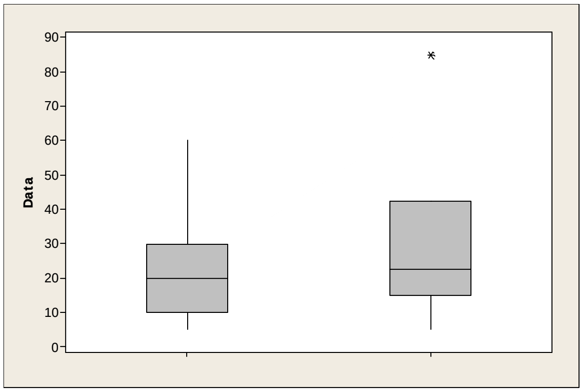
CHAPTER

2



Describing Distributions with Numbers

Numerical Descriptions

Numerical measures are often used to describe distributions. Select

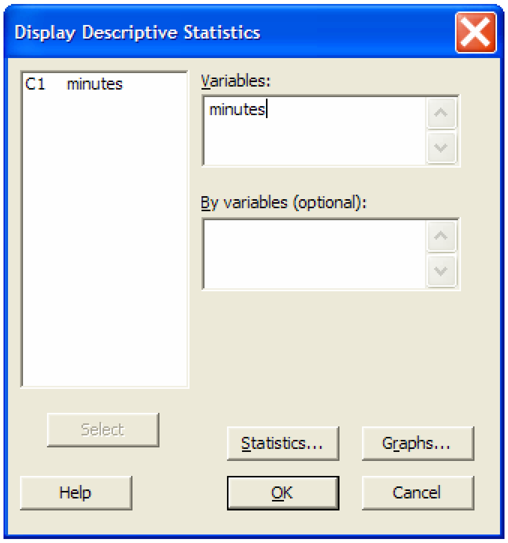
Stat ➤ Basic Statistics ➤ Display Descriptive Statistics

from the menu to obtain descriptive statistics. The command summarizes several different measures of both the center and spread of a distribution. The command prints the statistics N, N*, Mean, SE Mean, StDev, Minimum, Q1, Median, Q3, and Maximum for each column specified.

If we want descriptive information for the travel times in minutes for 15 workers in North Carolina, chosen at random by the Census Bureau, enter the following data into a Minitab worksheet or open EG02_01.MTW.

30 20 10 40 25 20 10 60 15 40 5 30 12 10 10

Descriptive information for the travel times is obtained by selecting **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics** from the menu. We select C1 or earnings for our variable in the dialog box as follows.



Descriptive Statistics: minutes

```
Variable   N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
minutes   15   0  22.47     3.93  15.23     5.00  10.00   20.00  30.00    60.00
```

N is the number of actual values in the column (missing values are not counted). N* is the number (if any) of missing values. Mean is the average of the values. To find the median, the data first must be ordered. If N is odd, the median is the value in the middle. If N is even, the median is the average of the two middle values. StDev is the standard deviation computed as

$$\text{StDev} = \sqrt{\frac{\sum (x_i - \overline{x})^2}{\text{N} - 1}}$$

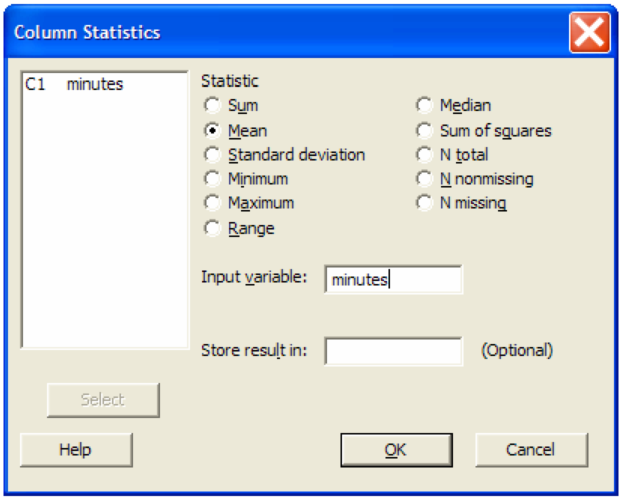
SE Mean is the standard error of the mean. It is calculated as $\text{StDev}/\sqrt{\text{N}}$. Q3 is the third quartile and Q1 is the first quartile. Minitab doesn't use exactly the same algorithm to calculate quartiles as BPS, so minor differences in results will sometimes occur.

The By variables box can be filled in to display descriptive statistics separately for each value of the specified variable. Column lengths must be equal to use the By variables box.

Column statistics can also be obtained individually by selecting

Calc ➤ Column Statistics

from the menu. The variable to be described and the descriptive measure or measures are selected on the dialog box.



Column Statistics

C1 minutes

Statistic

○ Sum	○ Median
● Mean	○ Sum of squares
○ Standard deviation	○ N total
○ Minimum	○ N nonmissing
○ Maximum	○ N missing
○ Range	

Input variable: minutes

Store result in: (Optional)

The statistics just described are also available for rows. These commands compute summaries across rows rather than down columns. The answers are always stored in a column. Rowwise statistics are obtained by selecting

Calc ➤ Row Statistics

from the menu. For the rowwise statistics as well as the column statistics, missing observations are omitted from the calculations.

Boxplots

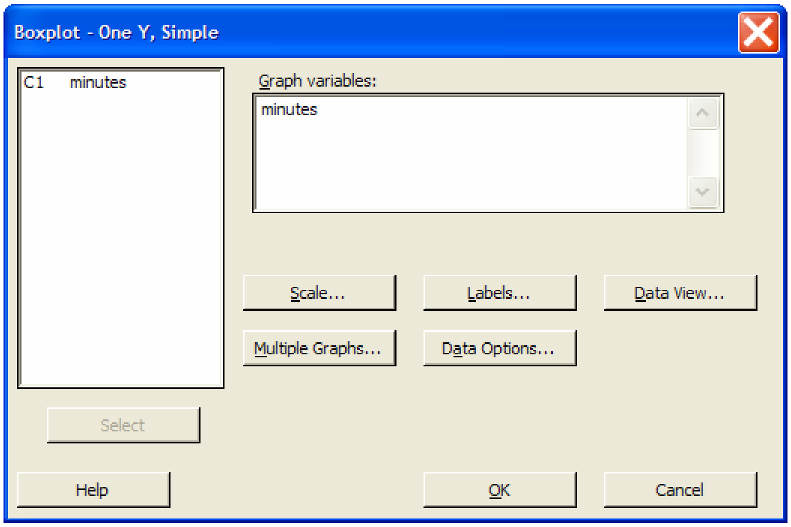
The five-number summary consisting of the median, quartiles, and minimum and maximum values provides a quick overall description of a distribution. Boxplots based on the five-number summary display the main features of a column of data. Boxplots can be obtained by selecting

Graph ➤ Boxplot

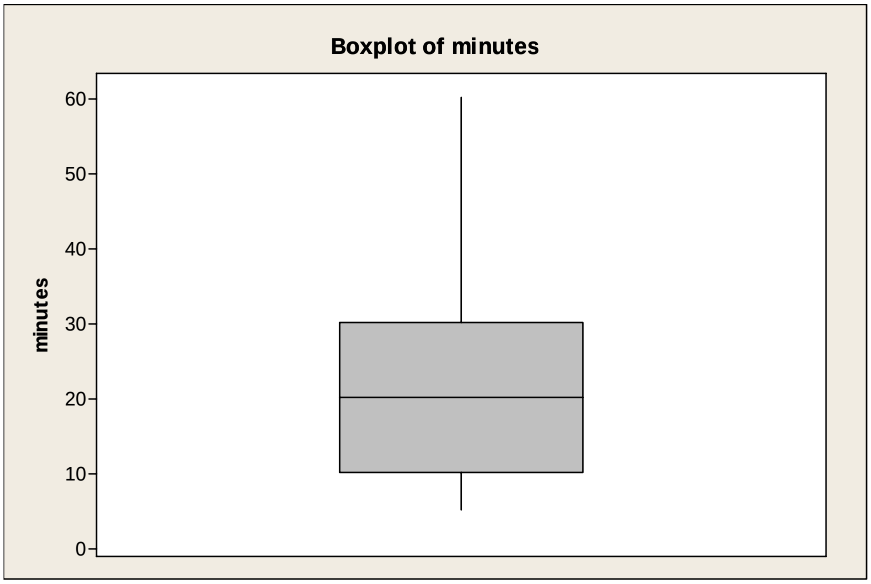
from the menu.

A boxplot graphically displays the main features of data from a single variable. A boxplot is illustrated on the following page on the hourly bank workers. The boxplot consists of a box, whiskers, and outliers. Minitab draws a line across the box at the median. By default, the bottom of the box is at the first quartile (Q1) and the top is at the third quartile (Q3). The whiskers are the lines that extend from the top and bottom of the box to the adjacent values, the lowest and highest observations still inside the region defined by the lower limit $Q1 - 1.5(Q3 - Q1)$ and the upper limit $Q3 + 1.5(Q3 - Q1)$. Outliers are points outside the lower and upper limits, plotted with asterisks (*).

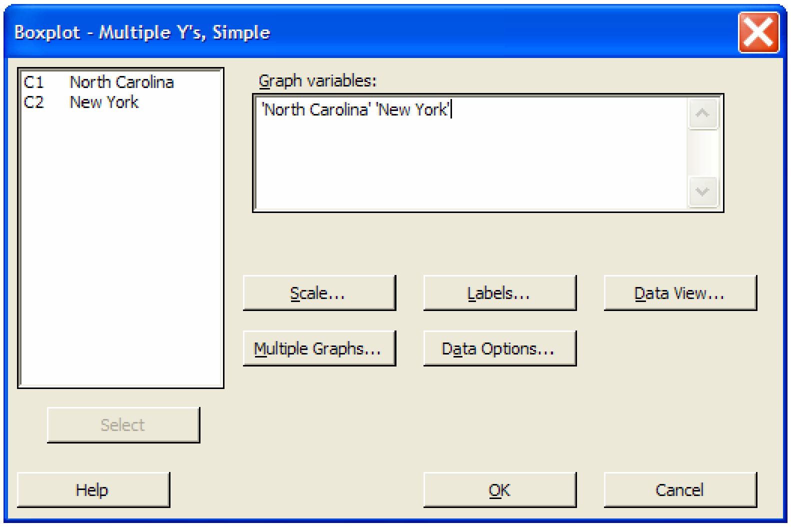
Selecting **Graph ➤ Boxplot** from the menu and filling in the dialog box will produce a boxplot(s). In the first dialog box, click on Simple for the boxplot of a single variable and then select a column for the box below Y. This is illustrated for the travel time data from Example 2.1 in BPS and EG02-01.MTW.



The boxplot is illustrated below.



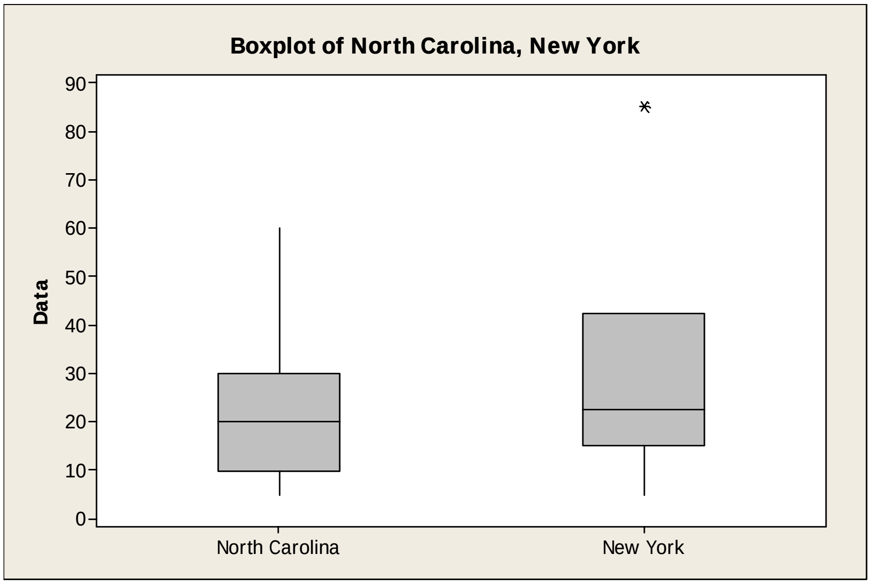
Example 2.5 in BPS illustrates travel times for both New York and North Carolina. To construct side-by-side boxplots comparing the travel times with for drivers in North Carolina to the travel times for drivers in New York, we have the data entered into a separate column for each state. Select **Graph ➤ Boxplot** from the menu and click on Multiple Y's Simple in the first dialog box. The next dialog box will then give space for the graph variables as follows. We select the columns with the North Carolina and New York data and click OK to obtain side-by-side boxplots.



C1 North Carolina
C2 New York

Graph variables: 'North Carolina' 'New York'

Side-by-side boxplots comparing the travel times for North Carolina with the travel times for New York appear below. By default, the upper whisker extends to the highest data value within the upper limit. $Q3 + 1.5(Q3 - Q1)$. Values beyond the whiskers are identified as outliers as shown for New York.



As an alternative, data for side-by-side boxplots can be arranged with the measurements (travel times) in one column and a categorical variable (location) in another. In this case, select **Graph ➤ Boxplots** from the menu and then One Y With Groups in the first dialog box to make side-by-side boxplots.

EXERCISES

2.1 Table 1.2 and TA01-02.MTW give the gas mileages for the 22 two-seater cars listed in the government's fuel economy guide.

(a) Select **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics** from the menu to find the mean $\bar{x}$ and the standard deviation *s*.

(b) The Honda Insight is an outlier that doesn't belong with the other cars. Find $\bar{x}$ and *s* for the observations that remain when you leave out the outlier. How does the outlier affect the values of $\bar{x}$ and *s*?

2.2 Select **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics** from the menu to find the median highway mileage for the 22 two−seater cars listed in Table 1.2 of BPS and TA01-02.MTW. What is the median of the 21 cars that remain if we remove the Honda Insight? Compare the effect of the Insight on mean mileage (Exercise 2.1) and on the median mileage. What general fact about the mean and median does this comparison illustrate?

2.4 The major league baseball single-season home run record is held by Barry Bonds of the San Francisco Giants, who hit 73 in 2001. Here and in EX02-04.MTW are Bonds's home run totals from 1986 (his first year) to 2002:

16 25 24 19 33 25 34 46 37
33 42 40 37 34 49 73 46

Select **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics** from the menu to find the mean $\bar{x}$ and the median. Bonds's record year is a high outlier. How do his career mean and median number of home runs change when we drop the record 73? What general fact about the mean and median does your result illustrate?

2.5 Example 1.8 in BPS and EG01-08.MTW give the breaking strengths of 20 pieces of Douglas fir.

(a) Select **Graph ➤ Stem-and-Leaf** to make a stemplot. The stemplot shows that the distribution is skewed to the left.

(b) Select **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics** from the menu to find the five-number summary of the distribution of breaking strengths. Does the five−number summary show the skew? Remember that only a graph gives a clear picture of the shape of a distribution.

2.6 Table 2.1 in BPS and TA02-01.MTW give the city and highway gas mileage for 36 midsize cars from the 2002 model year. There is one low outlier, the 12-cylinder Rolls-Royce. We wonder if midsize sedans get better mileage than sports cars.

(a) Select **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics** from the menu to find the five-number summaries for both city and highway mileage for the midsize cars in TA02-01.MTW and for the two-seater cars in TA01-02.MTW. (Leave out the Honda Insight.)

(b) Copy the data from TA01-02.MTW onto TA02-01.MTW. Select **Graph ➤ Boxplot** to make four side-by-side boxplots to display the summaries. Write a brief description of city versus highway and two-seaters versus midsize cars.

2.7 How old are presidents at their inauguration? Was Bill Clinton, at age 46, unusually young? Table 2.2 in BPS and TA02-02.MTW give the data, the ages of all U.S. presidents when they took offce.

(a) Select **Graph ➤ Stem-and-Leaf** to make a stemplot of the distribution of ages. From the shape of the distribution, do you expect the median to be much less than the mean, about the same as the mean, or much greater than the mean?

(b) Select **Stat ➤ Basics Statistics ➤ Desplay Descriptive Statistics** to find the mean and the five-number summary. Verify your expectation about the median.

(c) What is the range of the middle half of the ages of new presidents? Was Bill Clinton in the youngest 25%?

2.9 The mean $\bar{x}$ and standard deviation s measure center and spread but are not a complete description of a distribution. Data sets with different shapes can have the same mean and standard deviation. To demonstrate this fact, select **➤ Basics Statistics ➤ Desplay Descriptive Statistics** to find $\bar{x}$ and s the two small data sets found below and in EX02-09.MTW. Then select **Graph ➤ Stem-and-Leaf** to make a stemplot of each and comment on the shape of each distribution.

Data A	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data B	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

2.12 How much do users pay for Internet access? Here and in EX02-12.MTW are the monthly fees (in dollars) paid by a random sample of 50 users of commercial Internet service providers in August 2000:

20	40	22	22	21	21	20	10	20	20
20	13	18	50	20	18	15	8	22	25
22	10	20	22	22	21	15	23	30	12
9	20	40	22	29	19	15	20	20	20
20	15	19	21	14	22	21	35	20	22

(a) Select **Graph ➤ Stem-and-Leaf** to make a stemplot of these data. Briefly describe the pattern you see. About how much do you think America Online and its larger competitors were charging in August 2000? Are there any outliers?

(b) To report a quick summary of how much people pay for Internet service, do you prefer $\bar{x}$ and s or the five-number summary? Why? Select **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics** to calculate your preferred summary.

2.13 Table 1.6 in BPS and TA01-06.MTW give the number of medical doctors per 100,000 people in each state. Exercise 1.29 asked you to plot the data. The distribution is right-skewed with several high outliers.

(a) Do you expect the mean to be greater than the median, about equal to the median, or less than the median? Why? Select **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics** to calculate $\bar{x}$ and M and verify your expectation.

(b) The District of Columbia is a high outlier at 758 M.D.'s per 100,000 residents. If you remove D.C. because it is a city rather than a state, do you expect $\bar{x}$ or M to change more? Why? Select **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics** to calculate both measures for the 50 states (omitting D.C.) and verify your expectation.

2.14 Exercises 1.25 and 1.26 give the numbers of home runs hit each season by Babe Ruth and Mark McGwire. Enter the data for each player into a separate column on the same Minitab worksheet. Select **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics** to find the five-number summaries for each hitter. Select **Graph ➤ Boxplot** and make side-by-side box plots to compare these two home run hitters. What do your plots show?

2.15 Here are the scores of 18 first-year college women on the Survey of Study Habits and Attitudes (SSHA):

154	109	137	115	152	140	154	178	101
103	126	126	137	165	165	129	200	148

(a) Enter the data into a Mintab worksheet and select **Calc ➤ Column Statistics** to obtain the mean.

(b) A stemplot (Exercise 1.8) suggests that the score 200 is an outlier. Select **Calc ➤ Column Statistics** to find the mean for the 17 observations that remain when you drop the outlier. How does the outlier change the mean?

2.17 In Exercise 2.15 you found the mean of the SSHA scores of 18 first-year college women. Now select **Calc ➤ Column Statistics** to find the median of these scores. Is the median smaller or larger than the mean? Explain why this is so.

2.19 Does breast-feeding weaken bones? Breast-feeding mothers secrete calcium into their milk. Some of the calcium may come from their bones, so mothers may lose bone mineral. Researchers compared 47 breast-feeding 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. Here and in EX02-19.MTW are the data:

Breast-feeding women						Other women					
−4.7	−2.5	−4.9	−2.7	−0.8	−5.3	2.4	0.0	0.9	−0.2	1.0	1.7
−8.3	−2.1	−6.8	−4.3	2.2	−7.8	2.9	−0.6	1.1	−0.1	−0.4	0.3
−3.1	−1.0	−6.5	−1.8	−5.2	−5.7	1.2	−1.6	−0.1	−1.5	0.7	−0.4
−7.0	−2.2	−6.5	−1.0	−3.0	−3.6	2.2	−0.4	−2.2	−0.1		
−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							

Select **Graph ➤ Stem-and-Leaf** from the menu and use the change in mineral content for the Graph variable. Enter the column containing the group for the By variable. Minitab will produce a separate plot for each group. Also, select **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics** to compare the two distributions. Do the data show distinctly greater bone mineral loss among the breast-feeding women?

2.20 In 1798 the English scientist Henry Cavendish measured the density of the earth with great care. It is common practice to repeat careful measurements several times and use the mean as the final result. Cavendish repeated his work 29 times. Here and in EX02-20.MTW are his results (the data give the density of the earth as a multiple of the density of water):

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

Present these measurements with a graph of your choice. Scientists usually give the mean and standard deviation to summarize a set of measurements. Does the shape of this distribution suggest that $\bar{x}$ and s are adequate summaries? Calculate $\bar{x}$ and s.

2.21 You are planning a sample survey of households in California. You decide to select households separately within each county and to choose more households from the more populous counties. To aid in the planning, Table 2.3 in BPS and TA02-03.MTW give the populations of California counties from the 2000 census. Examine the distribution of county populations both graphically and numerically, using whatever tools are most suitable. Write a brief description of the main features of this distribution. Sample surveys often select households from all of the most populous counties but from only some of the less populous. How would you divide California counties into three groups according to population, with the intent of including all of the first group, half of the second, and a smaller fraction of the third in your survey?

2.22 The University of Miami Hurricanes have been among the more successful teams in college football. Table 2.4 gives the weights in pounds and positions of the players on the 2002 team. The positions are quarterback (QB), running back (RB), offensive line (OL), wide receiver (WR), tight end (TE), kicker/punter (KP), defensive back (DB), linebacker (LB), and defensive line (DL).

(a) Select **Graph** ➤ **Boxplots** to make side-by-side boxplots of the weights for running backs, wide receivers, offensive linemen, defensive linemen, linebackers, and defensive backs.

(b) Briefly compare the weight distributions. Which position has the heaviest players overall? Which has the lightest?

(c) Are any individual players outliers within their position?

2.23 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. Survival times, whether of machines under stress or cancer patients after treatment, usually have distributions that are skewed to the right.

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) Graph the distribution and describe its main features. Does it show the expected right skew?

(b) Which numerical summary would you choose for these data? Select **Stat** ➤ **Basic Statistics** ➤ **Display Descriptive Statistics** to calculate your chosen summary. How does it reflect the skewness of the distribution?

2.25 In 2002, the Chicago Cubs failed once again to reach the National League playoffs. Table 2.5 in BPS and TA02-05.MTW give the salaries of the Cubs' players as of opening day of the season. Describe the distribution of salaries both with a graph and with a numerical summary. Then write a brief description of the important features of the distribution.