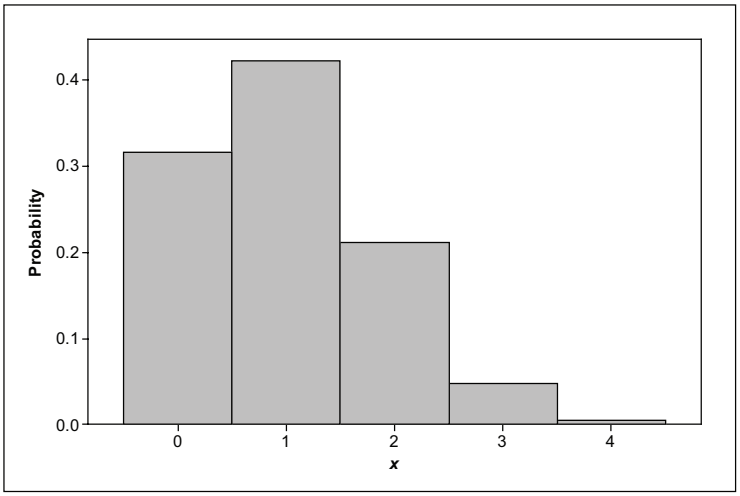
b. Sample answer:

Number of Successes, *x*	Number of Families with *x* successes	Proportion of Families with *x* Successes	Theoretical Probability
0	6	0.2	0.316
1	17	0.567	0.422
2	3	0.1	0.211
3	4	0.133	0.047
4	0	0	0.004

3. a. Sample answer:



b.

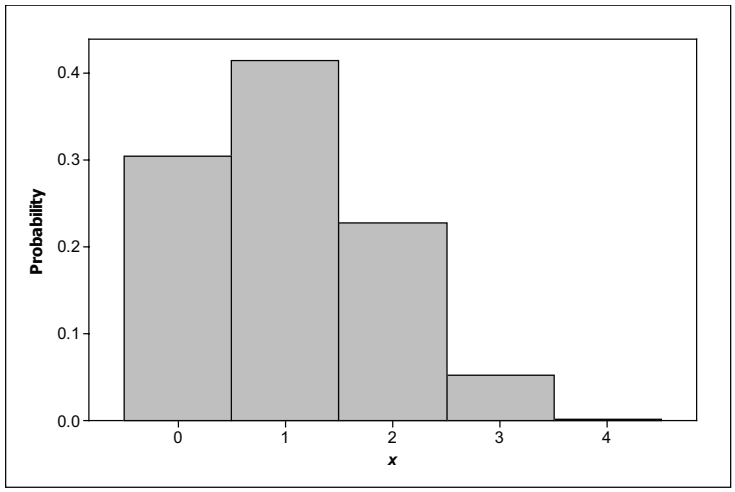


c. Sample answer: They don't look a lot alike. However, the tallest bar is above 1 and the second tallest is above 0 on both graphs.

4. Sample class data is based on 480 samples (from 16 students):

Number of Successes, *x*	Number of Families with *x* successes	Proportion of Families with *x* Successes	Theoretical Probability
0	146	0.304	0.316
1	199	0.415	0.422
2	109	0.227	0.211
3	25	0.521	0.047
4	1	0.002	0.004

a. Sample answer (based on sample class data):



b. The histogram of the class data closely resembles the probability histogram in 3(b).

EXERCISE SOLUTIONS

1. a. Each roll is a trial. Outcomes from one trial do not affect another, so the trials are independent. The probability of success is p = 1/6. However, the number of trials is not fixed. So, this is not a binomial setting.

b. This is an example of a binomial setting (at least approximately). Each individual in the sample is a trial. The trials are independent because the sample was randomly selected. There are a fixed number of trials, n = 10. If an individual believes in extraterrestrial life, that is a success. The probability of success is 0.60. So x has a binomial distribution with n = 10 and p = 0.6.

c. This is not an example of a binomial setting. Success is drawing a red card. The probability of drawing the first red card is 1/2. However, the probability of drawing the second red card is either 26/51 or 25/51 but not 1/2. So, the probability changes as each card is drawn.

2. a. The probabilities are calculated as follows:

$$p(0) = \binom{4}{0}(0.374)^0(0.626)^4 \approx 0.15$$

$$p(1) = \binom{4}{1}(0.374)^1(0.626)^3 = (4)(0.374)^1(0.626)^3 \approx 0.37$$

$$p(2) = \binom{4}{2}(0.374)^2(0.626)^2 = (6)(0.374)^2(0.626)^2 \approx 0.33$$

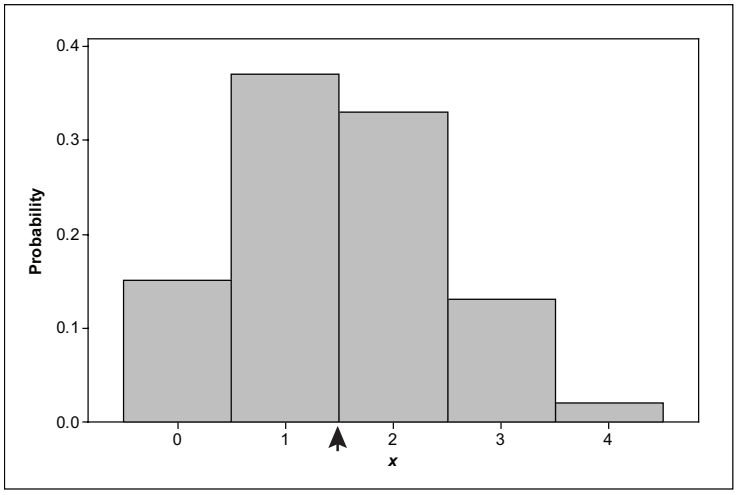
$$p(3) = \binom{4}{3}(0.374)^3(0.626)^1 = (4)(0.374)^3(0.626) \approx 0.13$$

$$p(4) = \binom{4}{4}(0.374)^4(0.626)^0 = (0.374)^4 \approx 0.02$$

b. μ = (0)(0.15) + (1)(0.37) + (2)(0.33) + (3)(0.13) + (4)(0.02) = 1.5

c. μ = (4)(0.374) = 1.496. This answer is slightly less than my answer to (b). The discrepancy is due to the fact that the probabilities in (a) were rounded to two decimals.

d. The mean of 0.1496 is marked with an arrow.



3. a. The three probability distributions appear in the table below.

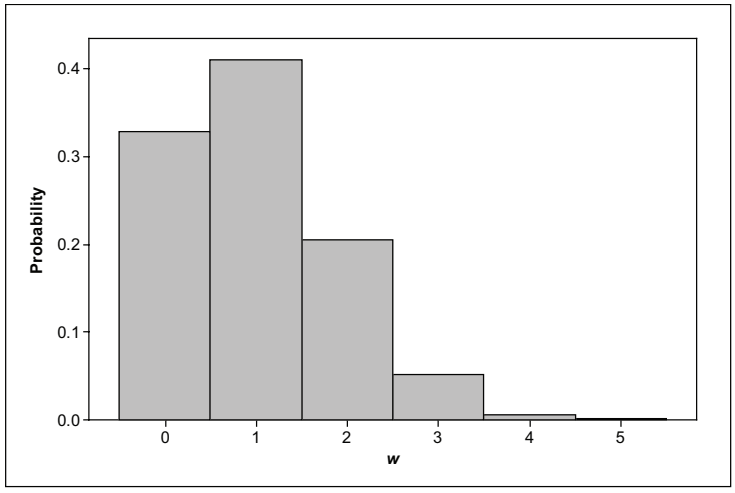
w	0	1	2	3	4	5
$p(w)$	0.3277	0.4096	0.2048	0.0512	0.0064	0.0003
$p(x)$	0.0313	0.1563	0.3125	0.3125	0.1563	0.0313
$p(y)$	0.0003	0.0064	0.0512	0.2048	0.4096	0.3277

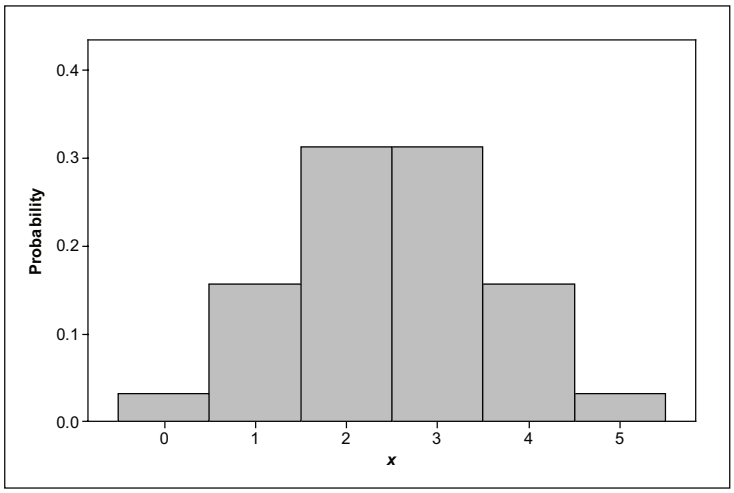
b. $\mu_w = (5)(0.2) = 1.0$; $\sigma_w^2 = (5)(0.2)(0.8) = 0.8$ and $\sigma_w \approx 0.894$

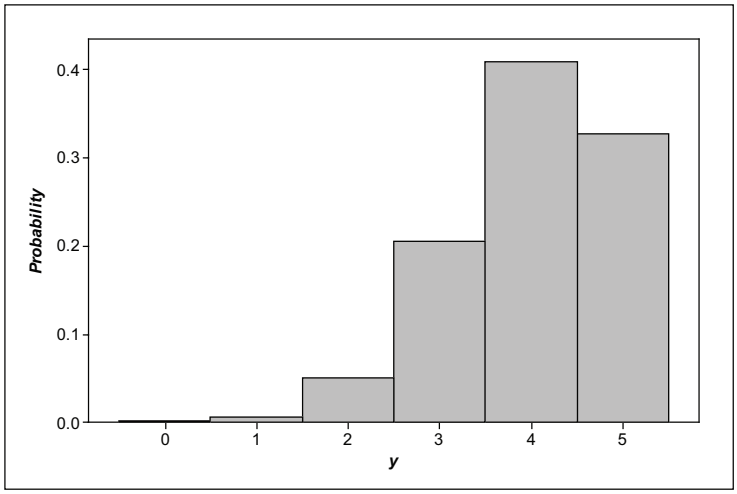
$\mu_x = (5)(0.5) = 2.5$; $\sigma_x^2 = (5)(0.5)(0.5) = 1.25$ and $\sigma_x \approx 1.118$

$\mu_y = (5)(0.8) = 4.0$; $\sigma_y^2 = (5)(0.8)(0.2) = 0.8$ and $\sigma_y \approx 0.894$

c.







d. When p = 0.2, the probability histogram is skewed to the right. When p = 0.5, the probability histogram is symmetric. When p = 0.8 the histogram is skewed to the left. The histograms corresponding to p = 0.2 and p = 0.8 are mirror images of each other about a vertical line drawn at their means. Hence, the values of the random variables are equally spread about their means for both distributions.

4. a. b(200, 0.85)

b. μ = (200)(0.85) = 170 ; $\sigma = \sqrt{(200)(0.85)(0.15)} = \sqrt{25.5} \approx 5.05$

c. We can approximate the binomial distribution by the normal distribution with μ = 170 and $\sigma = 5.05$. Using statistical software we find, $P(165 < x < 180) \approx 0.8151$.

REVIEW QUESTIONS SOLUTIONS

1. a. In this case, n = 100. The sample is randomly drawn so the trials are independent. Let the probability that a randomly selected college student routinely eats breakfast be p – we could estimate this from data, but this value is unknown. It is, however, the same for each of the students in the sample. There are two outcomes to this survey: the student either says yes or he/she does not say yes. So, x is a binomial distribution with n = 100 and p unknown.

b. This is not a binomial setting. We could view each accident as a trial, and an accident that involved alcohol as a success. However, the number of trials is not fixed; it varies from week to week.

c. This is a binomial setting. Random variable x has a binomial distribution with n = 5 and p = 0.75.

2. a. $p(0) = \binom{5}{0}(0.066)^0(0.934)^5 = (0.934)^5 \approx 0.71$

b. $p(1) = \binom{5}{1}(0.066)^1(0.934)^4 = (5)(0.066)(0.934)^4 \approx 0.25$

c. P(no more than one has blood type O-) = $P(x = 0 \text{ or } x = 1) = p(0) + p(1)$ = 0.71 + 0.25 = 0.96

d. P(at least one has blood type O-) $= 1 - p(0) \approx 1 - 0.71 = 0.29$

3. a. $p(0) = \binom{3}{0}(0.25)^0(0.75)^3 \approx 0.42$

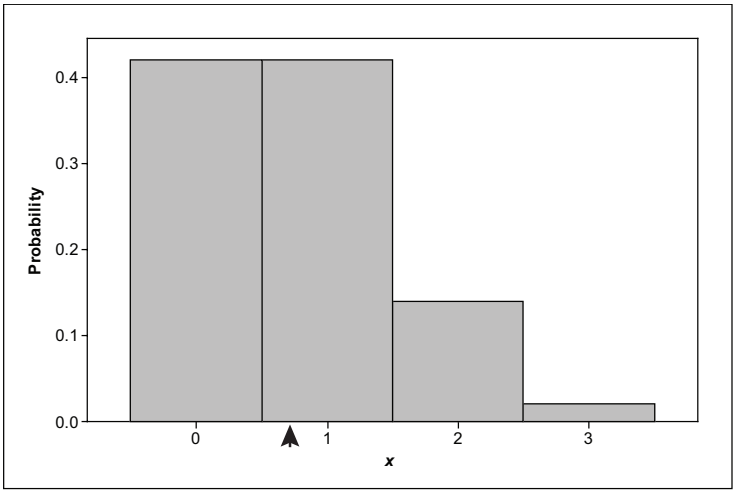
$p(1) = \binom{3}{1}(0.25)^1(0.75)^2 \approx 0.42$

$p(2) = \binom{3}{2}(0.25)^2(0.75) \approx 0.14$

$p(3) = \binom{3}{3}(0.25)^3(0.75)^0 \approx 0.02$

b. μ = (3)(0.25) = 0.75. This means that out of many, many families with three children where both parents are carriers, the average number of children per family who have sickle cell disease is 0.75.

c. The arrow on the horizontal axis marks the mean.



4. a. b(100, 0.31)

b. μ = (100)(0.31) = 31; $\sigma = \sqrt{(100)(0.31)(0.69)} \approx 4.62$

c. Approximate the distribution with a normal distribution with mean 31 and standard deviation 4.62. Using technology, $P(x < 25) \approx 0.09702$.