Chapter 5
Exploring Data: Distributions

Chapter Objectives

Check off these skills when you feel that you have mastered them.

☐ Construct a histogram for a small data set.

☐ List and describe two types of distributions for a histogram.

☐ Identify from a histogram possible outliers of a data set.

☐ Construct a stemplot for a small data set.

☐ Calculate the mean of a set of data.

☐ Sort a set of data from smallest to largest and then determine its median.

☐ Determine the upper and lower quartiles for a data set.

☐ Calculate the five-number summary for a data set.

☐ Construct the diagram of a boxplot from the data set's five-number summary.

☐ Calculate the standard deviation of a small data set.

☐ Describe a normal curve.

☐ Given the mean and standard deviation of a normally distributed data set, compute the first and third quartiles.

☐ Explain the 68–95–99.7 rule.

☐ Sketch the graph of a normal curve given its mean and standard deviation.

☐ Given the mean and standard deviation of a normally distributed data set, compute the intervals in which the data set fall into a given percentage by applying the 68–95–99.7 rule.

Guided Reading

Introduction

Data, or numerical facts, are essential for making decisions in almost every area of our lives. But to use them for our purposes, huge collection of data must be organized and distilled into a few comprehensible summary numbers and visual images. This will clarify the results of our study and allow us to draw reasonable conclusions. The analysis and display of data are thus the groundwork for statistical inference.

☞ **Key idea**

In a data set there are **individuals**. These individuals may be people, cars, cities, or anything to be examined.

☞ **Key idea**

The characteristic of an individual is a **variable**. For different individuals, a variable can take on different values.

✍ **Example A**

Identify the individuals and the variables in the following data set from a class roster.

Name	Age	Sex
Dan	16	Male
Edwin	17	Male
Adam	16	Male
Nadia	15	Female

Solution

The individuals are the names of the people on the class roster. The variables are their ages and sex.

☞ **Key idea**

In this chapter, you will be doing **exploratory data analysis**. This combines numerical summaries with graphical display to see patterns in a set of data. The organizing principles of data analysis are as follows.

1) Examine individual variables, and then look for relationships among variables.

2) Draw a graph or graphs and add to it numerical summaries.

Section 5.1   Displaying Distributions: Histograms

☞ **Key idea**

The **distribution** of a variable tells us what values the variable takes and how often it takes these values.
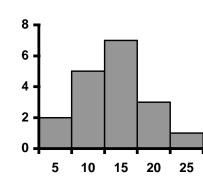
☞ **Key idea**

The most common graph of a distribution with one numerical variable is called a **histogram**.

⤳ **Example B**

Construct a histogram given the following data. How many pieces of data are there?

Value	Count
5	2
10	5
15	7
20	3
25	1

Solution

There are $2+5+7+3+1=18$ piec........

⚷ **Key idea**

When constructing a histogram, each piece of data must fall into one **class**. Each class must be of equal width. For any given data set, there is more than one way to define the classes. Either you are instructed as to how to define the classes, or you must determine class based on some criteria.

⤳ **Example C**

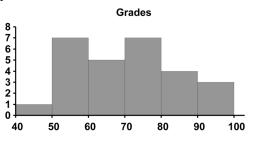Given the following exam scores, construct a histogram with classes of length 10 points.

40	50	50	53	55	55	55	58	60
60	63	65	68	70	70	73	75	75
78	78	83	85	85	88	90	95	96

Solution

It is helpful to first put the data into classes and count the individual pieces of data in each class. Since the smallest piece of data is 40, it makes sense to make the first class 40 to 49, inclusive.
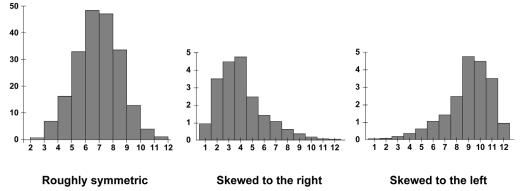
Notice that the sum of the values in the count column should be 27 (total number of pieces of data). Also notice that some of the details of the scores are lost when raw data are placed in classes.

Class	Count
40 – 49	1
50 – 59	7
60 – 69	5
70 – 79	7
80 – 89	4
90 – 99	3

Section 5.2   Interpreting Histograms

⊶ Key idea

An important feature of a histogram is its overall **shape**. Although there are many shapes and overall patterns, a distribution may be **symmetric**, or it may be **skewed to the right** or **skewed to the left**.

| Roughly symmetric | Skewed to the right | Skewed to the left |

If a distribution is skewed to the right, then the larger values extend out much further to the right. If a distribution is skewed to the left, then the smaller values extend out much further to the left. The easiest way to keep the two terms from being confused is to think of the direction of the "tail". If the tail points left, it is skewed to the left. If the tail points right, it is skewed to the right.

⊶ Key idea

Another way to describe a distribution is by its **center**. For now, we can think of the center of a distribution as the midpoint.

⊶ Key idea

Another way to describe a distribution is by its **spread**. The spread of a distribution is stating its smallest and largest values.

⊶ Key idea

In a distribution, we may also observe **outliers**; that is, a piece or pieces of data that fall outside the overall pattern. Often times determining an outlier is a matter of judgment. There are no hard and fast rules for determining outliers.
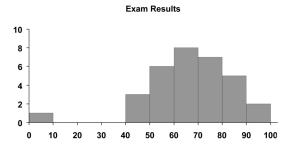
⌇ Example D

Given the following data regarding exam scores, construct a histogram. Describe its overall shape and identify any outliers.

Class	Count	Class	Count
0 – 9	1	50 – 59	6
10 – 19	0	60 – 69	8
20 – 29	0	70 – 79	7
30 – 39	0	80 – 89	5
40 – 49	3	90 – 99	2

Solution on next page
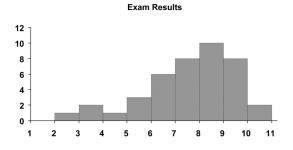
Solution

Exam Results

The shape is roughly symmetric. The score in the class 0 – 9, inclusive, is clearly an outlier. With a 0 on an exam, the most likely explanation is that the student missed the exam. It is also possible that the student was completely unprepared and performed poorly to obtain a very low score.

ᏏᎵ Example E

Given the following data regarding exam scores, construct a histogram. Describe its overall shape and identify any outliers.

Class	Count	Class	Count
0 – 9	0	50 – 59	6
10 – 19	1	60 – 69	8
20 – 29	2	70 – 79	10
30 – 39	1	80 – 89	8
40 – 49	3	90 – 99	2

Solution

Exam Results

The shape is skewed to the left. There doesn't appear to be any outliers.

✎ Question 1

Given the following exam scores, describe the overall shape of the distribution and identify any outliers. In your solution, construct a histogram with class length of 5 points.

21	22	59	60	61	62	63	64	65
65	66	67	68	68	69	69	70	72
73	74	74	75	76	77	78	80	81
82	85	86	89	91	92	95		

Answer

The distribution appears to be skewed to the right. The scores of 21 and 22 appear to be outliers.

Section 5.3   Displaying Distributions: Stemplots

☞ Key idea

A **stemplot** is a good way to represent data for small data sets. Stemplots are quicker to create than histograms and give more detailed information. Each value in the data set is represented as a stem and a leaf. The stem consists of all but the rightmost digit, and the leaf is the rightmost digit. A stemplot resembles a histogram turned sideways.

✆ Example F

Given the following exam scores, construct a stemplot.

40	50	50	53	55	55	55	58	60
60	63	65	68	70	70	73	75	75
78	78	83	85	85	88	90	95	96

Solution

In the stemplot, the tens digit will be the stem and the ones digit will be the leaf.

```
4 | 0
5 | 0035558
6 | 00358
7 | 0035588
8 | 3558
9 | 056
```

✎ Question 2

The following are the percentages of salt concentrate taken from lab mixture samples. Describe the shape of the distribution and any possible outliers. This should be done by first rounding each piece of data to the nearest percent and then constructing a stemplot.

Sample	1	2	3	4	5	6	7
Percent	39.8	65.7	64.7	20.1	40.8	53.4	70.8
Sample	8	9	10	11	12	13	14
Percent	50.7	68.7	74.3	82.6	58.5	68.0	72.2

Answer

The distribution appears to be roughly symmetric with 20 as a possible outlier.

Section 5.4   Describing Center: Mean and Median

☞ Key idea

The **mean** of a data set is obtained by adding the values of the observations in the data set and dividing by the number of data. If the observations are listed as values of a variable x (namely x_1, x_2,..., x_n), then the mean is written as $\bar{x}$. The formula for the mean is $\bar{x} = \dfrac{x_1 + x_2 + ... + x_n}{n}$, where n represents the number of pieces of data.

✍ Example G

Calculate the mean of each data set.
a)   123, 111, 105, 115, 112, 113, 117, 119, 114, 118, 111, 150, 147, 129, 138
b)   17, 15, 13, 2, 14, 15, 10, 1, 16, 16, 17, 22

Solution

a) $\bar{x} = \dfrac{123+111+105+115+112+113+117+119+114+118+111+150+147+129+138}{15}$

$= \dfrac{1822}{15} \approx 121.5$

b) $\bar{x} = \dfrac{17+15+13+2+14+15+10+1+16+16+17+22}{12} = \dfrac{158}{12} = 13.2$

✎ Question 3

Given the following stemplot, determine the mean. Round to the nearest tenth, if necessary.

1	259
2	3478
3	0334679
4	01259
5	46
6	1
7	3

Answer

$\bar{x} = 37$

⌇ Key idea

The **median**, **M**, of a distribution is a number in the middle of the data, so that half of the data are above the median, and the other half are below it. When determining the median, the data should be placed in order, typically smallest to largest. When there are n pieces of data, then the piece of data $\frac{n+1}{2}$ observations up from the bottom of the list is the median. This is fairly straightforward when n is odd. When there are n pieces of data and n is even, then you must find the average (add together and divide by two) of the two center pieces of data. The smaller of these two pieces of data is located $\frac{n}{2}$ observations up from the bottom of the list. The second, larger, of the two pieces of data is the next one in order or, $\frac{n}{2}+1$ observations up from the bottom of the list.

✍ Example H

Determine the median of each data set below.
a)   123, 111, 105, 115, 112, 113, 117, 119, 114, 118, 111, 150, 147, 129, 138
b)   17, 15, 13, 2, 14, 15, 10, 1, 16, 16, 17, 22

Solution

For each of the data sets, the first step is to place the data in order from smallest to largest.
a)   105, 111, 111, 112, 113, 114, 115, **117**, 118, 119, 123, 129, 138, 147, 150

Since there are 15 pieces of data, the $\frac{15+1}{2} = \frac{16}{2} = 8^{\text{th}}$ piece of data, namely 117, is the median.

b)   1, 2, 10, 13, 14, **15**, **15**, 16, 16, 17, 17, 22

Since there are 12 pieces of data, the mean of the $\frac{12}{2} = 6^{\text{th}}$ and 7th pieces of data will be the median. Thus, the median is $\frac{15+15}{2} = \frac{30}{2} = 15$. Notice, if you use the general formula $\frac{n+1}{2}$, you would be looking for a value $\frac{12+1}{2} = \frac{13}{2} = 6.5$ "observations" from the bottom. This would imply halfway between the actual 6th observation and the 7th observation.

✐ Question 4

Given the following stemplot, determine the median.

1	029
2	3478
3	03345679
4	012359
5	16
6	012

Answer

$M = 36.5$

Section 5.5   Describing Spread: The Quartiles

⌐⊶ Key idea

The **quartiles** Q_1 (the point below which 25% of the observations lie) and Q_3 (the point below which 75% of the observations lie) give a better indication of the true spread of the data. More specifically, Q_1 is the median of the data to the left of M (the median of the data set). Q_3 is the median of the data to the right of M.

↷ Example I

Determine the quartiles Q_1 and Q_3 of each data set below.
a)   123, 111, 105, 115, 112, 113, 117, 119, 114, 118, 111, 150, 147, 129, 138

b)   17, 15, 13, 2, 14, 15, 10, 1, 16, 16, 17, 22

Solution

For each of the data sets, the first step is to place the data in order from smallest to largest.
a)   105, 111, 111, **112**, 113, 114, 115, 117, 118, 119, 123, **129**, 138, 147, 150

From Example H we know that the median is the 8th piece of data. Thus, there are 7 pieces of data below M. We therefore can determine Q_1 to be the $\frac{7+1}{2} = \frac{8}{2} = 4^{th}$ piece of data. Thus, $Q_1 = 112$. Now since there are 7 pieces of data above M, Q_3 will be the 4th piece of data to the right of M. Thus, $Q_3 = 129$.

b)   1, 2, **10**, **13**, 14, 15, ↓ 15, 16, **16**, **17**, 17, 22

From Example H we know that the median is between the 6th and 7th pieces of data. Thus, there are 6 pieces of data below M. Since $\frac{6+1}{2} = \frac{7}{2} = 3.5$, Q_1 will be the mean of 3rd and 4th pieces of data, namely $\frac{10+13}{2} = \frac{23}{2} = 11.5$. Now since there are 6 pieces of data above M, Q_3 will be the mean of the 3rd and 4th pieces of data to the right of M. Thus, $Q_3 = \frac{16+17}{2} = \frac{33}{2} = 16.5$.

✐ Question 5

Determine the quartiles Q_1 and Q_3 of each data set below.
a)   21, 16, 20, 6, 8, 9, 12, 15, 3, 15, 7, 8, 19

b)   14, 12, 11, 12, 24, 8, 6, 4, 8, 10

Answer

a) $Q_1 = 7.5$ and $Q_3 = 17.5$

b) $Q_1 = 8$ and $Q_3 = 12$

Section 5.6   The Five-Number Summary and Boxplots

☞ **Key idea**

The **five-number summary** consists of the median (M), quartiles (Q_1 and Q_3), and extremes (high and low).
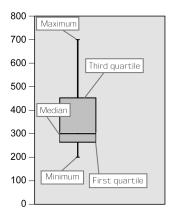
☞ **Key idea**

A **boxplot** is a graphical (visual) representation of the five-number summary. A central box spans quartiles Q_1 and Q_3. A line in the middle of the central box marks the median, M. Two lines extend from the box to represent the extreme values.

✑ **Example J**

Given the following five-number summary, draw the boxplot.

$$200, 250, 300, 450, 700$$

Solution
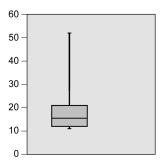

✎ **Question 6**

Given the following data, find the five-number summary and draw the boxplot.

$$12, 11, 52, 12, 15, 21, 17, 35, 16, 12$$

Answer

The five-number summary is 11, 12, 15.5, 21, 52.

The boxplot is as follows.

Section 5.7   Describing Spread: The Standard Deviation

✋ Key idea

The **variance**, s^2, of a set of observations is an average of the squared differences between the individual observations and their mean value. In symbols, the variance of n observations

$$\left(x_1, x_2, ..., x_n\right) \text{ is } s^2 = \frac{\left(x_1 - \bar{x}\right)^2 + \left(x_2 - \bar{x}\right)^2 + ... + \left(x_n - \bar{x}\right)^2}{n-1}. \text{ Notice we divide by } n-1.$$

✋ Key idea

The **standard deviation**, s, of a set of observations is the square root of the variance and measures the spread of the data around the mean in the same units of measurement as the original data set. You should be instructed as to the method (spreadsheet, calculator with statistical capabilities, or by hand) required for calculating the variance and in turn the standard deviation.

✍ Example K

Given the following data set, find the variance and standard deviation.

$$8.6, 7.2, 9.2, 5.6, 5.5, 4.4$$

Solution

Placing the data in order (not required, but helpful) we have the following hand calculations. Notice that $\bar{x} = \frac{40.5}{6} = 6.75$.

Observations x_i	Deviations $x_i - \bar{x}$	Squared deviations $\left(x_i - \bar{x}\right)^2$
4.4	$4.4 - 6.75 = -2.35$	$(-2.35)^2 = 5.5225$
5.5	$5.5 - 6.75 = -1.25$	$(-1.25)^2 = 1.5625$
5.6	$5.6 - 6.75 = -1.15$	$(-1.15)^2 = 1.3225$
7.2	$7.2 - 6.75 = 0.45$	$(0.45)^2 = 0.2025$
8.6	$8.6 - 6.75 = 1.85$	$(1.85)^2 = 3.4225$
9.2	$9.2 - 6.75 = 2.45$	$(2.45)^2 = 6.0025$
sum = 40.5	**sum = 0.00**	**sum = 18.035**

Thus, $s^2 = \dfrac{18.035}{6-1} = \dfrac{18.035}{5} = 3.607$ and $s = \sqrt{3.607} \approx 1.90$.

✏ Question 7

Given the following data set, find the variance and standard deviation.

$$3.41, 2.78, 5.26, 6.49, 7.61, 7.92, 8.21, 5.51$$

Answer

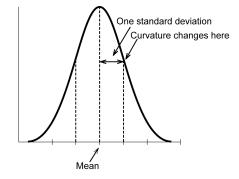$s^2 \approx 4.169$ and $s = \sqrt{4.169} \approx 2.04$.

Section 5.8   Normal Distributions

⌐ Key idea

Sampling distributions, and many other types of probability distributions, approximate a bell curve in shape and symmetry. This kind of shape is called a normal curve, and can represent a **normal distribution**, in which the area of a section of the curve over an interval coincides with the proportion of all values in that interval. The area under any normal curve is 1.

⌐ Key idea

A normal curve is uniquely determined by its mean and standard deviation. The **mean** of a normal distribution is the center of the curve. The symbol μ will be used for the mean. The **standard deviation** of a normal distribution is the distance from the mean to the point on the curve where the curvature changes. The symbol σ will be use for the standard deviation.

⌐ Key idea

The first quartile is located 0.67 standard deviation below the mean, and the third quartile is located 0.67 standard deviation above the mean. In other words, we have the following formulas.

$$Q_1 = \mu - 0.67\sigma \text{ and } Q_3 = \mu + 0.67\sigma$$

⌐ Example L

The scores on a marketing exam were normally distributed with a mean of 73 and a standard deviation of 12.

a)   Find the third quartile (Q_3) for the test scores.

b)   Find a range containing exactly half of the students' scores.

Solution
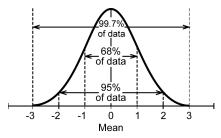
a)   Since $Q_3 = \mu + 0.67\sigma = 73 + 0.67(12) = 73 + 8.04 = 81.04$, we would say the third quartile is 81.

b)   Since 25% of the data lie below the first quartile and 25% of the data fall above the third quartile, 50% of the data would fall between the first and third quartiles. Thus, we must find the first quartile. Since $Q_1 = \mu - 0.67\sigma = 73 - 0.67(12) = 73 - 8.04 = 64.96$, we would say an interval would be $[65, 81]$.

Section 5.9   The 68 – 95 – 99.7 Rule

⌐ Key idea

The **68–95–99.7 rule** applies to a normal distribution. It is useful in determining the proportion of a population with values falling in certain ranges. For a normal curve, the following rules apply:
- The proportion of the population within one standard deviation of the mean is 68%.
- The proportion of the population within two standard deviations of the mean is 95%.
- The proportion of the population within three standard deviations of the mean is 99.7%.

𝒢⌒ Example M

The amount of coffee a certain dispenser fills 16 oz coffee cups with is normally distributed with a mean of 14.5 oz and a standard deviation of 0.4 oz.

a) Almost all (99.7%) cups dispensed fall within what range of ounces?

b) What percent of cups dispense less than 13.7 oz?

Solution
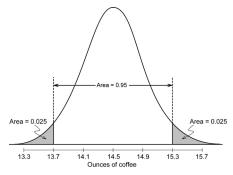
a)   Since 99.7% of all cups fall within 3 standard deviations of the mean, we find the following.

$$\mu \pm 3\sigma = 14.5 \pm 3(0.4) = 14.5 \pm 1.2$$

Thus, the range of ounces is 13.3 to 15.7.

b)   Make a sketch: 13.7 oz is two σ below μ; 95% are within 2σ of μ.

5% lie farther than 2σ. Thus, half of these, or 2.5%, lie below 13.7.

✎ Question 8

Look again at the marketing exam in which scores were normally distributed with a mean of 73 and a standard deviation of 12.

a)   Find a range containing 34% of the students' scores.

b)   What percentage of the exam scores were between 61 and 97?

Answer

a)   Either of the intervals [61, 73] or [73, 85]

b)   81.5%

Homework Help

Exercise 1
Carefully read the Introduction before responding to this exercise.

Exercises 2 – 3
Carefully read Section 5.2 before responding to these exercises. Pay special attention to the description of skewed distributions.

Exercise 4
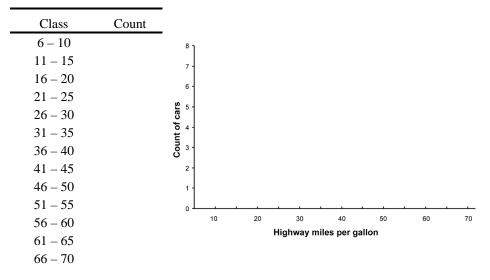Carefully read Sections 5.1 – 5.3 before responding to this exercise. First construct your classes and count individuals as described in Example 2 of your text. Include the outlier in your histogram. The following may be helpful in constructing your histogram. One possibility is to make the first class $6 \leq$ gas mileage < 11 or $11 \leq$ gas mileage < 16.

Class	Count
6 – 10	
11 – 15	
16 – 20	
21 – 25	
26 – 30	
31 – 35	
36 – 40	
41 – 45	
46 – 50	
51 – 55	
56 – 60	
61 – 65	
66 – 70	

Exercise 5
Carefully read Sections 5.1 – 5.2 before responding to this exercise. First construct your classes and count individuals as described in Example 2 of your text. Include the outliers in your histogram. The following may be helpful in constructing your histogram. One possibility is to make the first class $0.0 \leq$ emmissions < 2.0.

Class	Count
0.0 – 1.9	
2.0 – 3.9	
4.0 – 5.9	
6.0 – 7.9	
8.0 – 9.9	
10.0 – 11.9	
12.0 – 13.9	
14.0 – 15.9	
16.0 – 17.9	
18.0 – 19.9	

Pay special attention to the description of skewed distributions and outliers.

Exercise 6
Carefully read Section 5.2 before responding to this exercise. Pay special attention to the description of symmetric and skewed distributions. Think about how gender and right/left-handedness are distributed in real life.

Exercises 7 – 10
Carefully read Section 5.3 before responding to these exercises. Carefully read the description of how to describe each piece of data in Exercise 8. You may choose to use the following stems in the exercises.

Exercise 8	Exercise 9	Exercise 10
0	10	48
1	11	49
2	12	50
3	13	51
	14	52
	15	53
	16	54
	17	55
	18	56
	19	57
	20	58

Exercise 11
Carefully read Section 5.4 before responding to this exercise. Make sure to show all steps in your calculations, unless otherwise instructed.

Exercise 12
(a) Make the stemplot, with the outlier.

1
2
3
4
5
6
7

(b) Calculate the mean. Use the stemplot to put the data in order from smallest to largest in order to find the median. Since there is an even number of pieces of data, you will need to examine two pieces of data to determine the median. Remove the outlier and recalculate the mean and determine the median of the 17 pieces of data. Compare the results with and without the outlier.

Exercise 13 – 14
Carefully read Section 5.2 before responding to these exercises. The following drawings may be helpful to show the relative locations of the median and the mean.

Skewed to the right   Skewed to the left

Exercises 15 – 16
Examples will vary.

Exercises 17 – 20
Carefully read Section 5.5 – 5.6 before responding to these exercises. Make sure to first put data in order from smallest to largest. Double check that you have accounted for all pieces of data. Pay special attention when you are dealing with an even number of pieces of data in determining the median. When determining quartiles, remember if there is an even number of pieces to the left of the mean, there will also be an even number of pieces to the right of the mean.

Exercise 21
Carefully read Section 5.5 – 5.6 before responding to this exercise. It would be helpful to create a stem plot to organize your data from smallest to largest.

```
 0 |
 1 |
 2 |
 3 |
 4 |
 5 |
 6 |
 7 |
 8 |
 9 |
10 |
11 |
12 |
13 |
14 |
15 |
16 |
17 |
18 |
19 |
```

Exercise 22
Carefully read Sections 5.5 – 5.6 before responding to this exercise. The data are already organized from smallest to largest.

Exercise 23
Look carefully at the referenced figure and compare as many features as possible.

Exercise 24
Your values of the median and quartiles may differ slightly from another student. Try rounding to the nearest thousand.

Exercise 25
Carefully read Sections 5.1 – 5.6 before responding to this exercise. Make sure to first put data in order from smallest to largest and round to the nearest whole number. Double check that you have accounted for all pieces of data. In part a, you can either create a histogram or a stemplot. Pay special attention when you are dealing with an even number of pieces of data in determining the median. When determining quartiles, remember if there is an even number of pieces to the left of the mean, there will also be an even number of pieces to the right of the mean.

Exercise 26
Approximate the bar heights. You will need to determine in which bar the 25%, 50%, and 75% marks occur for Q_1, M, and Q_3.

Exercises 27 – 28
Both of these exercises rely on the description of *interquartile range* given in Exercise 27.

Exercise 29

(a) Placing the data in order is helpful, but not required. If you are performing the calculations by hand, the following table may be helpful.

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $\left(x_i - \bar{x}\right)^2$
sum =		sum =	0.00	sum =	

(b) If your data are in order above, the median can easily be determined. If you have already worked Exercise 10, then the mean has already been calculated for comparison in this exercise.

Exercise 30
Placing the data in order is helpful, but not required. When performing the calculations by hand, the following table may be helpful.

	Observations x_i		Deviations $x_i - \overline{x}$		Squared deviations $\left(x_i - \overline{x}\right)^2$
sum =		sum =	0.00	sum =	

Exercise 31
Since the square root of the variance is the standard deviation, the square of the standard deviation is the variance.

Exercise 32
Think about what each of the measures represents before answering each part of this exercise.

Exercise 33
Round the data to the nearest tenth to create the stemplots.

Data A

3 |
4 |
5 |
6 |
7 |
8 |
9 |

Data B

5 |
6 |
7 |
8 |
9 |
10 |
11 |
12 |

Exercise 34
Look carefully at each figure before answering each part of this exercise.

Exercise 35
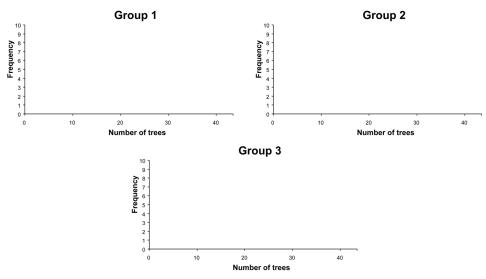Answers will vary on this exercise depending on what is used for calculations.

Exercise 36
The following may be helpful in creating stemplots.

Group 1

0 |
1 |
2 |
3 |

Group 2

0 |
1 |
2 |
3 |

Group 3

0 |
1 |
2 |
3 |

Continued on next page

Exercise 36 continued
The following may be helpful in creating histograms using class widths of 10 trees per 0.1 hectare in area.

Group 1

Group 2

Group 3

Placing the data in order is helpful, but not required. When performing the calculations by hand, the following table may be helpful.

Group 1:

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $\left(x_i - \bar{x}\right)^2$
sum =		sum =		sum =	

Continued on next page

Exercise 36 continued

In Group 2 and Group 3, you may choose to use 6 decimal place accuracy for $\bar{x}$ in order to calculate $x_i - \bar{x}$. You may also choose to round $\left(x_i - \bar{x}\right)^2$ to five decimal places. Round s^2 to four decimal places and s to three.

Group 2:

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $\left(x_i - \bar{x}\right)^2$
sum =		sum =		sum =	

Group 3:

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $\left(x_i - \bar{x}\right)^2$
sum =		sum =		sum =	

Exercise 37
(a) Since the standard deviation cannot be negative, think about what the smallest value it can be and its implication.
(b) Think about a way to make the data set as spread out as possible. Try with two pieces of data and see if that helps to determine four pieces of data. You may choose to try different data sets to to convince yourself of your answer.
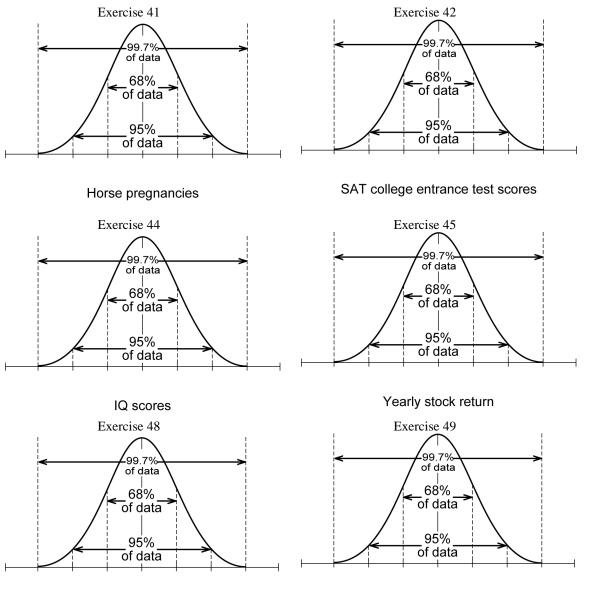
Exercise 38
Carefully read Section 5.8 before responding to this exercise. Sketch a normal curve, mark the axis with the mean as the center of the curve and one standard deviation to the right and left will mark the change-of-curvature points. These three points set the proper scale.

Exercise 39
Refer to exercises 13 – 14.

Exercise 40
Think of the letter M.

Exercises 41 – 42 and 44 – 45 and 48 – 49
Carefully read Section 5.9 before responding to these exercises. Make a sketch for each exercise by drawing a normal curve, placing the mean and 3 standard deviations to the right and left. The following may be helpful. Make sure to use symmetry when determining percentages above or below a value.

Exercise 41

Horse pregnancies

Exercise 42

SAT college entrance test scores

Exercise 44

IQ scores

Exercise 45

Yearly stock return

Exercise 48

Concentration of a painkiller

Exercise 49

Portion of people who stay home
for fear of crime

Exercises 43 and 46
Carefully read Section 5.8 before responding to these exercises regarding quartiles of a normal distribution.

Exercise 47
Carefully read Sections 5.8 – 5.9 before responding to this exercise.

Exercise 50
Apply the formula given in part a for parts a and b and compare the results in part c.

Exercise 51
The following may be helpful in creating the stemplots.

Lengths of red flowers	Lengths of yellow flowers
37	34
38	35
39	36
40	37
41	38
42	
43	

Exercise 52
Arrange the data (separately) in order from smallest to largest in order to determine the five-number summary for each variety. Draw the boxplots and compare the skewness and the variabilities.

Exercises 53
Placing the data in order is helpful, but not required. When performing the calculations by hand, the following tables may be helpful. Note the order of red and yellow were switched for room considerations.

Yellow:

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $\left(x_i - \bar{x}\right)^2$
sum =		sum =		sum =	

Continued on next page

Exercises 53 continued
Red:
In the red data, you may choose to use 6 decimal place accuracy for $\bar{x}$ in order to calculate $x_i - \bar{x}$.

You may also choose to round $\left(x_i - \bar{x}\right)^2$ to five decimal places. Round s^2 to four decimal places

and s to three.

	Observations x_i		Deviations $x_i - \bar{x}$		Squared deviations $\left(x_i - \bar{x}\right)^2$
sum =		sum =		sum =	

Exercises 54 – 55
Carefully read Sections 5.8 – 5.9 before responding to these exercises.

Do You Know the Terms?

Cut out the following 19 flashcards to test yourself on Review Vocabulary. You can also find these flashcards at http://www.whfreeman.com/fapp7e.

Chapter 5 **Exploring Data: Distributions** **Boxplot**	**Chapter 5** **Exploring Data: Distributions** **Distribution**
Chapter 5 **Exploring Data: Distributions** **Exploratory data analysis**	**Chapter 5** **Exploring Data: Distributions** **Five-number summary**
Chapter 5 **Exploring Data: Distributions** **Histogram**	**Chapter 5** **Exploring Data: Distributions** **Individuals**
Chapter 5 **Exploring Data: Distributions** **Mean**	**Chapter 5** **Exploring Data: Distributions** **Median**
Chapter 5 **Exploring Data: Distributions** **Normal distributions**	**Chapter 5** **Exploring Data: Distributions** **Outlier**

The pattern of outcomes of a variable. The distribution describes what values the variable takes and how often each value occurs.	A graph of the five-number summary. A box spans the quartiles, with an interior line marking the median. Lines extend out from this box to the extreme high and low observations.
A summary of a distribution that gives the median, the first and third quartiles, and the largest and smallest observations.	The practice of examining data for overall patterns and special features, without necessarily seeking answers to specific questions.
The people, animals, or things described by a data set.	A graph of the distribution of outcomes (often divided into classes) for a single variable. The height of each bar is the number of observations in the class of outcomes covered by the base of the bar. All classes should have the same width.
The midpoint of a set of observations. Half the observations fall below the median and half fall above.	The ordinary arithmetic average of a set of observations. To find the mean, add all the observations and divide the sum by the number of observations summed.
A data point that falls clearly outside the overall pattern of a set of data.	A family of distributions that describe how often a variable takes its values by areas under a curve. The normal curves are symmetric and bell-shaped. A specific normal curve is completely described by giving its mean and its standard deviation.

Chapter 5
Exploring Data: Distributions

Quartiles

Chapter 5
Exploring Data: Distributions

68 – 95 - 99.7 rule

Chapter 5
Exploring Data: Distributions

Skewed distribution

Chapter 5
Exploring Data: Distributions

Standard deviation

Chapter 5
Exploring Data: Distributions

Standard deviation of a normal curve

Chapter 5
Exploring Data: Distributions

Stemplot

Chapter 5
Exploring Data: Distributions

Symmetric distribution

Chapter 5
Exploring Data: Distributions

Variable

Chapter 5
Exploring Data: Distributions

Variance

In any normal distribution, 68% of the observations lie within 1 standard deviation on either side of the mean; 95% lie within 2 standard deviations of the mean; and 99.7% lie within 3 standard deviations of the mean.	**The first quartile of a distribution is the point with 25% of the observations falling below it; the third quartile is the point with 75% below it.**
A measure of the spread of a distribution about its mean as center. It is the square root of the average squared deviation of the observations from their mean.	**A distribution in which observations on one side of the median extend notably farther from the median than do observations on the other side. In a right-skewed distribution, the larger observations extend farther to the right of the median than the smaller observations extend to the left.**
A display of the distribution of a variable that attaches the final digits of the observations as leaves on stems made up of all but the final digit.	**The standard deviation of a normal curve is the distance from the mean to the change-of-curvature points on either side.**
Any characteristic of an individual.	**A distribution with a histogram or stemplot in which the part to the left of the median is roughly a mirror image of the part to the right of the median.**
	A measure of the spread of a distribution about its mean. It is the average squared deviation of the observations from their mean. The square root of the variance is the standard deviation.

Learning the Calculator

Example 1

Construct a histogram given the following.

Value	Count
12	2
13	4
15	6
16	8
20	3

Solution

First enter the data by pressing the [STAT] button. The following screen will appear.

If there is data already stored, you may wish the clear it out. For example, if you wish to remove the data in L1, toggle to the top of the data and press [CLEAR] then [ENTER]. Repeat for any other data sets you wish to clear. Enter the new data being sure to press [ENTER] after each piece of data is displayed.

In order to display a histogram, you press [2nd] then [Y=]. This is equivalent to [STAT PLOT]. The following screen (or similar) will appear.

You will need to turn a stat plot On and choose the histogram option (⊞⊞). You will also need to make sure Xlist and Freq reference the correct data. In this case L1 and L2, respectively.
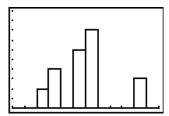
Next, you will need to make sure that no other graphs appear on your histogram. Press [Y=] and if another relation is present, either toggle to = and press enter to deselect or delete the relation.

You will next need to choose an appropriate window. By pressing [WINDOW] you need to enter an appropriate window that includes your smallest and largest pieces of data. These values dictate your choices of Xmin and Xmax. Your choice of Xscl is determined by the kind of data you are given. In this case, the appropriate choice is 1. If you are given data such as 10, 12, 14, 16, and values such as 11, 13, and 15 are not considered then the appropriate choice would be 2 in order to make the vertical bars touch. In terms of choices for frequency, Ymin should be set at zero. Ymax should be at least as large as the highest frequency value. Your choice of Yscl is determined by how large the maximum frequency value is from your table.

Next, we display the histogram by pressing the [GRAPH] button.

Notice that the histogram differs slightly from how a hand drawing should be. Ideally, the base of each rectangle should be shifted left by half of a unit.

Example 2

Given the following data, construct a histogram.

Class	Count
0 – 9	2
10 – 19	1
20 – 29	3
30 – 39	6
40 – 49	2

Solution

Follow the instructions in Example 1 in order to input data and set up the window to display the histogram. The width of the classes should be the Xscl in order to make the vertical bars touch. Also, in a case like this where you are given classes, use the left endpoint of the class as data pieces.

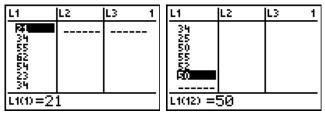## Example 3

Consider the following data.

$$21, 34, 55, 62, 54, 23, 34, 25, 50, 55, 52, 50$$

- Arrange the data in order from smallest to largest.

- Find the mean.
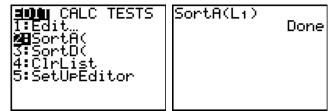
- Find the standard deviation.

- Find the five – number summary.

- Display the boxplot.

Enter the data, noting that there are 12 pieces of data. Make sure the location of the last entry corresponds to the total number of pieces of data.

To arrange the data in order from smallest to largest, press the [STAT] button and choose the SortA(option which sorts the data in ascending order. Choose the appropriate data set (in this case L1) and then press [ENTER] . The calculator will display Done indicating the data is sorted.

By pressing the [STAT] button, you can then view the data arranged in order by choosing the Edit option.

The data arranged from smallest to largest is as follows.

21, 23, 25, 34, 34, 50, 50, 52, 54, 55, 55, 62

To find the mean and standard deviation, press the [STAT] button. Toggle over to CALC and choose the 1-Var Stats option and then press the [ENTER]. You will get your home screen. Press [ENTER] again and you will then be able to determine the mean and standard deviation.

```
EDIT CALC TESTS        SortA(L₁)              1-Var Stats
■1:1-Var Stats                        Done    x̄=42.91666667
2:2-Var Stats          1-Var Stats            Σx=515
3:Med-Med                                     Σx²=24421
4:LinReg(ax+b)                                Sx=14.51931837
5:QuadReg                                     σx=13.901189
6:CubicReg                                    ↓n=12
7↓QuartReg
```

The mean is   (approximately 42.917) and the standard deviation is Sx (approximately 14.519).

To determine the five – number summary, from the last screen press the down arrow ([▼]) five times.

```
1-Var Stats
↑n=12
 minX=21
 Q₁=29.5
 Med=50
 Q₃=54.5
 maxX=62
```

The five – number summary is 21, 29.5, 50, 54.5, 62.

To display the box plot, press [2nd] then [Y=]. This is equivalent to [STAT PLOT]. You will need to choose ⊡⊢ for boxplot. Make sure the proper data are chosen for Xlist and Freq should be set at 1.

```
Plot1 Plot2 Plot3
On Off
Type: ⊡ ⊿ dⅲ
      ⊞⊷ ⊞ ⊿
Xlist:L₁
Freq:1
```

Choose an appropriate window for Xmin and Xmax based on the minimum and maximum values. The values you choose for Ymin and Ymax do not have an effect on the boxplot. You may choose values for Xscl and Yscl based on appearance of the axes. Display boxplot by pressing the [GRAPH] button.

```
WINDOW
 Xmin=20
 Xmax=70
 Xscl=10
 Ymin=0
 Ymax=1
 Yscl=1
 Xres=1
```

Practice Quiz

1. The weights (in pounds) of your cousins are: 120, 89, 108, 76, 21. Which are the outliers?

    a. 21 only.

    b. 120 only

    c. both 120 and 21

2. Below is a stemplot of the ages of adults on your block. Which statement is true?

2	1125
3	025788
4	15
5	257
6	25
7	8
8	1

    a. The stemplot is roughly symmetric.

    b. The stemplot is skewed to the higher ages.

    c. The two oldest people are outliers.

3. Here are 7 measured lengths (in inches): 13, 8, 5, 3, 8, 9, 12. Find their median.

    a. 3

    b. 8

    c. 8.3

4. Here are 7 measured lengths (in inches): 13, 8, 5, 3, 8, 9, 12. Find their mean.

    a. 3

    b. 8

    c. 8.3

5. The boxplot graph always includes the

    a. mean and median.

    b. quartiles and the standard deviation.

    c. quartiles and the median.

6. The percentage of scores on a standardized exam that lie between the first and third quartiles is

    a. 25%.

    b. 50%.

    c. 75%.

7. If the mean of the data 2, 4, 6, 3, 5, 8, 7 is 5, what is its standard deviation?

    a. $\sqrt{\frac{12}{7}}$

    b. 4

    c. 2

8. The scores on a marketing exam were normally distributed with a mean of 67 and a standard deviation of 9. Find the first quartile (Q_1) for the test scores.
   a. 58
   b. 61
   c. 25

9. Given the following data, find the five-number summary.

   5, 8, 12, 15, 11, 21, 9, 12

   a. 5, 8.5, 11.5, 13.5, 21
   b. 5, 8.5, 12, 13, 21
   c. 5, 12, 13, 21, 12

10. The amount of coffee a certain dispenser fills 12 oz coffee cups with is normally distributed with a mean of 10.9 oz and a standard deviation of 0.2 oz. What percent of cups dispense more than 11.1 oz?
   a. 68%
   b. 5%
   c. 16%

Word Search

Refer to pages 207 – 208 of your text to obtain the Review Vocabulary. There are 17 hidden vocabulary words/expressions in the word search below. *Standard deviation of a normal curve* and *68-95-99.7 rule* were both omitted from the word search. It should be noted that spaces and hyphens are removed.

```
I Y E Q T E L L I A T S O P M S T S Y V G N I I A S
M U E R E R F J U O R M T E I T Z C N F P E N G I A
G R G E N I P Q G V O E I J R P S O G Q S E I S N K
S R J R D P T F O E I A K N E H A T E T P H I K B T
I O H I S T O G R A M T C O Z F L I A L C A M E L H
X A S C S J Y O E L B A I R A V C N I K R Q H W L D
S J Y F T O Z H T X S P G M W O D H L O I A P E C E
O N M A E U D R G F L N T A N A T O L P X O B D U G
E I M I M T E G R S E I K L R E A M O L J I S D A K
E O E X P L O R A T O R Y D A T A A N A L Y S I S E
T F T U L I A M M S G K D I P E Q V S D H M J S S F
M T R O O E E L E E S E Y S T H E H L I E E T T X P
X Q I I T R T O D L V Z D T E C N E A S A A Q R G E
A D C I I S Z N I I D E N R M E S A U T O N H I F O
S F D D T X N T A T A T F I L C F R D R E E A B E F
Y I I E W K V T N R D F R B A E S T I I L C A U E N
O T S R K G I Z X A H E K U F H R O V B G N A T A J
Y E T O T O Z L E U Z C A T S I X T I U V A A I D I
E M R V N N I F X Q M I O I O U R H D T E I E O A E
I Y I E N L A P T E K R U O N S O L N I O R Z N N S
V D B S N E E A Q F S O E N M H C X I O J A C H H W
L O U A D N M Y R A M M U S R E B M U N E V I F J I
A I T Y L F O S T D B I O O S C M A F N N F U I G S
K S I P N K D S S Q G D S W H O U G N A R F R O M X
M N O M U A Z S I O E Y Q K N T Q S O H O H S L M F
P X N S A I Q P U R S R R D R Z E Q S A E K R R R H
```

1. _____

2. _____

3. _____

4. _____

5. _____

6. _____

7. _____

8. _____

9. _____

10. _____

11. _____

12. _____

13. _____

14. _____

15. _____

16. _____

17. _____