PART I

Descriptive Statistics

1.

The Histogram

> *Grown-ups love figures. When you tell them that you have made a new friend, they never ask you any questions about essential matters. They never say to you, "What does his voice sound like? What games does he love best? Does he collect butterflies?" Instead, they demand: "How old is he? How many brothers has he? How much does he weigh? How much money does his father make?" Only from these figures do they think they have learned anything about him.*
>
> —*The Little Prince*[1]

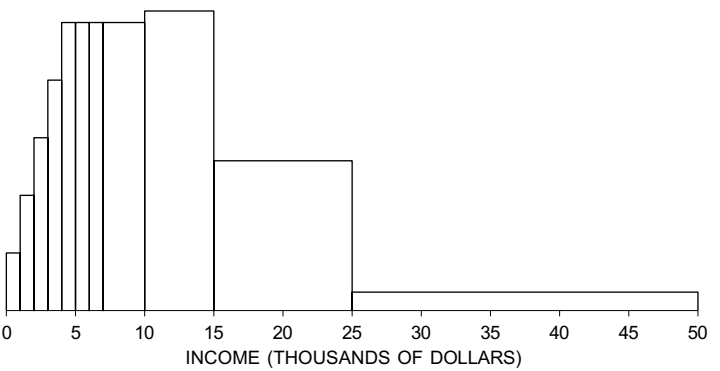
1. INTRODUCTION

In the U.S., how are incomes distributed? How much worse off are minority groups? Some information is provided by government statistics, obtained from the Current Population Survey. Each month, interviewers talk to a representative cross section of about 50,000 American families (for details, see part VI). In March, these families are asked to report their incomes for the previous year. We are going to look at the results for 1973. These data have to be summarized—nobody

wants to look at 50,000 numbers. To summarize data, statisticians often use a graph called a *histogram* (figure 1 on the next page).

This section explains how to read histograms. First of all, there is no vertical scale: unlike most other graphs, a histogram does not need a vertical scale. Now look at the horizontal scale. This shows income in thousands of dollars. The graph itself is just a set of blocks. The bottom edge of the first block covers the range from \$0 to \$1,000, the bottom edge of the second goes from \$1,000 to \$2,000;

Figure 1. A histogram. This graph shows the distribution of families by income in the U.S. in 1973.



Source: Current Population Survey.[2]

and so on until the last block, which covers the range from \$25,000 to \$50,000. These ranges are called *class intervals*. The graph is drawn so the area of a block is proportional to the number of families with incomes in the corresponding class interval.

To see how the blocks work, look more closely at figure 1. About what percentage of the families earned between \$10,000 and \$15,000? The block over this interval amounts to something like one-fourth of the total area. So about one-fourth, or 25%, of the families had incomes in that range.

Take another example. Were there more families with incomes between \$10,000 and \$15,000, or with incomes between \$15,000 and \$25,000? The block over the first interval is taller, but the block over the second interval is wider. The areas of the two blocks are about the same, so the percentage of families earning \$10,000 to \$15,000 is about the same as the percentage earning \$15,000 to \$25,000.

For a last example, take the percentage of families with incomes under \$7,000. Is this closest to 10%, 25%, or 50%? By eye, the area under the histogram between \$0 and \$7,000 is about a quarter of the total area, so the percentage is closest to 25%.

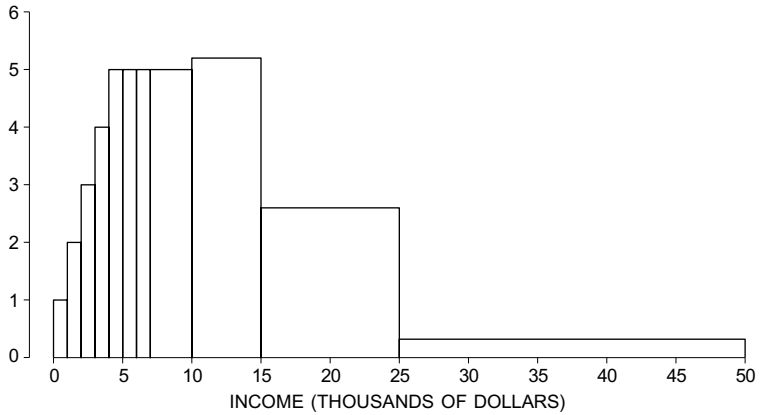
In a histogram, the areas of the blocks represent percentages.

The horizontal axis in figure 1 stops at \$50,000. What about the families earning

more than that? The histogram simply ignores them. In 1973, only 1% of American families had incomes above that level: most are represented in the figure.

At this point, a good way to learn more about histograms is to do some exercises. Figure 2 shows the same histogram as figure 1, but with a vertical scale supplied. This scale will be useful in working exercise 1. Exercise 8 compares the income data for 1973 and 2004.

Figure 2. The histogram from figure 1, with a vertical scale supplied.



2. DRAWING A HISTOGRAM

This section explains how to draw a histogram. The method is not difficult, but there are a couple of wrong turns to avoid. The starting point in drawing a histogram is a *distribution table*, which shows the percentage of families with incomes in each class interval (table 1). These percentages are found by going back to the original data—on the 50,000 families—and counting. Nowadays this sort of work is done by computer, and in fact table 1 was drawn up with the help of a computer at the Bureau of the Census.

The computer has to be told what to do with families that fall right on the boundary between two class intervals. This is called an *endpoint convention*. The convention followed in table 1 is indicated by the caption. The left endpoint is included in the class interval, the right endpoint is excluded. In the first line of the table, for example, \$0 is included and \$1,000 is excluded. This interval has the families that earn \$0 or more, but less than \$1,000. A family that earns \$1,000 exactly goes in the next interval.

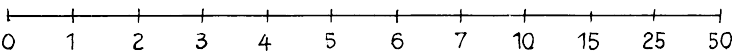
Table 1. Distribution of families by income in the U.S. in 1973. Class intervals include the left endpoint, but not the right endpoint.

Income level	*Percent*
\$0–\$1,000	1
\$1,000–\$2,000	2
\$2,000–\$3,000	3
\$3,000–\$4,000	4
\$4,000–\$5,000	5
\$5,000–\$6,000	5
\$6,000–\$7,000	5
\$7,000–\$10,000	15

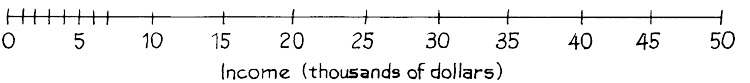
$10,000–$15,000	26
$15,000–$25,000	26
$25,000–$50,000	8
$50,000 and over	1

Note: Percents do not add to 100%, due to rounding.
Source: Current Population Survey.[4]

The first step in drawing a histogram is to put down a horizontal axis. For the income histogram, some people get

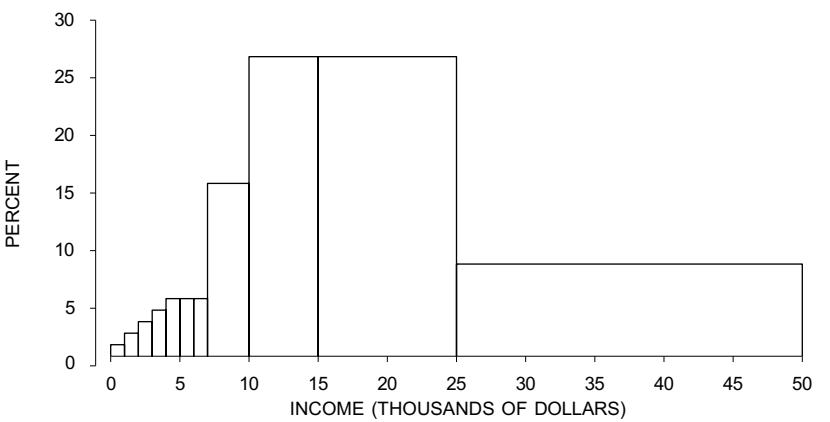


That is a mistake. The interval from $7,000 to $10,000 is three times as long as the interval from $6,000 to $7,000. So the horizontal axis should look like this:



The next step is to draw the blocks. It's tempting to make their heights equal to the percents in the table. Figure 3 shows what happens if you make that mistake. The graph gives much too rosy a picture of the income distribution. For example, figure 3 says there were many more families with incomes over $25,000 than under $7,000. The U.S. was a rich country in 1973, but not that rich.

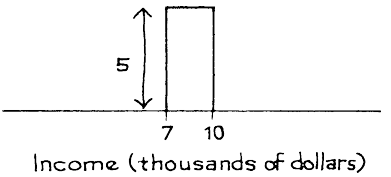
Figure 3. Don't plot the percents.



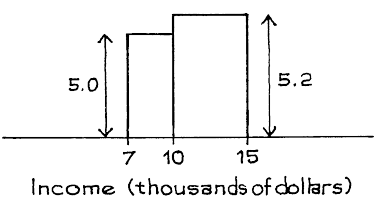
The source of the trouble is that some class intervals are longer than others, so the percents in table 1 are not on a par with one another. The 8% who earn $25,000 to $50,000, for instance, are spread over a much larger range of incomes than the 15% who earn $7,000 to $10,000. Plotting percents directly ignores this, and makes the blocks over the longer class intervals too big.

There is a simple way to compensate for the different lengths of the class intervals—use thousand-dollar intervals as a common unit. For example, the class interval from $7,000 to $10,000 contains three of these intervals: $7,000 to $8,000, $8,000 to $9,000, and $9,000 to $10,000. From table 1, 15% of the

families had incomes in the whole interval. Within each of the thousand-dollar sub-intervals, there will only be about 5% of the families. This 5, not the 15, is what should be plotted above the interval $7,000 to $10,000.



For a second example, take the interval from $10,000 to $15,000. This contains 5 of the thousand-dollar intervals. According to table 1, 26% of the families had incomes in the whole interval. Within each of the 5 smaller intervals there will be about 5.2% of the families: 26/5= 5.2. The height of the block over the interval $10,000 to $15,000 is 5.2.

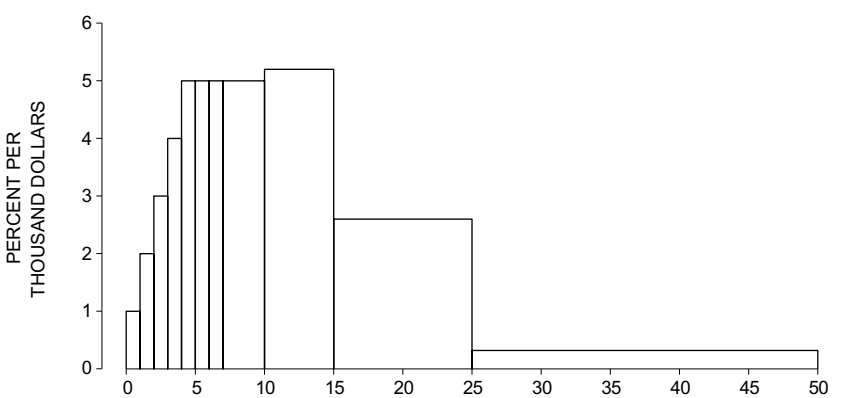


The work is done for two of the lines in table 1. To complete the histogram, do the same thing for the rest of the class intervals. Figure 4 (below) is the result.

> To figure out the height of a block over a class interval, divide the percentage by the length of the interval.

That way, the area of the block equals the percentage of families in the class interval. The histogram represents the distribution as if the percent is spread evenly over the class interval. Often, this is a good first approximation.

Figure 4. Distribution of families by income in the U.S. in 1973.



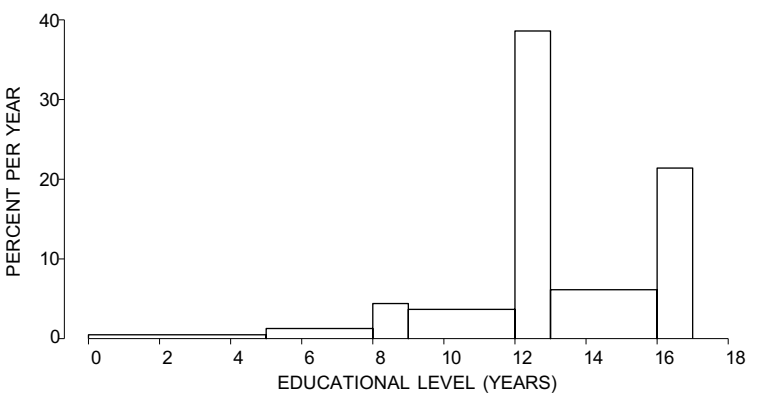
The procedure is straightforward, but the units on the vertical scale are a little complicated. For instance, to get the height of the block over the interval \$7,000 to \$10,000, you divide 15 percent by 3 thousand dollars. So the units for the answer are percent per thousand dollars. Think about the "per" just as you would when reading that there are 50,000 people per square mile in Tokyo: in each square mile of the city, there are about 50,000 people. It is the same with histograms. The height of the block over the interval \$7,000 to \$10,000 is 5% per thousand dollars: in each thousand-dollar interval between \$7,000 and \$10,000, there are about 5% of the families. Figure 4 shows the complete histogram with these units on the vertical scale.

3. THE DENSITY SCALE

When reading areas off a histogram, it is convenient to have a vertical scale. The income histogram in the previous section was drawn using the *density scale*.[5] The unit on the horizontal axis was \$1,000 of family income, and the vertical axis showed the percentage of families per \$1,000 of income. Figure 5 is another example of a histogram with a density scale. This is a histogram for educational level of persons age 25 and over in the U.S. in 1991. "Educational level" means years of schooling completed; kindergarten doesn't count.

The endpoint convention followed in this histogram is a bit fussy. The block over the interval 8–9 years, for example, represents all the people who finished eighth grade, but not ninth grade; people who dropped out part way through ninth

Figure 5. Distribution of persons age 25 and over in the U.S. in 1991 by educational level.



Source: *Statistical Abstract*, 1992, Table 220.

grade are included. The units on the horizontal axis of the histogram are years, so the units on the vertical axis are percent per year. For instance, the height of the histogram over the interval 13–16 years is 6% per year. In other words, about 6% of the population finished the first year of college, another 6% finished the second year, and another 6% finished the third year.

Section 1 described how area in a histogram represents percent. If one block covers a larger area than another, it represents a larger percent of the cases. What does the height of a block represent? Look at the horizontal axis in figure 5. Imag-

ine the people lined up on this axis, with each person stationed at his or her educational level. Some parts of the axis—years—will be more crowded than others. The height of the histogram shows the crowding.

The histogram is highest over the interval 12–13 years, so the crowding is greatest there. This interval has all the people with high-school degrees. (Some people in this interval may have gone on to college, but they did not even finish the first year.) There are two other peaks, a small one at 8–9 years (finishing middle school) and a big one at 16–17 years—finishing college. The peaks show how people tend to stop their schooling at one of the three possible graduations rather than dropping out in between.

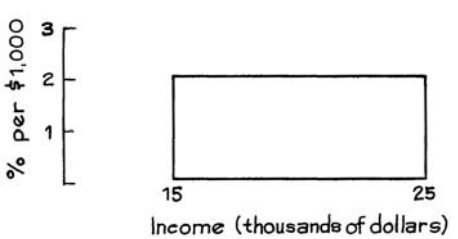
At first, it may be difficult to keep apart the notion of the crowding in an interval, represented by the height of the block, and the number in an interval, represented by the area of the block. An example will help. Look at the blocks over the intervals 8–9 years and 9–12 years in figure 5. The first block is a little taller, so this interval is a little more crowded. However, the block over 9–12 years has a much larger area, so this interval has many more people. Of course, there is more room in the second interval—it's 3 times as long. The two intervals are like the Netherlands and the U.S. The Netherlands is more crowded, but the U.S. has more people.

> In a histogram, the height of a block represents crowding—percentage per horizontal unit.

By contrast, the area of the block represents the percentage of cases in the corresponding class interval (section 1).

Once you learn how to use it, the density scale can be quite helpful. For example, take the interval from 9 to 12 years in figure 5—the people who got through their first year of high school but didn't graduate. The height of the block over this interval is nearly 4% per year. In other words, each of the three one-year intervals 9–10, 10–11, and 11–12 holds nearly 4% of the people. So the whole three-year interval must hold nearly $3 \times 4\% = 12\%$ of the people. Nearly 12% of the population age 25 and over got through at least one year of high school, but failed to graduate.

Example 1. The sketch below shows one block of the family-income histogram for a certain city. About what percent of the families in the city had incomes between \$15,000 and \$25,000?



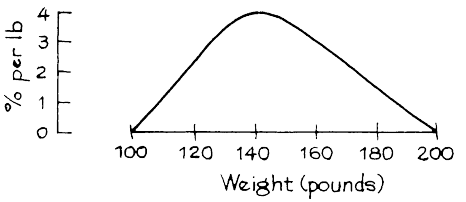
Solution. The height of the block is 2% per thousand dollars. Each thousand-dollar interval between \$15,000 and \$25,000 contains about 2% of the families in the city. There are 10 of these thousand-dollar intervals between

$15,000 and $25,000. The answer is $10 \times 2\% = 20\%$. About 20% of the families in the city had incomes between $15,000 and $25,000.

The example shows that with the density scale, the areas of the blocks come out in percent. The horizontal units—thousands of dollars—cancel:

$$2\% \text{ per thousand dollars} \times 10 \text{ thousand dollars} = 20\%.$$

Example 2. Someone has sketched a histogram for the weights of some people, using the density scale. What's wrong?



Solution. The total area is 200%, and should only be 100%. The area can be calculated as follows. The histogram is almost a triangle, whose height is 4% per pound and whose base is 200 lb − 100 lb = 100 lb. The area is

$$\frac{1}{2} \times \text{base} \times \text{height} = \frac{1}{2} \times 100 \text{ lb} \times 4\% \text{ per lb} = 200\%.$$

> With the density scale on the vertical axis, the areas of the blocks come out in percent. The area under the histogram over an interval equals the percentage of cases in that interval.[6] The total area under the histogram is 100%.

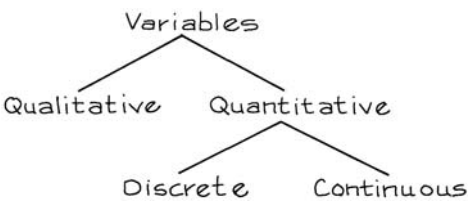
Since 1991, the educational level in the U.S. has continued to increase. Then, 21% of the population had a bachelor's degree or better (the "population" means people age 25 and over). In 2005, the corresponding figure was 28%.

4. VARIABLES

The Current Population Survey covers many other variables besides income. A *variable* is a characteristic which changes from person to person in a study. Interviewers for the survey use a battery of questions: How old are you? How many people are there in your family? What is your family's total income? Are you married? Do you have a job? The corresponding variables would be: age, fam- ily size, family income, marital status, and employment status. Some questions are answered by giving a number: the corresponding variables are *quantitative*. Age, family size, and family income are examples of quantitative variables. Some questions are answered with a descriptive word or phrase, and the corresponding variables are *qualitative*: examples are marital status (single, married, widowed,

divorced, separated) and employment status (employed, unemployed, not in the labor force).

Quantitative variables may be *discrete* or *continuous*. This is not a hard- and-fast distinction, but it is a useful one.[8] For a discrete variable, the values can only differ by fixed amounts. Family size is discrete. Two families can differ in size by 0 or 1 or 2, and so on. Nothing in between is possible. Age, on the other hand, is a continuous variable. This doesn't refer to the fact that a person is continuously getting older; it just means that the difference in age between two people can be arbitrarily small—a year, a month, a day, an hour, … Finally, the terms *qualitative*, *quantitative*, *discrete*, and *continuous* are also used to describe data—qualitative data are collected on a qualitative variable, and so on.



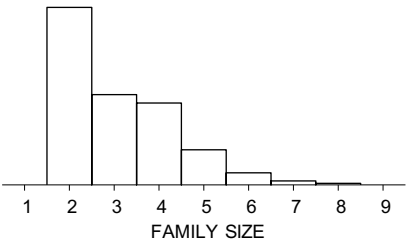
Section 2 showed how to plot a histogram starting with a distribution table. Often the starting point is the raw data—a list of cases (individuals, families, schools, etc.) and the corresponding values of the variable. In order to draw the histogram, a distribution table must be prepared. The first step is to choose the class intervals. With too many or too few, the histogram will not be informative. There is no rule, it is a matter of judgment or trial and error. It is common to start with ten or fifteen class intervals and work from there. In this book, the class intervals will always be given.[9]

When plotting a histogram for a continuous variable, investigators also have to decide on the endpoint convention—what to do with cases that fall right on the boundary. With a discrete variable, there is a convention which gets around this nuisance: center the class intervals at the possible values. For instance, family size can be 2 or 3 or 4, and so on. (The Census does not recognize one person as a family.) The corresponding class intervals in the distribution table would be

Center	*Class interval*
2	1.5 to 2.5
3	2.5 to 3.5
4	3.5 to 4.5
.	.
.	.
.	.

Since a family cannot have 2.5 members, there is no problem with endpoints.
Figure 6 (on the next page) shows the histogram for family size.
The bars seem to stop at 8;
that is because there are so few families with 9 or more people.

Figure 6. Histogram showing distribution of families by size in 2005. With a discrete variable, the class intervals are centered at the possible values.



Source: March 2005 Current Population Survey; CD-ROM supplied by the Bureau of the Census.

SUMMARY

1. A *histogram* represents percents by area. It consists of a set of blocks. The area of each block represents the percentage of cases in the corresponding *class interval*.

2. With the *density scale*, the height of each block equals the percentage of cases in the corresponding class interval, divided by the length of that interval.

3. With the density scale, area comes out in percent, and the total area is 100%. The area under the histogram between two values gives the percentage of cases falling in that interval.

4. A *variable* is a characteristic of the subjects in a study. It can be either *qualitative* or *quantitative*. A quantitative variable can be either *discrete* or *continuous*.

5. A confounding factor is sometimes controlled for by *cross-tabulation*.

2

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.

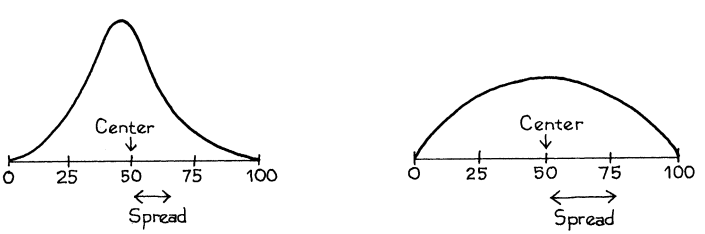
—SIR FRANCIS GALTON (ENGLAND, 1822–1911)[1]

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

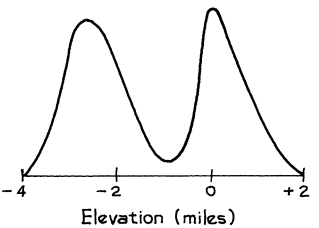
Figure 1. Center and spread. The centers of the two histograms are the same, but the second histogram is more spread out.



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.



1. 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.

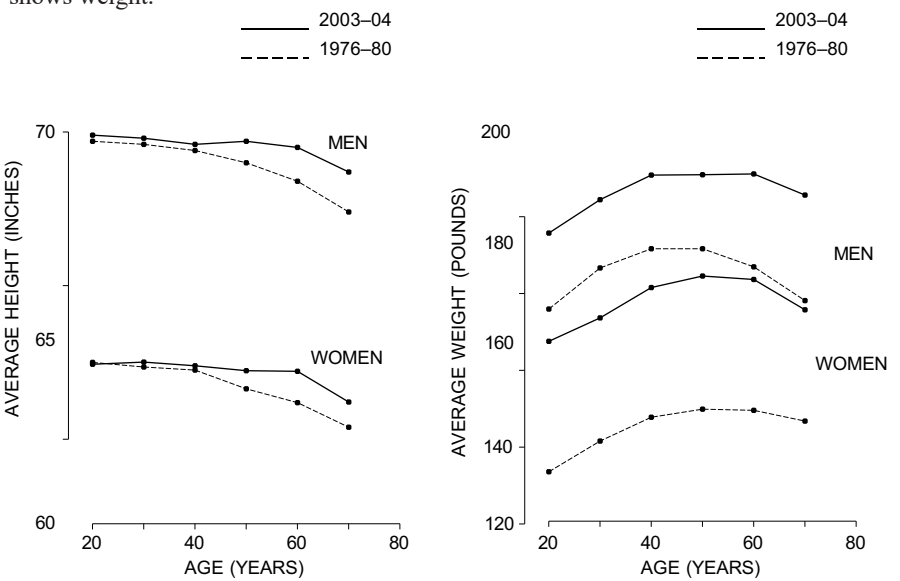
They're pretty chubby.

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.



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.

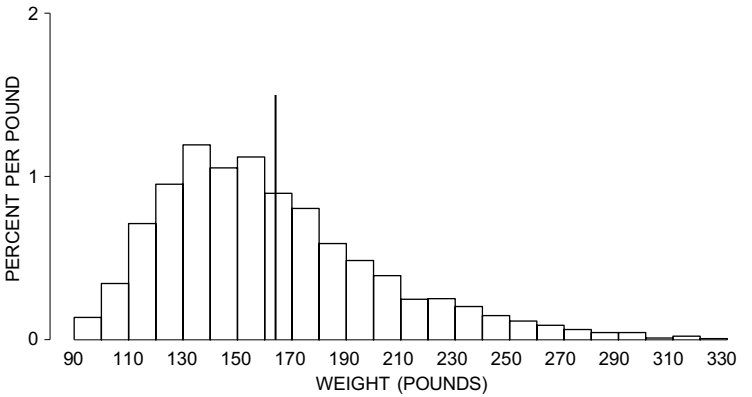
There is evidence to suggest that, over time, Americans have been getting taller. This is called the *secular trend* in height, and its effect is confounded with the effect of age in figure 3. Most of the two-inch drop in height seems to be due to the secular trend. The people age 65–74 were born around 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.

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.

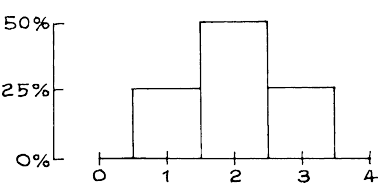


Source: www.cdc.gov/nchs/nhanes.htm.

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.)

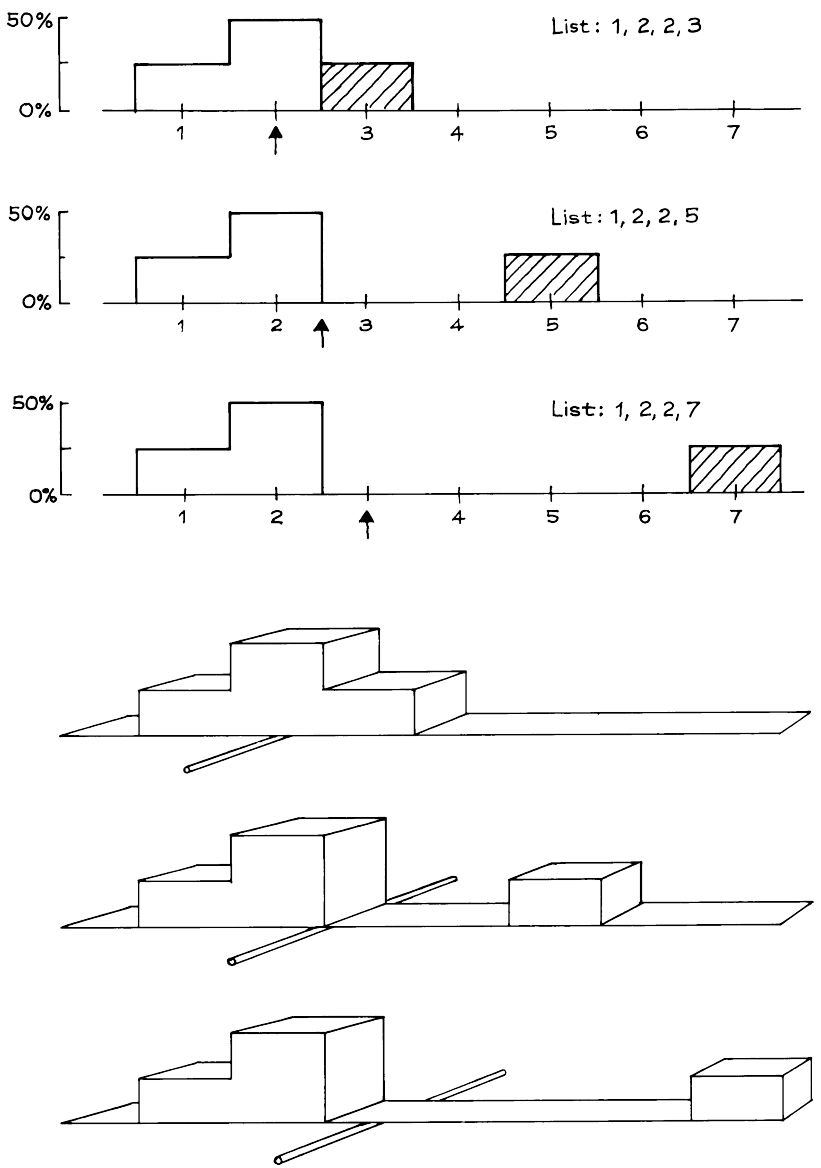
Figure 5. Histogram for the list 1, 2, 2, 3. The histogram is symmetric around 2, the average: 50% of the area is to the left of 2, and 50% is to the right.



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

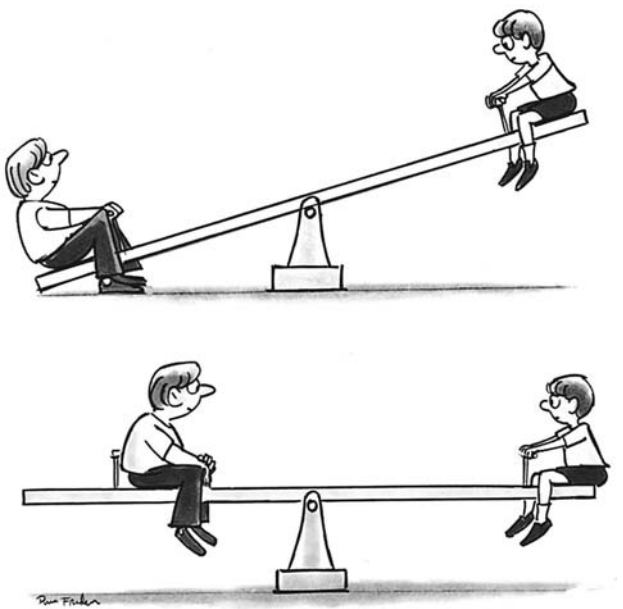
histogram will balance at the average.[7] A small area far away from the average can balance a large area close to the average, because areas are weighted by their distance from the balance point.

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.



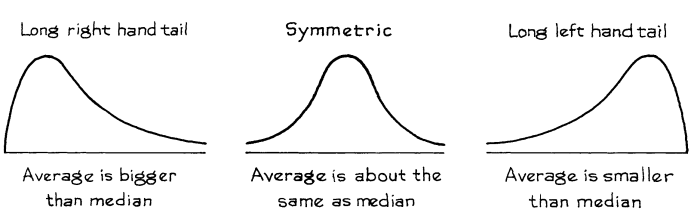
A histogram balances when supported at the average.

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 at- tention to the extreme tail of the distribution. We return to this point in the next chapter.

2. 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.

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

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 av- erage 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 av- erage. 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.

The SD of 3 inches says that many of the women differed from the average height by 1 or 2 or 3 inches: 1 inch is a third of an SD, and 3 inches is an SD. Few women differed from the average height by more than 6 inches (two SDs).

There is a rule of thumb which makes this idea more quantitative, and which applies to many data sets.

> Roughly 68% of the entries on a list (two in three) are within one SD of the average, the other 32% are further away. Roughly 95% (19 in 20) are within two SDs of the average, the other 5% are further away. This is so for many lists, but not all.

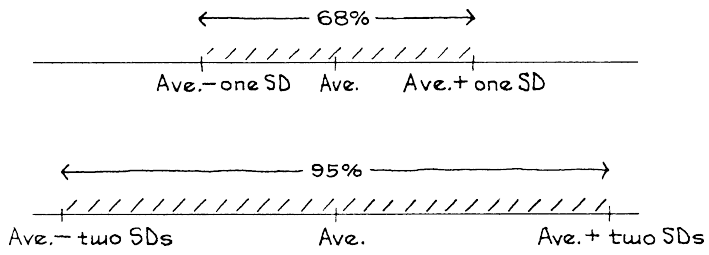


Figure 8 shows the histogram for the heights of women age 18 and over in HANES5. The average is marked by a vertical line, and the region within one SD of the average is shaded. This shaded area represents the women who differed from average height by one SD or less. The area is about 72%. About 72% of the women differed from the average height by one SD or less.

Figure 8. The SD and the histogram. Heights of 2,696 women age 18 and over in HANES5. The average of 63.5 inches is marked by a vertical line. The region within one SD of the average is shaded: 72% of the women differed from average by one SD (3 inches) or less.

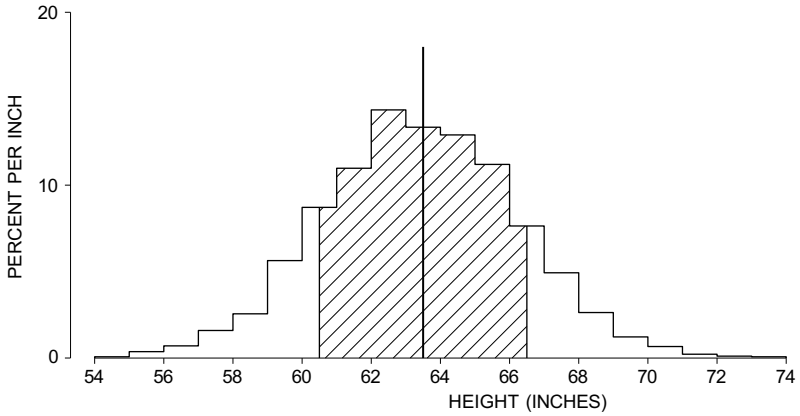
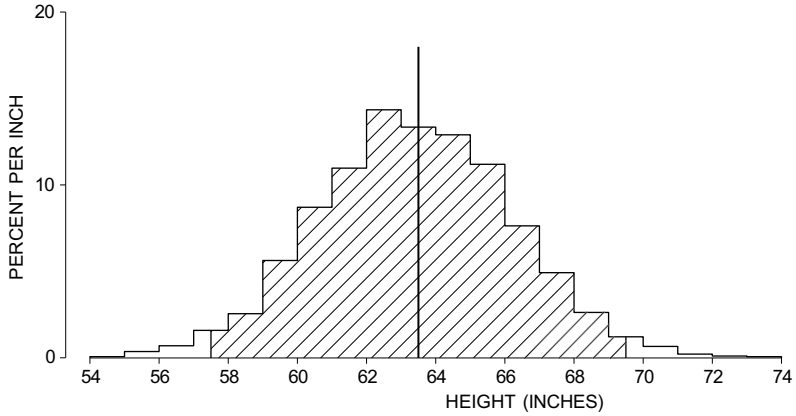


Figure 9 shows the same histogram. Now the area within two SDs of average is shaded. This shaded area represents the women who differed from average height by two SDs or less. The area is about 97%. About 97% of the women differed from the average height by two SDs or less.

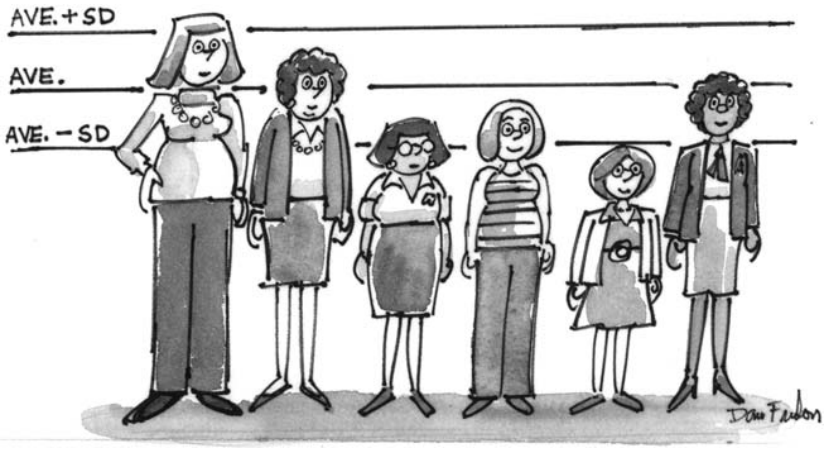
Figure 9. The SD and the histogram. Heights of 2,696 women age 18 and over in HANES5. The average of 63.5 inches is marked by a vertical line. The region within two SDs of the average is shaded: 97% of the women differed from average by two SDs (6 inches) or less.



To sum up, about 72% of the women differed from average by one SD or less, and 97% differed

from average by two SDs or less. There was only one woman in the sample who was more than three SDs away from the average, and none were more than four SDs away. For this data set, the 68%–95% rule works quite well. Where do the 68% and 95% come from?

About two-thirds of the HANES women differed from the average by less than one SD.



COMPUTING THE STANDARD DEVIATION

To find the standard deviation of a list, take the entries one at a time. Each deviates from the average by some amount, perhaps 0:

deviation from average = entry − average.

The SD is the r.m.s. size of these deviations. (Reminder: "r.m.s." means root- mean-square. See p. 66.)

SD = r.m.s. deviation from average.