4

The Average and the Standard Deviation

It is difficult to understand why statisticians commonly limit their enquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once.

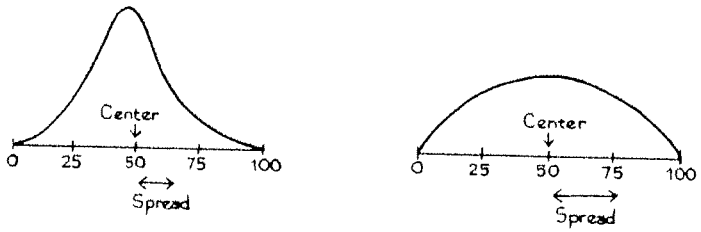
—SIR FRANCIS GALTON (ENGLAND, 1822–1911)[1]

1. INTRODUCTION

A histogram can be used to summarize large amounts of data. Often, an even more drastic summary is possible, giving just the center of the histogram and the spread around the center. ("Center" and "spread" are ordinary words here, without any special technical meaning.) Two histograms are sketched in figure 1 on the next page. The center and spread are shown. Both histograms have the same center, but the second one is more spread out—there is more area farther away from the center. For statistical work, precise definitions have to be given, and there are several ways to go about this. The *average* is often used to find the center, and so is the *median*.[2] The *standard deviation* measures spread around the average; the *interquartile range* is another measure of spread.

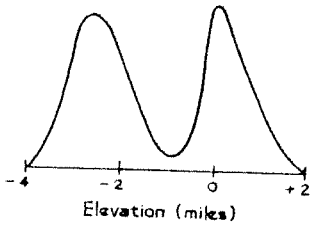
The histograms in figure 1 can be summarized by the center and the spread. However, things do not always work out so well. For instance, figure 2 gives the distribution of elevation over the earth's surface. Elevation is shown along the

Figure 1. Center and spread. The centers of the two histograms are the same, but the second histogram is more spread out.



horizontal axis, in miles above (+) or below (−) sea level. The area under the histogram between two elevations gives the percentage of the earth's surface area between those elevations. There are clear peaks in this histogram. Most of the surface area is taken up by the sea floors, around 3 miles below sea level; or the continental plains, around sea level. Reporting only the center and spread of this histogram would miss the two peaks.[3]

Figure 2. Distribution of the surface area of the earth by elevation above (+) or below (−) sea level.



2. THE AVERAGE

The object of this section is to review the average; the difference between *cross-sectional* and *longitudinal* surveys will also be discussed. The context is HANES—the Health and Nutrition Examination Survey, in which the Public Health Service examines a representative cross section of Americans. This survey has been done at irregular intervals since 1959 (when it was called the Health Examination Survey). The objective is to get baseline data about—

- demographic variables, like age, education, and income;
- physiological variables like height, weight, blood pressure, and serum cholesterol levels;
- dietary habits;
- prevalence of diseases.

Subsequent analysis focuses on the interrelationships among the variables, and has some impact on health policy.[4]

The HANES2 sample was taken during the period 1976–80. Before looking at the data, let's make a quick review of averages.

> The average of a list of numbers equals their sum, divided by how many there are.

For instance, the list 9, 1, 2, 2, 0 has 5 entries, the first being 9. The average of the list is

$$\frac{9+1+2+2+0}{5} = \frac{14}{5} = 2.8$$

Let's get back to HANES. What did the men and women in the sample (age 18–74) look like?

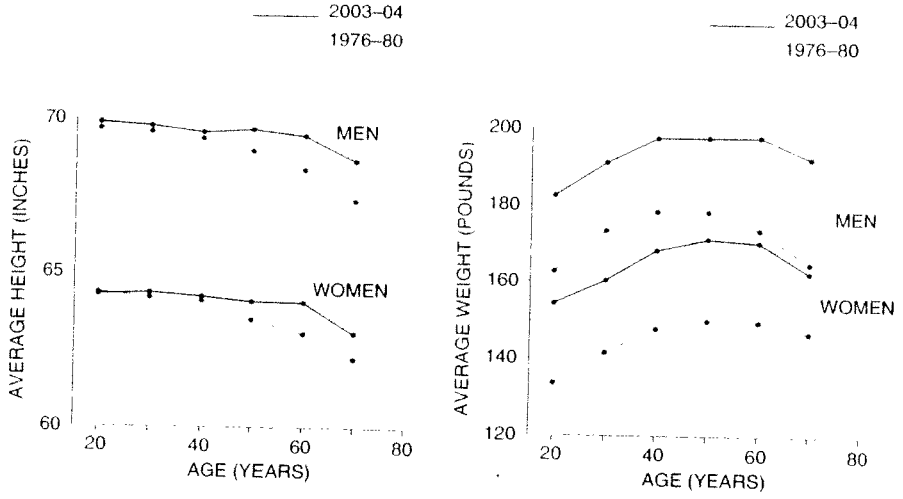
- The average height of the men was 5 feet 9 inches, and their average weight was 171 pounds.
- The average height of the women was 5 feet 3.5 inches, and their average weight was 146 pounds.

They're pretty chubby.

What's happened since 1980? The survey was done again in 2003–04 (HANES5). Average heights went up by a fraction of an inch, while weights went up by nearly 20 pounds—both for men and for women.

Figure 3 shows the averages for men and women, and for each age group; averages are joined by straight lines. From HANES2 to HANES5, average heights went up a little in each group—but average weights went up a lot. This could become a serious public-health problem, because excess weight is associated with many diseases, including heart disease, cancer, and diabetes.

Figure 3. Age-specific average heights and weights for men and women 18–74 in the HANES sample. The panel on the left shows height, the panel on the right shows weight.



Source: www.cdc.gov/nchs/nhanes.htm

The average is a powerful way of summarizing data—many histograms are compressed into the four curves. But this compression is achieved only by smoothing away individual differences. For instance, in 2003–04, the average height of the men age 18–24 was 5 feet 10 inches. But 15% of them were taller than 6 feet 1 inch; another 15% were shorter than 5 feet 6 inches. This diversity is hidden by the average.

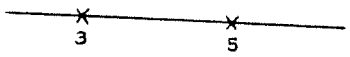
For a moment, we return to design issues (chapter 2). In the 1976–80 data, the average height of men appears to decrease after age 20, dropping about two inches in 50 years. Similarly for women. Should you conclude that an average person got shorter at this rate? Not really. HANES is *cross-sectional*, not *longitudinal*. In a cross-sectional study, different subjects are compared to each other at one point in time. In a longitudinal study, subjects are followed over time, and compared with themselves at different points in time. The people age 18–24 in figure 3 are completely different from those age 65–74. The first group was born a lot later than the second.

There is evidence to suggest that, over time, Americans have been getting taller. This is called the *secular trend* in height, and its effect is confounded with the effect of age in figure 3. Most of the two-inch drop in height seems to be due to the secular trend. The people age 65–74 were born around 50 years before those age 18–24, and are an inch or two shorter for that reason.[5] On the other hand, the secular trend has slowed down. (Reasons are unclear.) Average heights only increased a little from 1976–80 to 2003–04. The slowing also explains why the height curves for 2003–04 are flatter than the curves for 1976–80.

> If a study draws conclusions about the effects of age, find out whether the data are cross-sectional or longitudinal.

Exercise Set A

1. (a) The numbers 3 and 5 are marked by crosses on the horizontal line below. Find the average of these two numbers and mark it by an arrow.



(b) Repeat (a) for the list 3, 5, 5.

(c) Two numbers are shown below by crosses on a horizontal axis. Draw an arrow pointing to their average.

2. A list has 10 entries. Each entry is either 1 or 2 or 3. What must the list be if the average is 1? If the average is 3? Can the average be 4?

3. Which of the following two lists has a bigger average? Or are they the same? Try to answer without doing any arithmetic.

   (i) 10, 7, 8, 3, 5, 9 (ii) 10, 7, 8, 3, 5, 9, 11

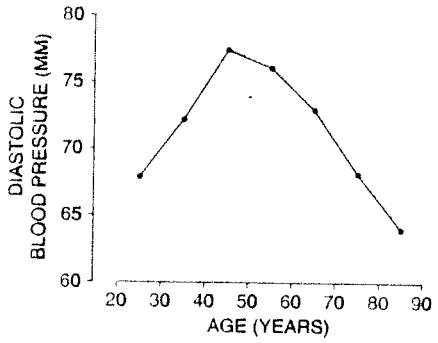
4. Ten people in a room have an average height of 5 feet 6 inches. An 11th person, who is 6 feet 5 inches tall, enters the room. Find the average height of all 11 people.

5. Twenty-one people in a room have an average height of 5 feet 6 inches. A 22nd person, who is 6 feet 5 inches tall, enters the room. Find the average height of all 22 people. Compare with exercise 4.

6. Twenty-one people in a room have an average height of 5 feet 6 inches. A 22nd person enters the room. How tall would he have to be to raise the average height by 1 inch?

7. In figure 2, are the Rocky Mountains plotted near the left end of the axis, the middle, or the right end? What about Kansas? What about the trenches in the sea floor, like the Marianas trench?

8. Diastolic blood pressure is considered a better indicator of heart trouble than systolic pressure. The figure below shows age-specific average diastolic blood pressure for the men age 20 and over in HANES5 (2003–04).[6] True or false: the data show that as men age, their diastolic blood pressure increases until age 45 or so, and then decreases. If false, how do you explain the pattern in the graph? (Blood pressure is measured in "mm," that is, millimeters of mercury.)



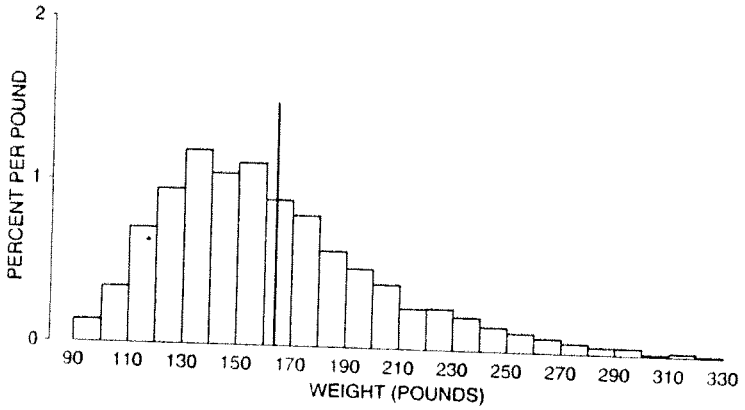
9. Average hourly earnings are computed each month by the Bureau of Labor Statistics using payroll data from commercial establishments. The Bureau figures the total wages paid out (to nonsupervisory personnel), and divides by the total hours worked. During recessions, average hourly earnings typically go up. When the recession ends, average hourly earnings often start going down. How can this be?

The answers to these exercises are on pp. A47–48.

3. THE AVERAGE AND THE HISTOGRAM

This section will indicate how the average and the median are related to histograms. To begin with an example, there were 2,696 women age 18 and over in HANES5 (2003–04). Their average weight was 164 pounds. It is natural to guess

Figure 4. Histogram for the weights of the 2,696 women in the HANES5 sample. The average is marked by a vertical line. Only 41% of the women were above average in weight.

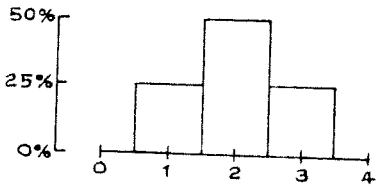


Source: www.cdc.gov/nchs/nhanes.htm.

that 50% of them were above average in weight, and 50% were below average. However, this guess is somewhat off. In fact, only 41% were above average, and 59% were below average. Figure 4 shows a histogram for the data: the average is marked by a vertical line. In other situations, the percentages can be even farther from 50%.

How is this possible? To find out, it is easiest to start with some hypothetical data—the list 1, 2, 2, 3. The histogram for this list (figure 5) is symmetric about the value 2. And the average equals 2. If the histogram is symmetric around a value, that value equals the average. Furthermore, half the area under the histogram lies to the left of that value, and half to the right. (What does symmetry mean? Imagine drawing a vertical line through the center of the histogram and folding the histogram in half around that line: the two halves should match up.)

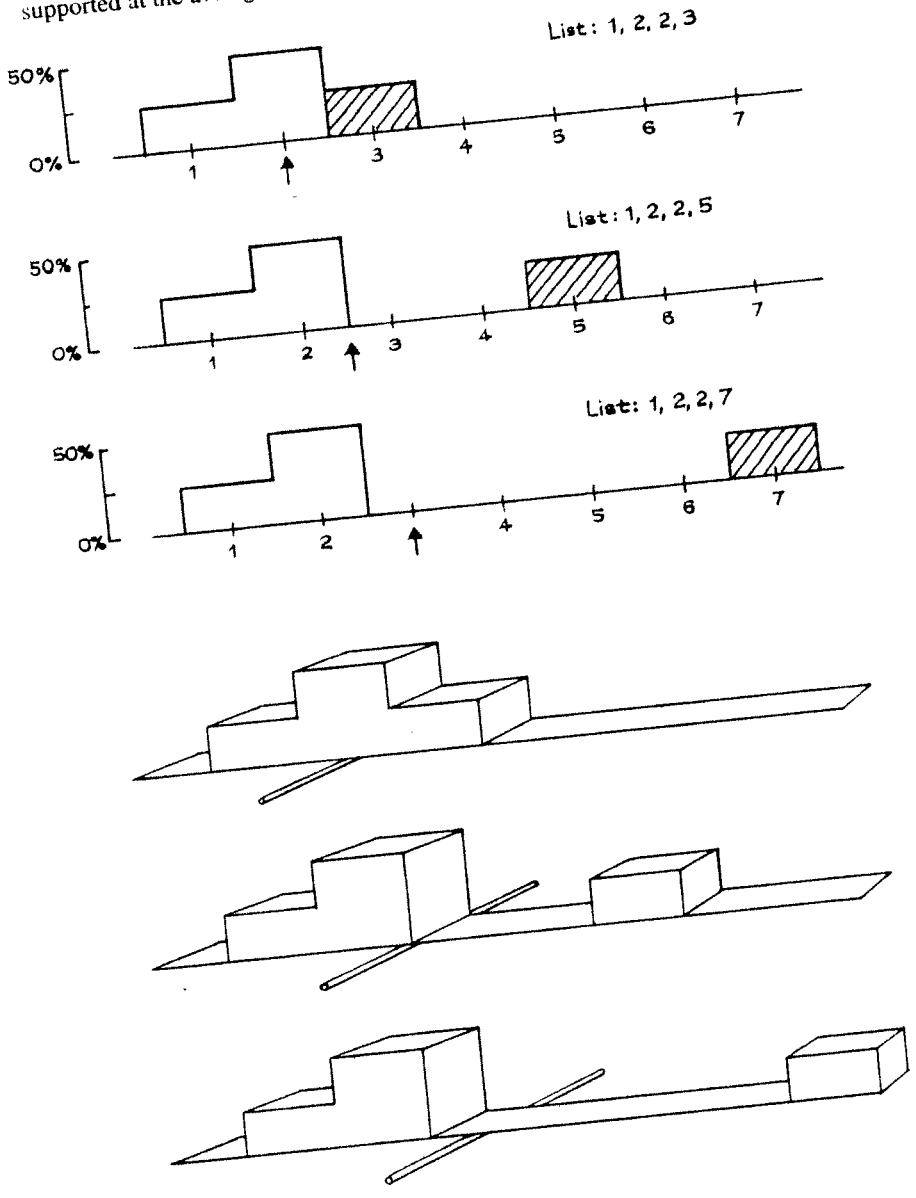
Figure 5. Histogram for the list 1, 2, 2, 3. The histogram is symmetric around 2, the average: 50% of the area is to the left of 2, and 50% is to the right.



What happens when the value 3 on the list 1, 2, 2, 3 is increased, say to 5 or 7? As shown in figure 6, the rectangle over that value moves off to the right, destroying the symmetry. The average for each histogram is marked with an arrow, and the arrow shifts to the right following the rectangle. To see why, imagine the histogram is made out of wooden blocks attached to a stiff, weightless board. Put the histogram across a taut wire, as illustrated in the bottom panel of figure 6. The

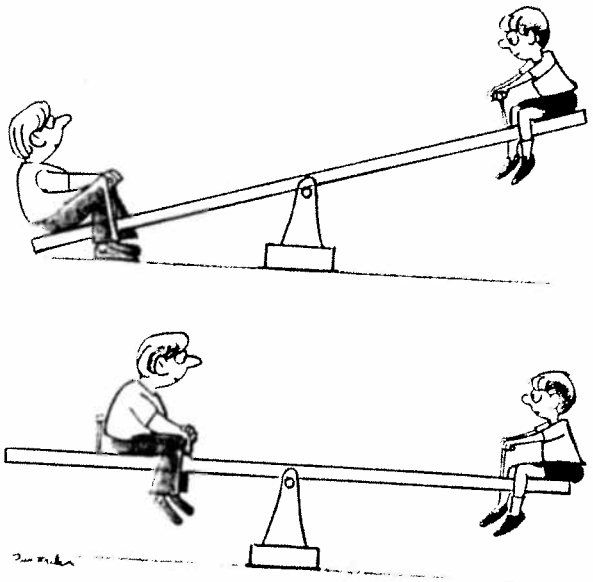
histogram will balance at the average.[7] A small area far away from the average can balance a large area close to the average, because areas are weighted by their distance from the balance point.

Figure 6. The average. The top panel shows three histograms; the averages are marked by arrows. As the shaded box moves to the right, it pulls the average along with it. The area to the left of the average gets up to 75%. The bottom panel shows the same three histograms made out of wooden blocks attached to a stiff, weightless board. The histograms balance when supported at the average.



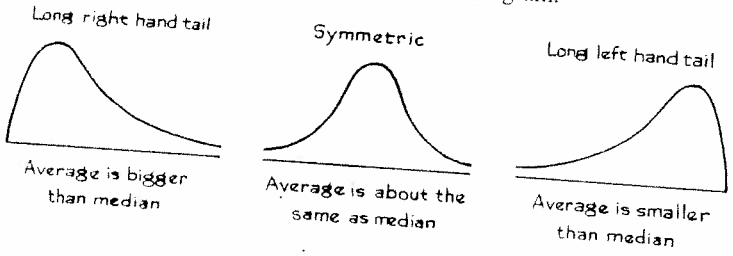
A histogram balances when supported at the average.

A small child sits farther away from the center of a seesaw in order to balance a large child sitting closer to the center. Blocks in a histogram work the same way. That is why the percentage of cases on either side of the average can differ from 50%.



The *median* of a histogram is the value with half the area to the left and half to the right. For all three histograms in figure 6, the median is 2. With the second and third histograms, the area to the right of the median is far away by comparison with the area to the left. Consequently, if you tried to balance one of those histograms at the median, it would tip to the right. More generally, the average is to the right of the median whenever the histogram has a long right-hand tail, as in figure 7. The weight histogram (figure 4 on p. 62) had an average of 164 lbs and a median of 155 lbs. The long right-hand tail is what made the average bigger than the median.

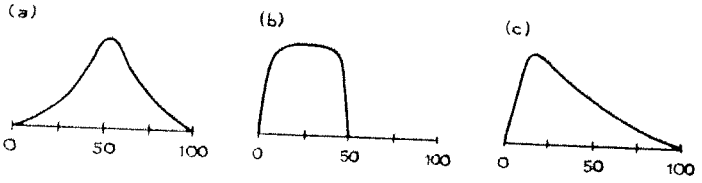
Figure 7. The tails of a histogram.



For another example, median family income in the U.S. in 2004 was about $54,000. The income histogram has a long right-hand tail, and the average was higher—$60,000.[8] When dealing with long-tailed distributions, statisticians might use the median rather than the average, if the average pays too much attention to the extreme tail of the distribution. We return to this point in the next chapter.

Exercise Set B

1. Below are sketches of histograms for three lists. Fill in the blank for each list: the average is around _____. Options: 25, 40, 50, 60, 75.



2. For each histogram in exercise 1, is the median equal to the average? or is it to the left? to the right?

3. Look back at the cigarette histogram on p. 42. The median is around _____. Fill in the blank. Options: 10, 20, 30, 40

4. For this cigarette histogram, is the average around 15, 20, or 25?

5. For registered students at universities in the U.S., which is larger: average age or median age?

6. For each of the following lists of numbers, say whether the entries are on the whole around 1, 5, or 10 in size. No arithmetic is needed.

   (a) 1.3, 0.9, 1.2, 0.8 (b) 13, 9, 12, 8
   (c) 7, 3, 6, 4 (d) 7, −3, −6, 4

The answers to these exercises are on pp. A48–49.

Technical note. The median of a list is defined so that half or more of the entries are at the median or bigger, and half or more are at the median or smaller. This will be illustrated on 4 lists—

(a) 1, 5, 7

(b) 1, 2, 5, 7

(c) 1, 2, 2, 7, 8

(d) 8, −3, 5, 0, 1, 4, −1

For list (a), the median is 5: two entries out of the three are 5 or more, and two are 5 or less. For list (b), any value between 2 and 5 is a median; if pressed, most statisticians would choose 3.5 (which is halfway between 2 and 5) as "the" median. For list (c), the median is 2: four entries out of five are 2 or more, and three are 2 or less. To find the median of list (d), arrange it in increasing order:

$$-3, \quad -1, \quad 0, \quad 1, \quad 4, \quad 5, \quad 8$$

There are seven entries on this list: four are 1 or more, and four are 1 or less. So, 1 is the median.

4. THE ROOT-MEAN-SQUARE

The next main topic in the chapter is the *standard deviation*, which is used to measure spread. This section presents a mathematical preliminary, illustrated on the list

$$0, \quad 5, \quad -8, \quad 7, \quad -3$$

How big are these five numbers? The average is 0.2, but this is a poor measure of size. It only means that to a large extent, the positives cancel the negatives. The simplest way around the problem would be to wipe out the signs and then take the average. However, statisticians do something else: they apply the *root-mean-square* operation to the list. The phrase "root-mean-square" says how to do the arithmetic, provided you remember to read it backwards:

- SQUARE all the entries, getting rid of the signs.
- Take the MEAN (average) of the squares.
- Take the square ROOT of the mean.

This can be expressed as an equation, with root-mean-square abbreviated to r.m.s.

$$\text{r.m.s. size of a list} = \sqrt{\text{average of (entries}^2\text{)}}.$$

Example 1. Find the average, the average neglecting signs, and the r.m.s. size of the list 0, 5, −8, 7, −3.

Solution.

$$\text{average} = \frac{0+5-8+7-3}{5} = 0.2$$

$$\text{average neglecting signs} = \frac{0+5+8+7+3}{5} = 4.6$$

$$\text{r.m.s. size} = \sqrt{\frac{0^2+5^2+(-8)^2+7^2+(-3)^2}{5}} = \sqrt{29.4} \approx 5.4$$

The r.m.s. size is a little bigger than the average neglecting signs. It always turns out like that—except in the trivial case when all the entries are the same size. The root and the square do not cancel, due to the intervening operation of taking the mean. (The "≈" means "nearly equal:" some rounding has been done.)

There doesn't seem to be much to choose between the 5.4 and the 4.6 as a measure of the overall size for the list in the example. Statisticians use the r.m.s. size because it fits in better with the algebra that they have to do.[9] Whether this explanation is appealing or not, don't worry. Everyone is suspicious of the r.m.s. at first, and gets used to it very quickly.

Exercise Set C

1. (a) Find the average and the r.m.s. size of the numbers on the list 1, −3, 5, −6, 3.
   (b) Do the same for the list −11, 8, −9, −3, 15.
2. Guess whether the r.m.s. size of each of the following lists of numbers is around 1, 10, or 20. No arithmetic is required.
   (a) 1, 5, −7, 8, −10, 9, −6, 5, 12, −17
   (b) 22, −18, −33, 7, 31, −12, 1, 24, −6, −16
   (c) 1, 2, 0, 0, −1, 0, 0, −3, 0, 1
3. (a) Find the r.m.s. size of the list 7, 7, 7, 7.
   (b) Repeat, for the list 7, −7, 7, −7.
4. Each of the numbers 103, 96, 101, 104 is almost 100 but is off by some amount. Find the r.m.s. size of the amounts off.
5. The list 103, 96, 101, 104 has an average. Find it. Each number in the list is off the average by some amount. Find the r.m.s. size of the amounts off.
6. A computer is programmed to predict test scores, compare them with actual scores, and find the r.m.s. size of the prediction errors. Glancing at the printout, you see the r.m.s. size of the prediction errors is 3.6, and the following results for the first ten students:

predicted score:	90	90	87	80	42	70	67	60	83	94
actual score:	88	70	81	85	63	77	66	49	71	69

Does the printout seem reasonable, or is something wrong with the computer?

The answers to these exercises are on p. A49.

5. THE STANDARD DEVIATION

As the quote at the beginning of the chapter suggests, it is often helpful to think of the way a list of numbers spreads out around the average. This spread is usually measured by a quantity called the *standard deviation*, or SD. The SD measures the size of deviations from the average: it is a sort of average deviation. The program is to interpret the SD in the context of real data, and then see how to calculate it.

There were 2,696 women age 18 and over in the HANES5 sample. The average height of these women was about 63.5 inches, and the SD was close to 3 inches. The average tells us that most of the women were somewhere around 63.5 inches tall. But there were deviations from the average. Some of the women were taller than average, some shorter. How big were these deviations? That is where the SD comes in.

> The SD says how far away numbers on a list are from their average. Most entries on the list will be somewhere around one SD away from the average. Very few will be more than two or three SDs away.