To help you master the basics, many of our exercises will continue to tell you what todo—make a histogram, find the five-number summary, and so on. Of course, real-world statistical problems don't come with detailed instructions. You will encounter some exercises that are more realistic, especially in the later chapters of the book. Use the four-step process as a guide to solving and reporting these problems. They are marked with the four-step icon, as the following example illustrates.

EXAMPLE 2.11 Highly Superior Autobiographical Memory

STATE: Some individuals have the ability to recall accurately vast amounts of autobiographical information without mnemonic tricks or extra practice. This ability is called Highly Superior Autobiographical Memory (HSAM). A study recruited 11 adults with confirmed HSAM and 15 control individuals of similar age without HSAM. All study participants were given a battery of cognitive and behavioral tests with the goal of finding out what might explain this extraordinary ability.[12]

First, autobiographical memory was assessed by asking each participant to recall in detail five important personal events. These events were selected by the researchers, and the participants did not know ahead of time which events they would have to recall. Answer accuracy was then verified from documents and interviews, and each correct detail was scored as one point. Individual total scores on this verifiable autobiographical memory task are displayed below.

HSAM	22	23	26	26	33	34	38	39	39	46	47				
Control	5	5	5	6	7	7	10	10	11	12	13	16	18	22	23

Do individuals with HSAM and without HSAM display distinct distributions of verifiable autobiographical memory scores? How do the mean scores compare?

PLAN: Use graphs and numerical descriptions to describe and compare these two distributions of verifiable autobiographical memory scores.

SOLVE: Dotplots work best for data sets of these sizes. Figure 2.8 displays stacked dotplots to facilitate comparison. The control group has a somewhat right-skewed distribution, so we might choose to compare the five-number summaries. But because the researchers plan to use $\bar{x}$ and s for further analysis, we instead calculate these measures:

Group	Mean score	Standard deviation
HSAM	33.9	8.8
Control	11.3	6.0

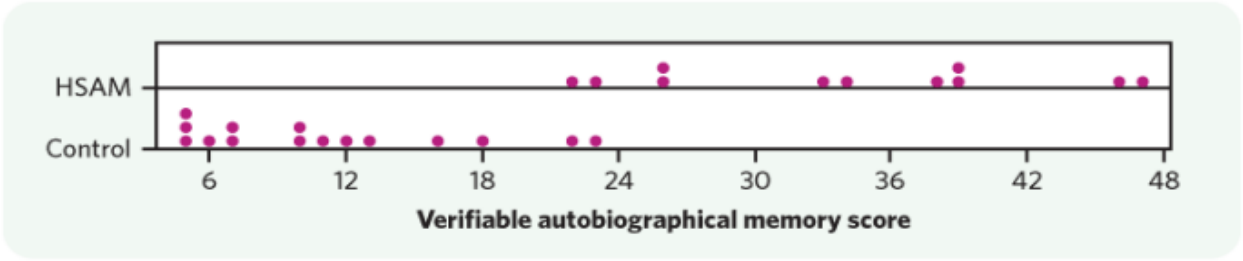

Figure 2.8
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.8 Dotplots comparing the distributions of scores on a test of verifiable autobiographical memory.

CONCLUDE: The two groups differ so much in verifiable autobiographical memory scores that there is little overlap among them. Overall, individuals with HSAM have higher scores than control individuals

without HSAM. The mean scores are 33.9 for the HSAM group compared with only 11.3 for the control group, despite similar variation in individual scores (standard deviations 8.8 and 6.0, respectively). This first test confirms that individuals with HSAM have better recall of autobiographical events than ordinary individuals do. The researchers can now proceed to compare cognitive and behavioral traits in both groups.

APPLY YOUR KNOWLEDGE

2.13 Highly Superior Autobiographical Memory, continued. The study in Example 2.11 assessed common obsessional symptoms in the HSAM and the control individuals, as previous findings had suggested a possible obsessional component to HSAM. The study participants completed the Leyton Obsessional Inventory test, which has a maximum score of 30 points. The results for the 11 HSAM individuals and for 14 of the controls are displayed below (one individual in the control group left before completing this test).

HSAM	2	4	5	8	8	9	9	10	11	11	12			
Control	1	2	2	2	2	4	4	4	5	7	8	8	12	12

How do individuals with and without HSAM compare in terms of obsessional score? Follow the four-step process in reporting your work.

CONCLUDE: The two groups differ so much in verifiable autobiographical memory scores that there is little overlap among them. Overall, individuals with HSAM have higher scores than control individuals without HSAM. The mean scores are 33.9 for the HSAM group compared with only 11.3 for the control group, despite similar variation in individual scores (standard deviations 8.8 and 6.0, respectively). This first test confirms that individuals with HSAM have better recall of autobiographical events than ordinary individuals do. The researchers can now proceed to compare cognitive and behavioral traits in both groups.

APPLY YOUR KNOWLEDGE

2.13 Highly Superior Autobiographical Memory, continued. The study in Example 2.11 assessed common obsessional symptoms in the HSAM and the control individuals, as previous findings had suggested a possible obsessional component to HSAM. The study participants completed the Leyton Obsessional Inventory test, which has a maximum score of 30 points. The results for the 11 HSAM individuals and for 14 of the controls are displayed below (one individual in the control group left before completing this test).

HSAM	2	4	5	8	8	9	9	10	11	11	12			
Control	1	2	2	2	2	4	4	4	5	7	8	8	12	12

How do individuals with and without HSAM compare in terms of obsessional score? Follow the four-step process in reporting your work.

CHAPTER 2 SUMMARY

- A numerical summary of a quantitative distribution should report at least its **center** and its **spread** or **variability.**
- The **mean** $\bar{x}$ and the **median** M describe the center of a distribution in different ways. The mean is the arithmetic average of *all* the observations, and the median is the midpoint of the values.
- When you use the median to indicate the center of the distribution, describe the spread by giving the **five-number summary** consisting of the median, the quartiles, and the smallest and largest individual observations. The **first quartile** Q_1 has one-quarter of the observations below it, and the **third quartile** Q_3 has three-quarters of the observations below it.
- When you use the mean to indicate the center of a sample distribution, describe the spread by giving the **standard deviation** s. The standard deviation is the square root of the **variance** s^2. They are both zero when there is no spread and become larger as the spread increases.
- A **resistant measure** of any aspect of a quantitative distribution is relatively unaffected by changes in the numerical value of a small proportion of the total number of observations, no matter how large these changes are.
- The median and quartiles are resistant, but the mean and standard deviation are not. The mean and standard deviation are good descriptions for symmetric distributions without outliers. The five-number summary is a better description for skewed distributions. However, numerical summaries do not fully describe the shape of a distribution. Always plot your data.
- **Boxplots** based on the five-number summary are useful for comparing several distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the extremes and show the full spread of the data.
- A statistical problem has a real-world setting. You can organize many problems by using the four-step method: **State, Plan, Solve,** and **Conclude.**

Boxplots based on the five-number summary are useful for comparing several distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the extremes and show the full spread of the data.

- A statistical problem has a real-world setting. You can organize many problems by using the four-step method: **State, Plan, Solve,** and **Conclude.**

Boxplots based on the five-number summary are useful for comparing several distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the extremes and show the full spread of the data.

- A statistical problem has a real-world setting. You can organize many problems by using the four-step method: **State, Plan, Solve,** and **Conclude.**

CHAPTER 2 EXERCISES

2.24 Food oils and health. Table 1.2 (page 34) gives the ratio of omega-3 to omega-6 fatty acids in common food oils. Exercise 1.36 asked you to plot the data. The distribution is strongly right-skewed with a high outlier. Do you expect the mean to be greater than the median, about equal to the median, or less than the median? Why? Calculate $\bar{x}$ and M and verify your expectation.

2.25 Lyme disease. Figure 1.8 (page 19) shows the age in years of 241,931 patients diagnosed with Lyme disease in the United States between 1992 and 2006. Give a brief description of the important features of the distribution. Explain why no numerical summary would appropriately describe this distribution.

2.26 Making resistance visible. In the *Mean and Median* applet, place three observations on the line by clicking below it: two close together near the center of the line, and one somewhat to the right of these two.

APPLET

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

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

2.27 Anorexia nervosa. Figure 1.9 (page 20) is a histogram of the distribution of age at onset of anorexia for 691 Canadian girls diagnosed with the disorder. If you round the age to whole numbers of years, the first bar of the histogram (the first class) would include all girls diagnosed during their 11th year. With a little care, you canfind the median and the quartiles from the histogram. What are these numbers? How did you find them?

2.28 Laughing behavior. A study of freely forming groups in bars throughout Europe examined the number of individuals found in groups whose members were laughing together. The study reported on a total of 501 laughing groups, distributed as follows:[14]

Number of individuals in group	2	3	4	5	6
Number of groups	254	168	52	21	6

a. Obtain the five-number summary for these data and display the results in a hand-drawn boxplot. Describe the distribution of the number of individuals in the freely forming laughing groups examined by this study.

b. Based on your boxplot, do you expect the mean laughing group size to be smaller, similar, or larger than the median group size? Explain your reasoning.

c. Compute the mean laughing group size, paying attention to the fact that the data are presented in frequencies (see Exercise 2.3 for reference). Does the computed mean fit your expectation as stated in part b?

2.29 Metabolic rate. In Example 2.7 you examined the metabolic rates of 7 men. Here are the metabolic rates for 12 women from the same study:

995 1425 1396 1418 1502 1256 1189
913 1124 1052 1347 1204

a. The most common methods for formal comparison of two groups use $\bar{x}$ and s to summarize the data. What kinds of distributions are best summarized by $\bar{x}$ and s?
b. Make a summary graph comparing the metabolic rates of the 7 men and 12 women, as in Figure 2.4. What can you conclude about these two groups from your graph?

2.30 Behavior of the median. Place five observations on the line in the *Mean and Median* applet by clicking below it.

a. Add one additional observation *without changing the median*. Where is your new point?
b. Use the applet to convince yourself that when you add yet another observation (there are now seven in all), the median does not change no matter where you put the seventh point. Explain why this must be true.

2.31 Nanomedicine. In Exercise 1.41 (page 36) you graphed the distribution of ovarian tumor increases under two experimental conditions: a new nanoparticle-based delivery system for a suicide gene therapy or an inactive buffer solution.

a. Make a boxplot comparing tumor increase under the two conditions and compute the mean and standard deviation for each condition.
b. Write a short description of the experimental results based on your work in a.

2.32 Guinea pig survival times. Here are the survival times (in days) of 72 guinea pigs after they were injected with infectious bacteria in a medical experiment:[15]

43	45	53	56	56	57	58	66	67	73	74	79
80	80	81	81	81	82	83	83	84	88	89	91
91	92	92	97	99	99	100	100	101	102	102	102
103	104	107	108	109	113	114	118	121	123	126	128
137	138	139	144	145	147	156	162	174	178	179	184
191	198	211	214	243	249	329	380	403	511	522	598

a. Plot and describe the distribution of survival times. As is often the case with survival times, this distribution is strongly skewed to the right.
b. Which numerical summary would you choose for these data? Explain why.
c. Obtain your chosen summary. How does it reflect the skewness of the distribution?

2.33 A standard deviation contest. This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 10, with repeats allowed.

a. Choose four numbers that have the smallest possible standard deviation.
b. Choose four numbers that have the largest possible standard deviation.
c. Is more than one choice possible in either a or b? Explain.

2.34 Does breastfeeding weaken bones? Breastfeeding mothers secrete calcium into their milk. Some of the calcium may come from their bones, so mothers may lose bone mineral content. Researchers compared 47 breastfeeding women with 22 women of similar age who were neither pregnant nor lactating. They measured the percent change in the mineral content of the women's spines over three months. A negative value indicates a loss in mineral content. Here are the data:[16]

Breastfeeding women					
−4.7	−2.5	−4.9	−2.7	−0.8	−5.3
−8.3	−2.1	−6.8	−4.3	2.2	−7.8

−3.1	−1.0	−6.5	−1.8	−5.2	−5.7
−7.0	−2.2	−6.5	−1.0	−3.0	−3.6
−5.2	−2.0	−2.1	−5.6	−4.4	−3.3
−4.0	−4.9	−4.7	−3.8	−5.9	−2.5
−0.3	−6.2	−6.8	1.7	0.3	−2.3
0.4	−5.3	0.2	−2.2	−5.1	
Other women					
2.4	0.0	0.9	−0.2	1.0	1.7
2.9	−0.6	1.1	−0.1	−0.4	0.3
1.2	−1.6	−0.1	−1.5	0.7	−0.4
2.2	−0.4	−2.2	−0.1		

Do the data show distinctly greater bone mineral loss among the breastfeeding women? Be sure to consider in your interpretation the understanding that a bone mineral loss is reflected by a negative value here. Follow the four-step process illustrated by Example 2.11.

2.35 Cicadas as fertilizer? Every 17 years, swarms of cicadas emerge from the ground in the eastern United States, live for about six weeks, then die. (There are several "broods," so we see cicada eruptions more often than every 17 years.) So many cicadas die that their bodies may serve as fertilizer and increase plant growth. In an experiment, a researcher added 10 cicadas under some plants in a natural plot of American bellflowers in a forest, leaving other plants undisturbed. One of the variables studied was the size of seeds produced by the plants. Here are data (seed mass in milligrams) for 39 plants fertilized with cicadas and 33 undisturbed plants:[17]

Cicada plants				Undisturbed plants			
0.237	0.277	0.241	0.142	0.212	0.188	0.263	0.253
0.109	0.209	0.238	0.277	0.261	0.265	0.135	0.170
0.261	0.227	0.171	0.235	0.203	0.241	0.257	0.155
0.276	0.234	0.255	0.296	0.215	0.285	0.198	0.266
0.239	0.266	0.296	0.217	0.178	0.244	0.190	0.212
0.238	0.210	0.295	0.193	0.290	0.253	0.249	0.253
0.218	0.263	0.305	0.257	0.268	0.190	0.196	0.220
0.351	0.245	0.226	0.276	0.246	0.145	0.247	0.140
0.317	0.310	0.223	0.229	0.241			
0.192	0.201	0.211					

Do these data support the idea that dead cicadas can serve as fertilizer? Follow the four-step process in your work.

2.36 Daily activity and obesity. People gain weight when they take in more energy from food than they expend. Table 2.1 compares volunteer subjects who were lean with others who were mildly obese. None of the subjects followed an exercise program. The subjects wore sensors that recorded every move for 10 days. The table shows the average minutes per day spent in activity (standing and walking) and in lying down.[18] Compare the distributions of time spent actively for lean and obese subjects as well as the distributions of time spent lying down. How does the behavior of lean and mildly obese people differ? Follow the four-step process in your work.

TABLE 2.1 Time (minutes per day) active and lying down by lean and obese subjects

Lean subjects			Obese subjects		
Subject	Stand/Walk	Lie	Subject	Stand/Walk	Lie
1	511.100	555.500	11	260.244	521.044
2	607.925	450.650	12	464.756	514.931
3	319.212	537.362	13	367.138	563.300
4	584.644	489.269	14	413.667	532.208
5	578.869	514.081	15	347.375	504.931
6	543.388	506.500	16	416.531	448.856
7	677.188	467.700	17	358.650	460.550
8	555.656	567.006	18	267.344	509.981
9	374.831	531.431	19	410.631	448.706
10	504.700	396.962	20	426.356	412.919

2.37 Logging in the rainforest. "Conservationists have despaired over destruction of tropical rainforest by logging, clearing, and burning." These words begin a report on a statistical study of the effects of logging in Borneo.[19] Researchers compared forest plots that had never been logged (Group 1) withsimilar plots nearby that had been logged 1 year earlier (Group 2) and 8 years earlier (Group 3). All plots were 0.1 hectare in area. Here are the counts of trees for plots in each group:

Group 1	27	22	29	21	19	33	16	20	24	27	28	19
Group 2	12	12	15	9	20	18	17	14	14	2	17	19
Group 3	18	4	22	15	18	19	22	12	12			

To what extent has logging affected the count of trees? Follow the four-step process in reporting your work.

Exercises 2.38 to 2.41 make use of the optional material on the $1.5 \times IQR$ *rule for suspected outliers.*

2.38 Mercury levels, continued. In Exercise 2.4 you obtained the five-number summary of the distribution of blood mercury levels among 4134 pregnant British women enrolled in a scientific study.

a. Use the $1.5 \times IQR$ rule to determine whether the smallest and the largest values quoted qualify as suspected outliers.

b. The published quote included a number of percentiles besides those making up the five-number summary. Do any of the quoted values beside the minimum and the maximum qualify as suspected outliers based on the $1.5 \times IQR$ rule?

c. Based on your analysis in parts a and b, does the distribution of blood mercury levels in this study include at least one actual outlier or is it simply skewed? Explain your reasoning.

2.39 Aggression and social status in macaque monkeys. The boxplots in Figure 2.9 summarize the number of aggressive behaviors directd at female macaque monkeys of differing social status living in experimentally arranged groups.[20] These are modified boxplots that indicate suspected outliers by a dot, using the $1.5 \times IQR$ rule. Based on these plots, describe how "aggressions received" by female macaque monkeys vary depending on their dominance rank within the group.

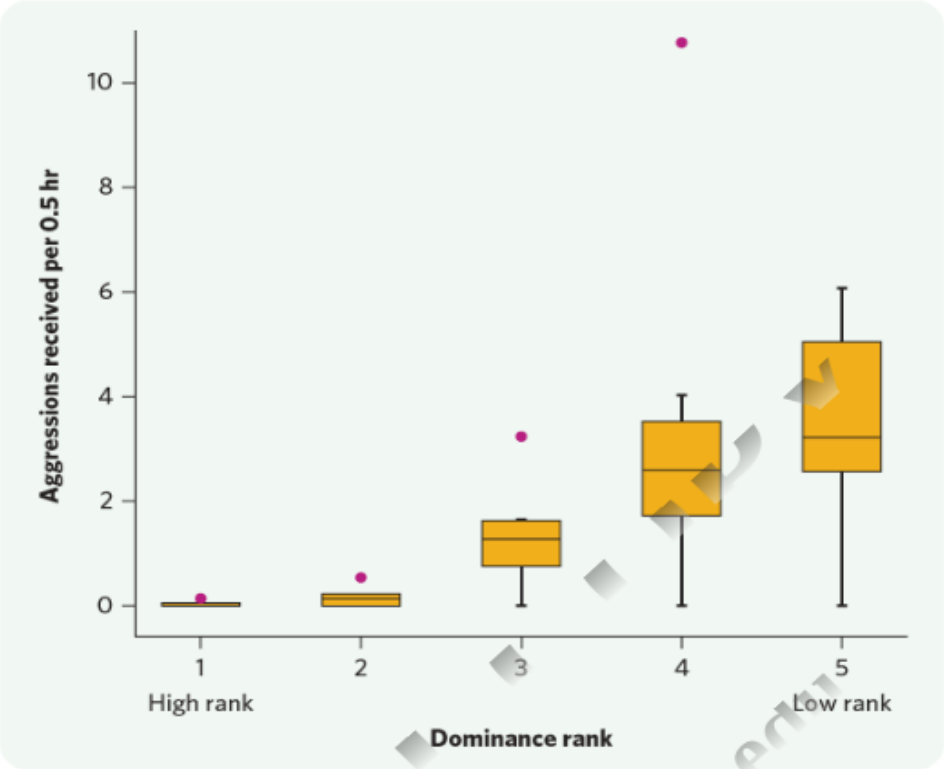

Figure 2.9
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.9 Boxplots of the distributions of number of aggressions received per half hour, for female macaque monkeys of differing social status.

2.40 Golden orb weavers. Exercise 2.1 described a study of female golden orb weaver spiders. The study also reported the body mass (in grams) for each of the 21 spiders. Here are the data:

0.04	0.11	0.16	0.07	0.13	0.1	0.17
0.25	0.36	0.33	0.29	0.14	0.32	0.57
0.31	0.79	0.49	0.64	0.6	0.99	0.81

a. Give the five-number summary, mean, and standard deviation of this distribution. How do the mean and the median compare?
b. Does the $1.5 \times IQR$ rule identify any suspected outliers?
c. What do the results in parts a and b suggest about the distribution of body mass in female golden orb weavers? Make a dotplot to confirm your interpretation.

2.41 Young Americans. The dotplot in Figure 1.16 (page 31) displays the distribution of the percents of individuals younger than age 18 in each of the 50 states and the District of Columbia. Note that the values in the dotplot are reported in increments of 0.5 percentage points.

a. Give the five-number summary of this distribution.
b. Does the $1.5 \times IQR$ rule identify the minimum (District of Columbia) and maximum (Utah) as suspected outliers? Does it also flag any other states?
c. The mean for these 51 values is 23.755. How does it compare with the median? Explain how this is possible.

Exercises marked with the Large Data Set *icon guide you through the comprehensive analysis of more complex data sets characterized by a large number of observations, a number of variables, or both. You can find many more such exercises in the book's review chapters (Chapters 8, 16, and 25). Note that short answers are not available for these exercises so that they can be used for class work or assignments.*

2.42 Elderly health. A study examined the medical records of elderly patients to determine whether there are differences between men and women in their calcium or inorganic phosphorus blood levels (both in millimoles per liter, mmol/l). The data file *Large.Calcium* contains the data.[21]

a. Make side-by-side boxplots comparing the calcium levels for men $(sex = 1)$ and for women $(sex = 2)$. Do the same for the inorganic phosphorus levels. What do you conclude?

b. Another purpose of the study was to determine whether different laboratories yield substantially different results. Use boxplots to compare the calcium levels obtained by the six different labs. Do you see notable differences? Do the same for the inorganic phosphorus levels. Note that the labs had each received blood samples from a different set of patients. How does that affect your interpretation of the boxplots?