Chapter 1

Introduction

Check Your Understanding, Page 5:

1. The cars in the student parking lot.
2. He measured the car's model (categorical), year (quantitative), color (categorical), number of cylinders (quantitative), gas mileage (quantitative), whether it has a navigation system (categorical), and weight (quantitative).

Exercises, page 7:

1.1 Type of wood, type of water repellent and paint color are categorical. Paint thickness and weathering time are quantitative.

1.2 Gender, Race, and Smoker status are categorical. Age, systolic blood pressure and level of calcium are quantitative.

1.3 (a) The individuals are the AP Statistics students who completed a questionnaire on the first day of class. (b) The categorical variables are gender (female or male), handedness (right or left), and favorite type of music (classical, gospel, rock, rap, country, R&B, top 40, oldies, etc.). The quantitative variables are height (in inches), amount of time the student is expecting to spend on homework (in minutes per week), and the total value of coins in a student's pocket (in cents). (c) The highlighted individual is a female who is right handed. She is 58 inches tall, spends 60 minutes on homework, prefers Alternative music and had 76 cents in her pocket.

1.4 (a) The individuals are roller coasters opened in 2009. (b) The categorical variables are Roller coaster (the name of the coaster), type (steel or wood) and the design (sit down, flying). The quantitative variables are height (in feet), speed (in mph) and duration (in seconds). (c) The highlighted roller coaster is the Prowler, a wood, sit-down type coaster. It's height is 102.3 feet, its speed is 51.2 mph and the duration of the ride is 150 seconds.

1.5 Student answers will vary; for comparison, recent *U.S. News* rankings have used measures such as academic reputation (measured by surveying college and university administrators), retention rate, graduation rate, class sizes, faculty salaries, student-faculty ratio, percentage of faculty with highest degree in their fields, quality of entering students (ACT/SAT scores, high school class rank, enrollment-to-admission ratio), financial resources, and the percentage of alumni who give to the school. Examples of categorical variables would include region of the country and type of institution (2-year college, 4-year college, university).

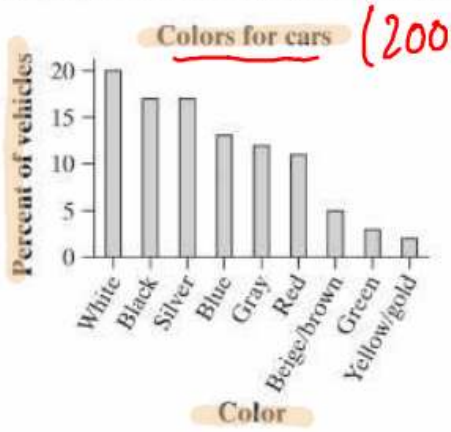
1.6 Student answers will vary. One possible answer is as follows. Reality shows (yes or no indicating whether the student watches reality shows), Music (yes or no indicating whether the student watches music videos, concerts, or documentaries about musicians and singers), Time (the average amount of time, in minutes per day, spent watching television), Network (the average number of network programs—shows, movies, sporting events, etc.— watched per week on ABC, CBS, NBC, and FOX). The categorical variables are Reality shows and Music, and the quantitative variables are Time and Network.

1.7 b

1.8 c

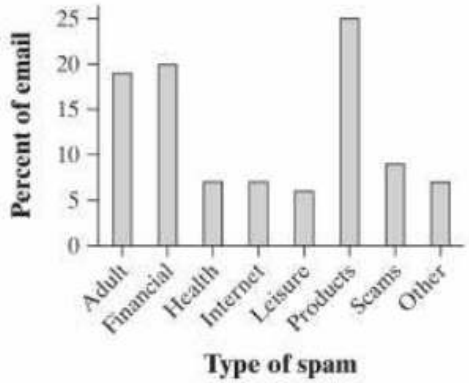
Exercises, page 22:

1.9 (a) The percent of cars with other colors is $100-20-17-17-13-12-11-5-3-2=0\%$. Apparently this list of colors constitutes nearly all possible car colors. (b) A bar graph is given below.



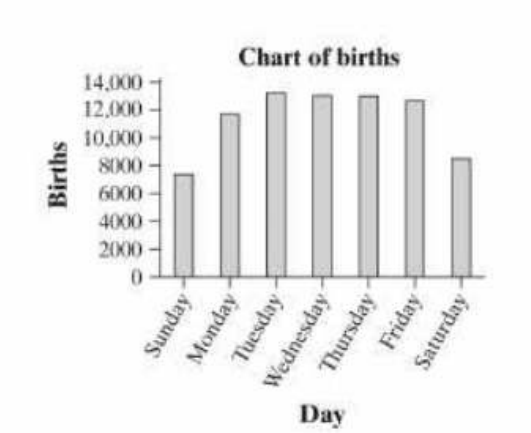
(c) It would be appropriate to make a pie chart of these data because we have all possible colors listed here.

1.10 (a) The percent of spam that occur in the "other" category is $100-19-20-7-7-6-25-9=7\%$. (b) A bar graph is given below.



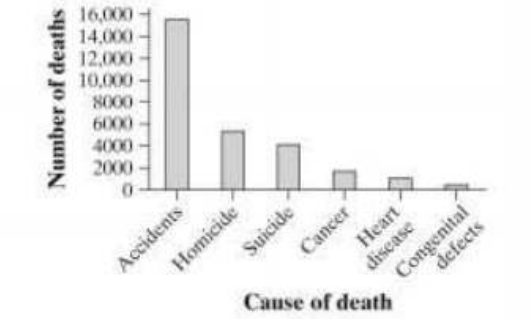
(c) If you include the category of "other," then it would be appropriate to make a pie chart, as all possible types of spam would be included. Without the category of "other," it would not be appropriate to make a pie chart.

1.11 (a) A pie chart would also be appropriate since all days are accounted for in the dataset. A bar graph is given below.



(b) Perhaps induced or C-section births are often scheduled for weekdays.

1.12 (a) A bar graph is given below.

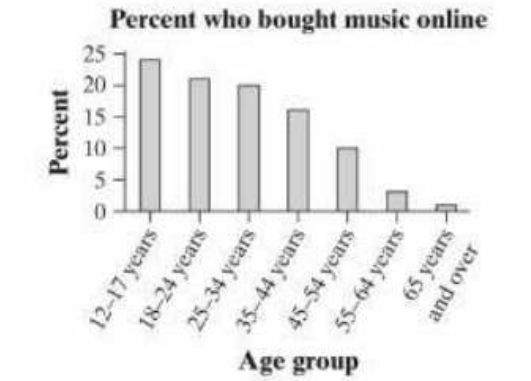


(b) You need to know the total number of deaths among people aged 15-24 in the US in 2005 so that you can figure out how many deaths would be categorized as "other."

1.13 Estimates will vary, but should be close to the actual reported numbers (which can be found at the Census Bureau Web site): 64% Mexican, 9% Puerto Rican.

1.14 Estimates will vary but should be close to: 20% Business, 12% Social Science.

1.15 (a) A pie chart could not be used for these data because the given percents represent fractions of different age groups, rather than parts of a single whole. (b) A bar graph is given below.



1.19 (a) This table describes 133 people, of which 36 were buyers of coffee filters made of recycled paper. (b) To find the marginal distribution of opinion we need to know the total numbers of people with each opinion: $\left(\frac{49}{133}\right)100\% = 36.84\%$ said "higher," $\left(\frac{32}{133}\right)100\% = 24.06\%$ said "the same," and $\left(\frac{52}{133}\right)100\% = 39.10\%$ said "lower." $36.84 + 24.06 = 60.9\%$ of consumers think the quality of the recycled product is the same or higher than the quality of other filters.

1.20 (a) The sum of the six counts is 5375 students. The proportion of these students who smoke is $\frac{1004}{5375} = 0.1868$, so the percent of smokers is 18.68%. (b) The marginal distribution of parents smoking behavior is shown below.

	Neither parent smokes	One parent smokes	Both parents smoke
Count	1356	2239	1780
Percent	25.23%	41.66%	33.12%

1.21 There were 36 buyers and 97 nonbuyers among the respondents, so (for example) $\left(\frac{20}{36}\right)100\% = 55.56\%$ of buyers rated the quality as higher. Similar arithmetic with the buyers and nonbuyers rows gives the two conditional distributions of opinion, shown in the table below. We see that buyers are much more likely to consider recycled filters higher in quality, though 25% still think they are lower in quality. We cannot draw any conclusion about causation: It may be that some people buy recycled filters because they start with a high opinion of recycled products, or it may be that use persuades people that the quality is high.

	Higher	The same	Lower
Buyers	55.56%	19.44%	25.00%
Nonbuyers	29.90%	25.77%	44.33%

1.22 The three conditional distributions are shown in the table below.

	Neither parent smokes	One parent smokes	Both parents smoke
Student does not smoke	86.14%	81.42%	77.53%
Student smokes	13.86%	18.58%	22.47%

The conditional distributions reveal what many people expect—parents have a ~~substantial~~ influence on their children. Students that smoke are more likely to come from families where one or more of their parents smoke.

1.23 The biggest difference between Europeans and Americans in their color choice for cars is the distinction between white/pearl and silver. Americans are much more likely to choose white/pearl, while Europeans are much more likely to choose silver. The only other differences worth mentioning are that Europeans are more likely to choose black or gray than Americans, while Americans are more likely to choose red than Europeans.

1.24 (a) Two side-by-side bar graphs are shown below. Each graph presents a slightly different view of the same percentages.

CATEGORICAL VARIABLES

A class was asked whether homework should count toward the course grade. The answers were simplified into two categories: *Yes* and *No*. The teacher's wife conjectured that the female students are more likely to answer *Yes*. To test this, the students were classified as male or female.

	Male	Female	Row Totals
Yes	4	6	10
No	3	2	5
Column Totals	7	8	15

MARGINAL DISTRIBUTIONS:

1. The marginal distribution of sex:

   Male $\frac{7}{15} = 47\%$ Female $\frac{8}{15} = 53\%$

2. The marginal distribution of responses:

   Yes $\frac{10}{15} = 67\%$ No $\frac{5}{15} = 33\%$

CONDITIONAL DISTRIBUTIONS:

1. Conditional distribution for response if the student was male:

   Yes $\frac{4}{7} = 57\%$ No $\frac{3}{7} = 43\%$

2. Conditional distribution for response if the student was female:

   Yes $\frac{6}{8} = 75\%$ No $\frac{2}{8} = 25\%$

3. Conditional distribution of sex if the response was "no":

   Male $\frac{3}{5} = 60\%$ Female $\frac{2}{5} = 40\%$

4. Conditional distribution of sex if the response was "yes":

   Male $\frac{4}{10} = 40\%$ Female $\frac{6}{10} = 60\%$

CONCLUSION:

20% more ... $\frac{|40-60|}{40} = \frac{20}{40} = 50\%$ inc

2. A simple random sample of adults living in a suburb of a large city was selected. The age and annual income of each adult in the sample were recorded. The resulting data are summarized in the table below.

	Annual Income			
Age Category	$25,000-$35,000	$35,001-$50,000	Over $50,000	Total
21-30	8	15	27	50
31-45	22	32	35	89
46-60	12	14	27	53
Over 60	5	3	7	15
Total	47	64	96	207

(a) What is the probability that a person chosen at random from those in this sample will be in the 31-45 age category? 89/207 = .4299 ≈ 43%

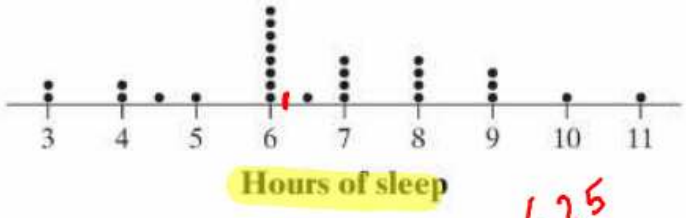
(b) What is the probability that a person chosen at random from those in this sample whose incomes are over $50,000 will be in the 31-45 age category? Show your work. 35/96 = .3645 ≈ 36%

(c) Based on your answers to parts (a) and (b), is annual income independent of age category for those in this sample? Explain.

No - Marginal distribution (.43) does not equal conditional distribution (.36)

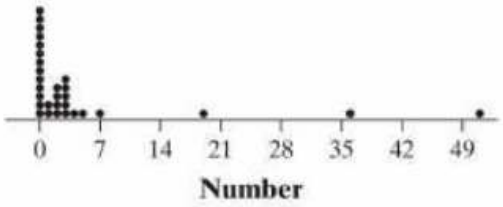
Exercises, page 42:

1.37 (a) The graph is shown below:



(b) The data is roughly symmetric with a center of 6.25 hours. The range is $11-3=8$ hours. There do not appear to be any outliers.

1.38 (a) A dotplot for the total number of gold medals for a sample of countries is shown below.

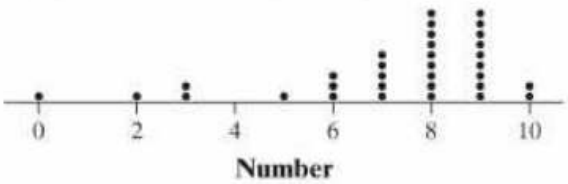


The overall distribution is skewed to the right with a mode of 0, which indicates that many countries did not win any gold medals. China, with 51 gold medals, is clearly unusual, as is the US with 36 and even Great Britain with 19. The rest of the countries earned 7 or fewer – the majority none. (b) No, this does not seem to be a representative sample since 17 out of the 30 countries in the sample (or about 57%) won medals. Overall, only about 27% won medals.

1.39 (a) The two dots above the -2 represent games where the opposing team won by 2 goals. (b) Only two of the 34 differentials are negative, which indicates that the U.S. women's soccer team had a very good season. The team scored at least as many goals as their opponents in 32 of 34 games. In one game they beat the other team by 8 goals, a very unusual event in soccer.

1.40 (a) The dot above 6 represents a car that got 6 mpg more on the highway than it did in city driving. (b) From the dotplot we see that the EPA mpg rating is higher on the highway than in the city for all of the cars. Most of the cars got at least 9 miles per gallon more on the highway than in the city. Five of the cars got 7 more miles per gallon on the highway than in the city while only two cars got less than 7 miles per gallon more on the highway than in the city.

1.41 (a) Answers will vary. One possible dotplot is given below:



(b) As coins get older they get taken out of circulation and new coins are introduced. So most coins in someone's pocket will be from recent years, but there may be a few from previous years.

1.42 The shape of this distribution is fairly uniform. That is, all of the numbers appear with about the same frequency in the last digit of telephone numbers.

1.43 Based on the dotplots, the average ratings were higher for the students in the internal reasons group. Both distributions are roughly symmetric with a range of about 20. But the internal reasons distribution has a center of 21 whereas the external reasons distribution has a center of 17.

1.44 Based on the dotplots, it appears that the claim is partially true. The middle shelf does have cereals with the most sugar. However, the bottom shelf has more cereals that have very little sugar and the top shelf has cereals with a wide range of sugar values.

1.45 (a) If we had not split the stems, most of the data would appear on just a few stems. (b) 16.0%. The high percentage for Utah may be due to the Mormon Church. (c) Ignoring Utah, the data is roughly symmetric around 13% with a spread of roughly 3.5%.

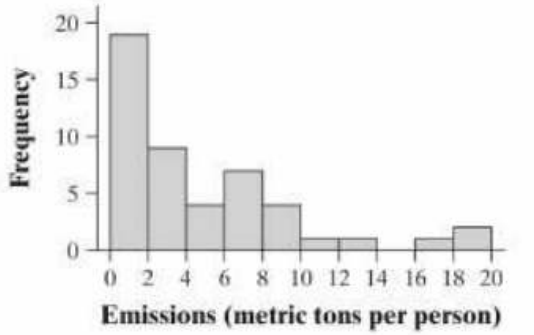
1.46 (a) If we had not split the stems, most of the data would appear on just a few stems. (b) Key: 2|3 means that an 8-ounce serving of that soft drink has 23 mg of caffeine. (c) This distribution is somewhat skewed to the right. The center is 28 mg and the values range from 15 mg to 47 mg. All of these drinks meet the USFDA's limit.

1.47 (a) The stemplots are given below:

Without splitting stems	With splitting stems
6|0 3 5 5 7	6|0 3
7|0 1 2 4 4 8 8 9 9 9	6|5 5 7
8|1 1 3 6 6 7	7|0 1 2 4 4
9|0 6	7|8 8 9 9 9
	8|1 1 3
	8|6 6 7
	9|0
	9|6

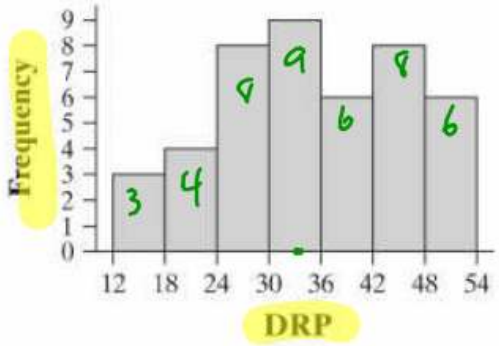
Student preferences may vary, but the split stems on the right show more detail. (b) The distribution is relatively symmetric, with center near 780 mm (the median is 784 mm), and range of $957 - 604 = 353$ mm. (c) Monsoon rainfall was below average in 18 of the 23 El Niño years, and only exceeded 900 mm in one of those years.

1.54 (a) The graph is given below:



(b) The distribution of the emissions is skewed to the right with center near 3 metric tons per person. The range is $19.6 - 0.1 = 19.5$ metric tons per person and there appear to be three outliers: Canada, Australia and the United States.

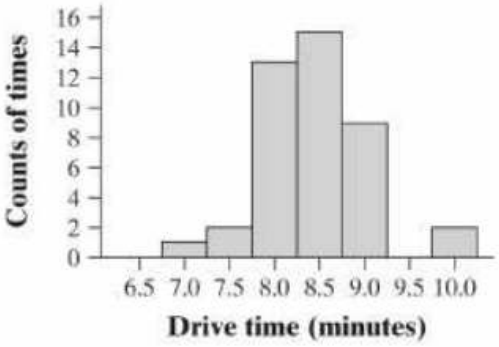
1.55 The graph is given below:



By Hand (classes = 6)

The distribution is somewhat skewed to the left with center at 35. The smallest DRP score is 14 and the largest DRP score is 54, so the scores have a range of 40. There are no gaps or outliers.

1.56 The graph is given below:



This distribution is roughly symmetric. There are no clear outliers, though some may suggest that the approximate 10 minute drive times are outliers.

about the ability of the students taking the exam. If we have a less able group of students, then scores would be lower, even on an easier exam.)

1.66 It appears that the prediction for China in 2050 suggests that until age 65 or so each age group will have more men than women. After age 65, it appears that each age group will have slightly more women than men. The largest group of Chinese will be in their late 50's and early 60's, in 2050. In fact, other than the late 50's and early 60's, the distribution looks quite uniform (equal numbers of people at different age levels) until the mid 80's and above.

1.67 (a) This histogram represents the amount of studying. We would expect most students to study some, but not a huge amount. Any outliers would likely be high outliers, leading to a right-skew distribution. (b) This graph represents the right vs. left-handed variable. About 90% of the population is right-handed and since 0 represents right-handed people we would expect a much higher bar at 0 than at 1. (c) This graph represents the gender of the students. We would expect a more even distribution among the males and females than we would for the right-handed and left-handed students. (d) This histogram represents the heights of the students. The distribution of heights is usually symmetric and bell-shaped.

1.68 (a) Radio stations are categorical, so use a bar graph – one bar for each station. (b) Since hours studied per week is quantitative, either use a dotplot, stemplot, or a histogram. (c) Since calories is quantitative, either use a dotplot, stemplot, or a histogram.

1.69 a

1.70 d

1.71 c

1.72 b

1.73 b

1.74 d

1.75 (a) The individuals are Major League Baseball players who were on the roster on opening day of the 2009 season. (b) There are six variables besides name. Two of them are categorical (team, position) and the other 4 are quantitative (age, height, weight and salary). (c) Age is measured in years, height in feet and inches, weight in pounds and salary in dollars.

1.76 (a) Generally, more people "love" the newer devices such as the iPod, Broadband and HDTV. Those less "loved" are older technologies like cable tv and pay tv. (b) It would not be appropriate to make a pie chart with these data because the categories are not dividing up the whole into pieces. Individuals could be represented in more than one bar.

1.77 (a) There were $10+9+24+61+206+548=858$ observations in all. Of those, $10+61=71$ were elite players. So $\left(\frac{71}{858}\right)100\%=8.28\%$ were elite soccer players. There were $10+9+24=43$ who had arthritis, which means that 5% of the people had arthritis. (b) 10 of the 71 elite players had arthritis. This means that 14.1% had arthritis. 10 of the 43 people who had

10
11
12
13-1
14-1
15
16-11
17-1
18-11
19
20-111
21-1111
22-1
23-1
24
25-1111
26-1111
27
28-111
29
30-111
31-1
32-111
33
34-1
35
36
37
38-11
39
40
41
42
43
44
45
46
47
48
49
50-1

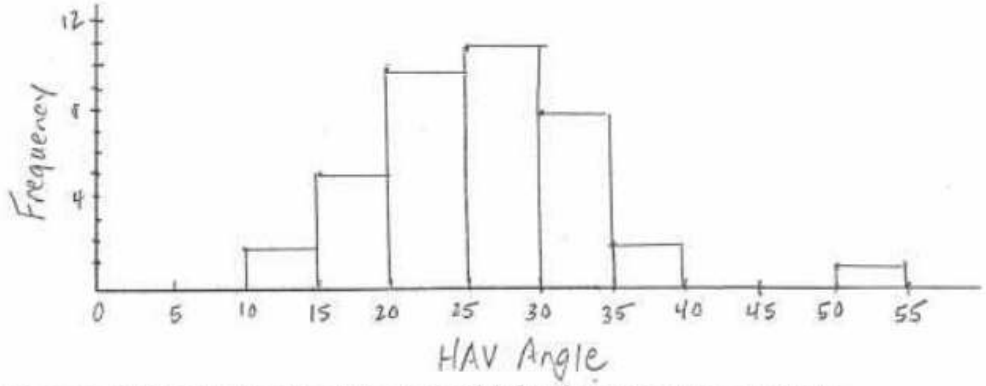
1.2 AP Statistics Name:

Hallux abducto valgus (call it HAV) is a deformation of the big toe that is not common among young people and often requires surgery. Doctors used X-rays to measure the angle (in degrees) of deformity in 38 consecutive patients under the age of 21 who came to a medical center for surgery to correct HAV. The higher the angle measure the more severe the deformity. Here are the data.

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

-5

1. Make a histogram of these data. Choose an appropriate bin width and scale, and label each axis.



2. Write a brief discussion of the distribution of the angle of deformity among young patients needing surgery for this condition.

10
11
12
2 13-1
14-1
15
16-11
5 17-1
18-11
19
20-111
21-1111
10 22-1
23-1
24
25-1111
26-1111
11 27
28-111
29
30-111
31-1
9 32-111
33
34-1
35
36
2 37
38-11
39
40
41
0 42
43
44
45
0 46
47
48
49
1 50-1

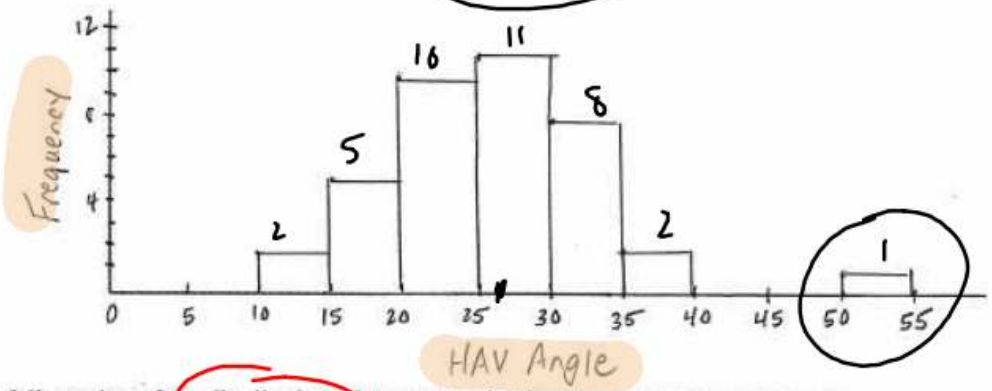
1.2 AP Statistics Name:

Hallux abducto valgus (call it HAV) is a deformation of the big toe that is not common among young people and often requires surgery. Doctors used X-rays to measure the angle (in degrees) of deformity in 38 consecutive patients under the age of 21 who came to a medical center for surgery to correct HAV. The higher the angle measure the more severe the deformity. Here are the data.

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

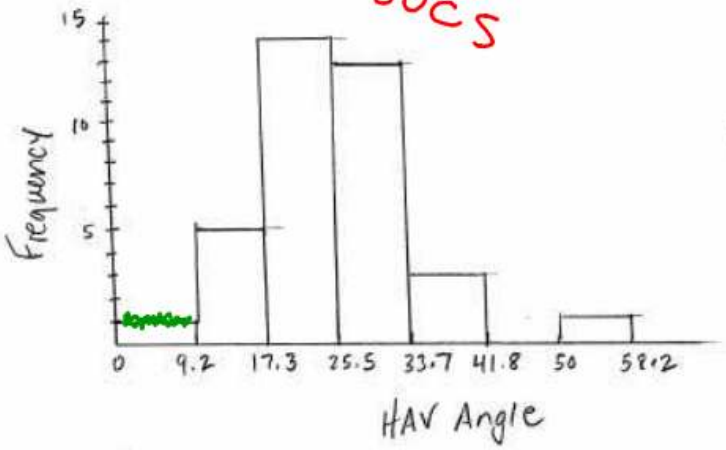
-5

★ **1.** Make a histogram of these data. Choose an appropriate bin width and scale, and label each axis.



2. Write a brief discussion of the distribution of the angle of deformity among young patients needing surgery for this condition.

SOCS



★ Calculator Generated

Shape - Symmetric (Outlier Removed); Skewed Right
Center ≈ Angle of 25°
Spread (Range) = 37°
Outlier at 50 degrees

Below are the resting heart rates of 26 ninth-grade biology students.

61	78	77	81	48	75	70	77	70	76	86	55	65
60	63	79	62	71	72	74	74	64	66	71	66	68

3. Make a stemplot of these data with split stems.

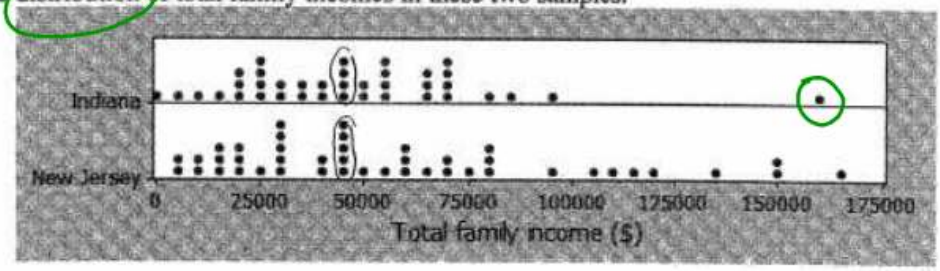
4	8
5	5
6	1 0 3 2 4 6 6 5 8
7	8 7 9 1 5 2 0 4 7 4 0 6 1
8	1 6

4	
4	8
5	
5	5
6	0 1 2 3 4
6	5 6 6 8
7	0 0 1 1 2 4 4
7	5 6 7 7 8 9
8	1
8	6

4|8 = 48 bpm

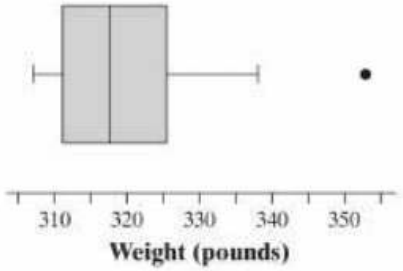
SOCS

4. The dotplots below show the total family income of randomly-chosen individuals from Indiana (38 individuals) and New Jersey (44 individuals). Write a few sentences comparing the distribution of total family incomes in these two samples.



	Indiana	New Jersey
Shape	Symmetric	Skewed Right
Center	Aprox $37,000	Aprox $38,000 > same
Spread	Most incomes close to each other	Incomes more spread-out; 8 households have incomes higher than all but 1 Indiana householders
Outliers	At about $160,000	None

2. The IQR is $326-311=15$. This is the range of the middle half of the data.
3. $1.5\text{IQR}=1.5(15)=22.5$. Any outliers occur below $311-22.5=288.5$ or above $326+22.5=348.5$. There are no observations below 288.5. However, there is one observation above 348.5 – the value 353. It is an outlier.
4. The graph is given below. Note that Minitab computes the quartiles differently and so does not find the highest point to be an outlier.



Check Your Understanding, page 64:

1. The mean is $\left(\frac{67+72+76+76+84}{5}\right)=75$. If the total of all of the heights was the same, but all players were the same height, they would be 75 inches tall.

2. The table is given below:

Observation	Deviation	Squared Deviation
67	$67-75=-8$	$(-8)^2=64$
72	$72-75=-3$	$(-3)^2=9$
76	$76-75=1$	$1^2=1$
76	$76-75=1$	$1^2=1$
84	$84-75=9$	$9^2=81$
Total	0	156

3. The variance is the sum of the squared deviations (taken from the "Total" line) divided by 4 (the number of observations -1). In this case that means that the variance is $\frac{156}{4}=39$ inches squared. The standard deviation is the square root of the variance. In this case that is 6.24 inches.
4. The players heights vary about 6.24 inches from the mean height of 75 inches on average.

Exercises, page 70:

1.79 The mean of Joey's first 14 quiz scores is $\frac{86+84+\ldots+93}{14}=\frac{1190}{14}=85$. If Joey had scored the same number of points on the first 14 quizzes, but the scores had all been the same, then he would have scored an 85 on each quiz.

1.80 The mean weight for the 7 defensive linemen on the 2009 Dallas Cowboys is $\frac{306+305+315+303+318+309+285}{7} = 305.86$ pounds. If the same number of pounds were spread equally among the 7 men, they would all weigh 305.86 pounds.

1.81 (a) Putting the scores in order: 74 75 76 78 80 82 84 86 87 90 91 93 96 98. Since there are 14 scores, the median is the mean of the 7th and 8th scores. Therefore the median is $\frac{84+86}{2} = 85$. About half of the scores are lower than 85 and about half are larger than 85. (b) If Joey had a 0 for the 15th quiz then the sum of his quiz scores would still be 1190 leading to a mean of $\frac{1190}{15} = 79.33$. To find the median, we add the 0 to the beginning of the list in part (a). Since there are now 15 measurements, the median would be the 8th measurement which is 84. Notice that the median did not change much but the mean did. This shows that the mean is not resistant to outliers, but the median is.

1.82 (a) Putting the weights in order: 285 303 305 306 309 315 318. Since there 7 measurements, the median is the 4th. Therefore the median is 306 pounds. About half of the men weigh less than 306 pounds and half weigh more than 306 pounds. (b) If the lowest weight were 265 instead of 285, then the mean would become smaller, but the median would not change. This is because the median is resistant to outliers, but the mean is not.

1.83 The mean is \$60,954 and the median is \$48,097. The distribution of salaries is likely to be quite right skewed because of a few people who have a very large income. When a distribution is skewed to the right, the mean is bigger since the tail values pull the mean toward them.

1.84 The mean house price is \$216,400 and the median is \$172,600. The distribution of house prices is likely to be quite skewed to the right because of a few very expensive homes. When a distribution is skewed to the right, the mean is bigger since the tail values pull the mean toward them.

1.85 The team's annual payroll is $1.2(25) = 30$ or \$30 million. No, you would not be able to calculate the team's annual payroll from the median because you cannot determine the sum of all 25 salaries from the median.

1.86 The mean salary is \$60,000. Seven of the eight employees (everyone but the owner) earned less than the mean. The median is \$22,000. An unethical recruiter would report the mean salary as the "typical" or "average" salary. The median is a more accurate depiction of a "typical" employee's earnings, because it is not influenced by the outlier of \$270,000.

1.87 (a) Estimate the frequencies of the bars (from left to right): 10, 40, 42, 58, 105, 60, 58, 38, 27, 18, 20, 10, 5,5,1 and 3 (although answers may vary slightly, the frequencies must sum to 500). Using these values, we can estimate the mean by adding 2 ten times, 3 forty times,…, and 17 three times. This is equivalent to multiplying the value of each bar (2 through 17) by its frequency or height. This gives us a sum of 3504. The mean is then estimated by dividing by the number of responses: $\bar{x} = \frac{3504}{500} = 7.01$. We estimate the median by finding the average of the 250th and 251st values. The median is found to be 6. (b) Since the median is less than the mean, we would use the median to argue that shorter domain names are more popular.

1.88 (a) Estimate the frequencies of the bars (from left to right): 15, 11, 15, 11, 8, 5, 3, 3, 3 (although answers may vary slightly, the frequencies must sum to 74). We estimate the median by finding the average of the 37th and 38th values. The median is found to be 2. The first quartile is the median of the lower 37 observations. This means that it is the value of the 19th observation. This is found to be 1. The third quartile is the median of the upper 37 observations, which means that it is the value of the 56th observation. This is found to be 4. (b) Using these values, we can estimate the mean by adding 0 fifteen times, 1 eleven times,…, and 8 three times. This is equivalent to multiplying the value of each bar (0 through 8) by its frequency or height. This gives us a sum of 194. The mean is then estimated by dividing by the number of responses:

$\bar{x} = \frac{194}{74} = 2.62.$

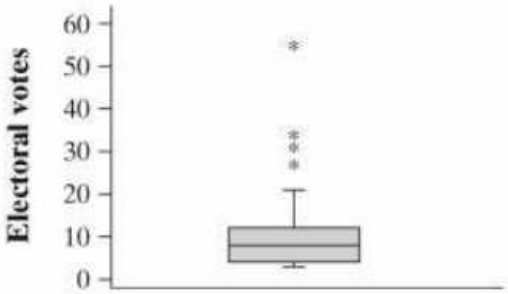
1.89 (a) Putting the data in order we get: 74 75 76 78 80 82 84 86 87 90 91 93 96 98. There are 14 observations here so the first quartile is the median of the bottom 7 observations. This means that it is the value of the 4th observation. We find it to be 78. The third quartile is the median of the top 7 observations, so it is the value of the 11th observation. We find it to be 91. So IQR = 91 − 78 = 13. The middle 50% of the data have a spread of 13 points. (b) Any outliers are below $Q_1 - 1.5\text{IQR}$ or above $Q_3 + 1.5\text{IQR}$. These are computed to be $78 - 1.5(13) = 58.5$ and $91 + 1.5(13) = 110.5$. There are no points outside of these bounds, so there are no outliers.

1.90 (a) Putting the data in order we get: 285 303 305 306 309 315 318. Since there are 7 data points, the median is the 4th and is not included in either the lower half or the upper half of the data set. The first quartile is the middle of the bottom 3 observations, or 303. The third quartile is the middle of the top 3 observations, or 315. Therefore IQR = 315 − 303 = 12. The middle 50% of the weights have a spread of 12 pounds. (b) Any outliers are below $Q_1 - 1.5\text{IQR}$ or above $Q_3 + 1.5\text{IQR}$. These are computed to be $303 - 1.5(12) = 285$ and $315 + 1.5(12) = 333$. While we do have an observation of exactly 285, it is not lower than the boundary we computed, so it is not designated as an outlier. This data set has no outliers.

1.91 (a) Using a stemplot to put the data in order:

```
0   0001133557889          1|4 represents 14 messages sent
1   4
2   5569
3
4   24
5   2
6
7   2
8
9   28
10
11  8
```

We now find that the median is 9, the first quartile is 3 and the third quartile is 43. The IQR is 40. So designate anything below $3 - 1.5(40) = -57$ or above $43 + 1.5(40) = 103$ as outliers. This means that the value of 118 is an outlier. The boxplot produced by computer software is shown below.



(b) Use the median and IQR rather than the mean and standard deviation because the distribution is right skewed.

1.95 (a) The stock fund varied between about -3.5% and 3%. (b) The median return for the stock fund was slightly positive, about 0.1%, while the median real estate fund return appears to be close to 0%. (c) The stock fund is much more variable. It has higher positive returns, but also higher negative returns.

1.96 All five income distributions are skewed to the right. As highest education level rises, the median, quartiles, and extremes rise—that is, all five points on the boxplot increase. Additionally, the width of the box (the IQR) and the distance from one extreme to the other (the difference between the 5th and 95th percentiles) also increase, meaning that the distributions become more and more spread out.

1.97 (a) The mean phosphate level is $\bar{x} = \frac{32.4}{6} = 5.4$ mg/dl. The standard deviation is $s_x = \sqrt{\frac{2.06}{5}} = 0.6419$ mg/dl. Details are provided below.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
5.6	0.2	0.04
5.2	-0.2	0.04
4.6	-0.8	0.64
4.9	-0.5	0.25
5.7	0.3	0.09
6.4	1.0	1.00
32.4	0	2.06

(b) The typical phosphate level is an average of 0.6419 mg/dl different from the mean level.

1.98 (a) Mean = $\bar{x} = \frac{7+7+9+9}{4} = 8$. The average amount of sleep that the first four students got last night was 8 hours. The deviations are $7-8=-1$, $7-8=-1$, $9-8=1$, $9-8=1$. The standard deviation is then $\sqrt{\frac{(-1)^2+(-1)^2+1^2+1^2}{4-1}} = \sqrt{\frac{4}{3}} = 1.15$. (b) The distance between a typical response and the mean response is 1.15 hours. (c) No, it would not be safe to make this generalization. This is not a random sample and it is not likely that the first 4 students to arrive in

the classroom are representative of the entire class in terms of the amount of sleep they got last night.

1.99 (a) It looks like the distribution is skewed to the right because the mean is much larger than the median. (b) The standard deviation is 21.6974. The distance between a typical response and the mean response is $21.6974. (c) The first quartile is 19.27 and the third quartile is 45.4 so the IQR is $45.4 - 19.27 = 26.13$. Any points below $19.27 - 1.5(26.13) = -19.925$ or above $45.4 + 1.5(26.13) = 84.595$ are outliers. Since the maximum point is 93.34, there are outliers.

1.100 (a) It would appear that the distribution for the female doctors is more likely to be symmetric since the mean and median are relatively close together (19.1 and 18.5 respectively). The mean and median for the male doctors are quite far apart (41.333 and 34 respectively). (b) The IQR measures the range of the middle 50% of the data. This does not take outliers into consideration. The standard deviation, however, uses every point and is not resistant to outliers. So, while the middle 50% of the data set may look very similar, if one data set has many more outliers, it will have a larger standard deviation. (c) It does appear that males perform more C-sections. Each of the numbers in the 5-number summary was larger for the males than for the females.

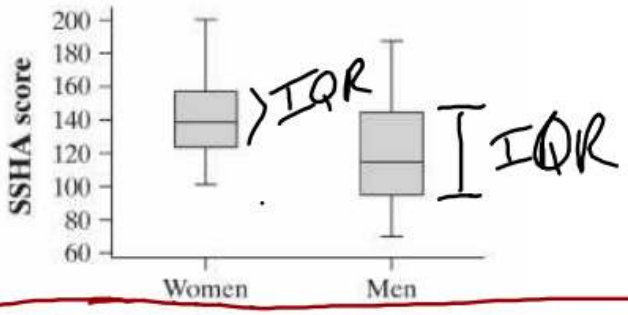
1.101 Yes, IQR is resistant. Answers will vary. Consider the simple data set 1, 2, 3, 4, 5, 6, 7, 8. The median = 4.5, $Q_1 = 2.5$, $Q_3 = 6.5$, and IQR = 4. Changing any value outside the interval between Q_1 and Q_3 will have no effect on the IQR. For example, if 8 is changed to 88, the IQR will still be 4.

1.102 Variable A has a larger standard deviation because more of the observations have values further from the mean. Because of the bell-shape to the distribution of variable B, more of the observations have values quite close to the mean.

1.103 (a) One possible answer is 1, 1, 1, 1. (b) 0, 0, 10, 10. (c) For (a), any set of four identical numbers will have $s_x = 0$. For (b), the answer is unique; here is a rough description of why. We want to maximize the "spread-out"-ness of the numbers (which is what standard deviation measures), so 0 and 10 seem to be reasonable choices based on that idea. We also want to make each individual squared deviation—$(x_1 - \bar{x})^2$, $(x_2 - \bar{x})^2$, $(x_3 - \bar{x})^2$ and $(x_4 - \bar{x})^2$—as large as possible. If we choose 0, 10, 10, 10—or 10, 0, 0, 0—we make the first squared deviation 7.5^2, but the other three are only 2.5^2. Our best choice is two at each extreme, which makes all four squared deviations equal to 5^2.

1.104 (a) This could be used to measure the center since we are averaging the 25th and 75th percentiles, effectively finding a middle point between these positions. It would be resistant to outliers, because any outliers would occur further out in the tail. (b) This could be used as a measure of spread since it finds the distance between the smallest and largest values and then divides by 2. It gives half of the range. This measure, however, would not be resistant to outliers, since if outliers exist, they would be, by definition, either the max, the min, or both.

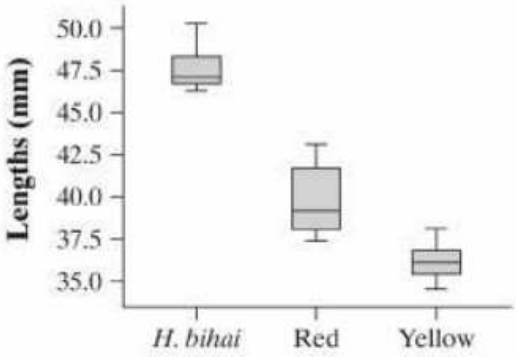
1.105 *State:* Do the data indicate that women have better study habits and attitudes towards learning than men? *Plan:* We will draw side-by-side boxplots for each group. We will compute the 5-number summary, the mean and the standard deviation for the scores of each group. Then we will compare the groups using both graphical and numerical summaries. *Do:* The boxplots are given below, as is a table of the numerical summaries.



Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Women	18	141.06	26.44	101.00	123.25	138.50	156.75	200.00
Men	20	121.25	32.85	70.00	95.00	114.50	144.50	187.00

Conclude: It appears from the boxplot and the numerical summaries that the women have higher values for all of the components of the 5- number summary. They also have a higher mean and a smaller standard deviation. This means that not only are their scores higher, but there is less variability to their scores.

1.106 *State:* Do the different types of flowers have different lengths? *Plan:* We will look at side-by-side boxplots, the 5-number summaries, the means and the standard deviations for all three kinds of flowers. *Do:* The boxplots are given below, as is a table of the numerical summaries.



Variable	Minimum	Q1	Median	Q3	Maximum
H. bihai	46.340	46.690	47.120	48.293	50.260
red	37.400	38.070	39.160	41.690	43.090
yellow	34.570	35.450	36.110	36.820	38.130

Variable	Mean	StDev
H. bihai	47.597	1.213
red	39.711	1.799
yellow	36.180	0.975

Conclude: H. bihai is clearly the tallest variety—the shortest *bihai* was over 3 mm taller than the tallest red. Red is generally taller than yellow, with a few exceptions. Another noteworthy fact: The red variety is more variable than either of the other varieties. Our overall conclusion, then, is