3

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.

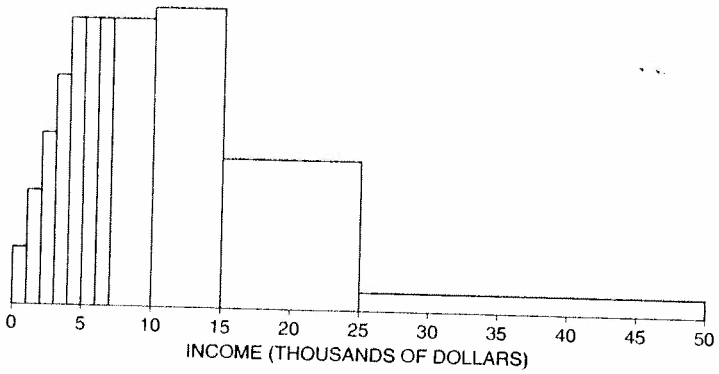
—*The Little Prince*[1]

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.

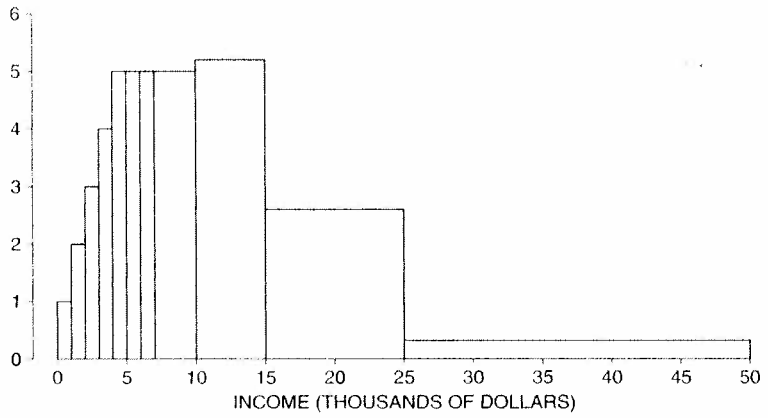
For a last example, take the percentage of families with incomes under $7,000. Is this closest to 10%, 25%, or 50%? By eye, the area under the histogram between $0 and $7,000 is about a quarter of the total area, so the percentage is closest to 25%.

> 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.

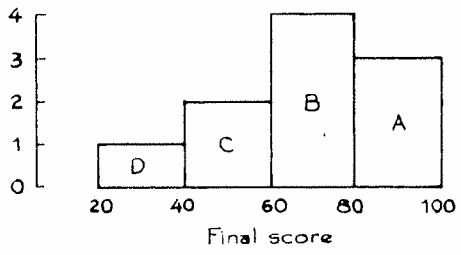


Exercise Set A

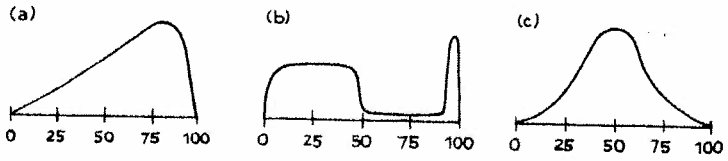
1. About 1% of the families in figure 2 had incomes between \$0 and \$1,000. Estimate the percentage who had incomes—
   (a) between \$1,000 and \$2,000
   (b) between \$2,000 and \$3,000
   (c) between \$3,000 and \$4,000
   (d) between \$4,000 and \$5,000
   (e) between \$4,000 and \$7,000
   (f) between \$7,000 and \$10,000

2. In figure 2, were there more families earning between \$10,000 and \$11,000 or between \$15,000 and \$16,000? Or were the numbers about the same? Make your best guess.

3. The histogram below shows the distribution of final scores in a certain class.
   (a) Which block represents the people who scored between 60 and 80?
   (b) Ten percent scored between 20 and 40. About what percentage scored between 40 and 60?
   (c) About what percentage scored over 60?



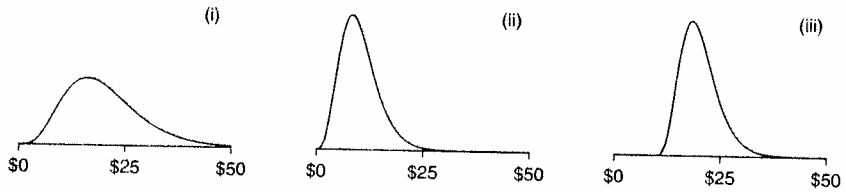
4. Below are sketches of histograms for test scores in three different classes. The scores range from 0 to 100; a passing score was 50. For each class, was the percent who passed about 50%, well over 50%, or well under 50%?



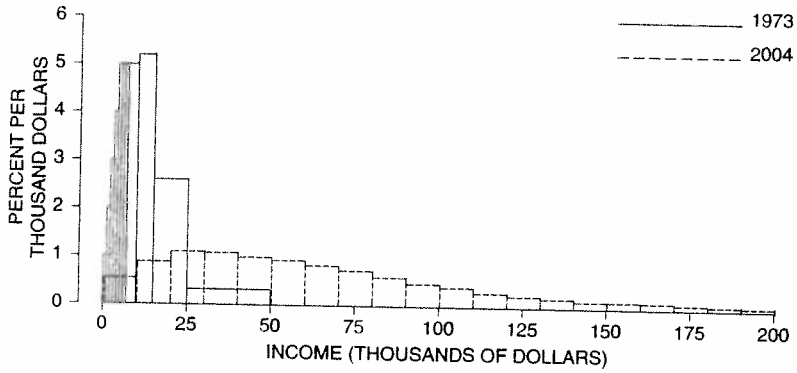
5. One class in exercise 4 had two quite distinct groups of students, with one group doing rather poorly on the test, and the other group doing very well. Which class was it?

6. In class (b) of exercise 4, were there more people with scores in the range 40–50 or 90–100?

7. An investigator collects data on hourly wage rates for three groups of people. Those in group B earn about twice as much as those in group A. Those in group C earn about $10 an hour more than those in group A. Which histogram belongs to which group? (The histograms don't show wages above $50 an hour.)



8. The figure below compares the histograms for family incomes in the U.S. in 1973 and in 2004. It looks as if family income went up by a factor of 4 over 30 years. Or did it? Discuss briefly.



Source: Current Population Survey.[3]

The answers to these exercises are on pp. A45–46.

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.

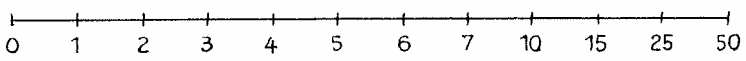
The computer has to be told what to do with families that fall right on the boundary between two class intervals. This is called an *endpoint convention*. The convention followed in table 1 is indicated by the caption. The left endpoint is included in the class interval, the right endpoint is excluded. 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.

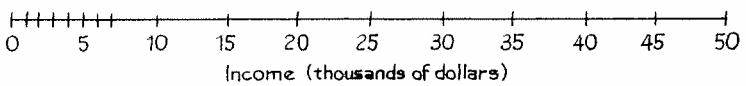
Income level	*Percent*
\$0–\$1,000	1
\$1,000–\$2,000	2
\$2,000–\$3,000	3
\$3,000–\$4,000	4
\$4,000–\$5,000	5
\$5,000–\$6,000	5
\$6,000–\$7,000	5
\$7,000–\$10,000	15
\$10,000–\$15,000	26
\$15,000–\$25,000	26
\$25,000–\$50,000	8
\$50,000 and over	1

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

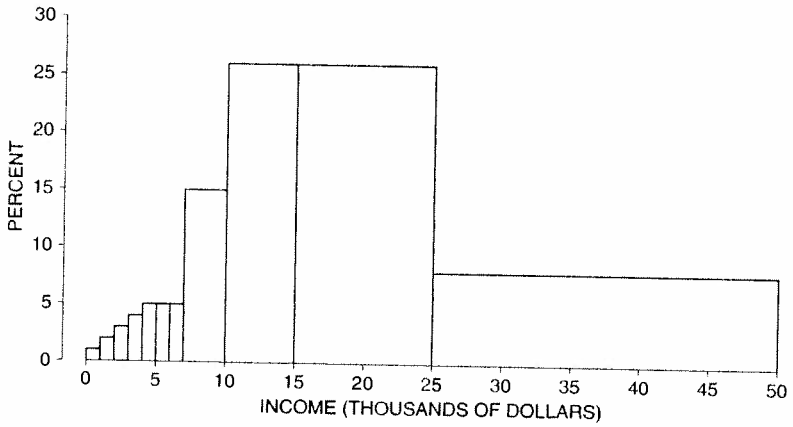


That is a mistake. The interval from \$7,000 to \$10,000 is three times as long as the interval from \$6,000 to \$7,000. So the horizontal axis should look like this:



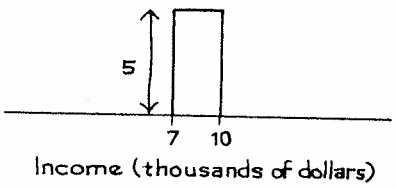
The next step is to draw the blocks. It's tempting to make their heights equal to the percents in the table. Figure 3 shows what happens if you make that mistake. The graph gives much too rosy a picture of the income distribution. For example, figure 3 says there were many more families with incomes 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.

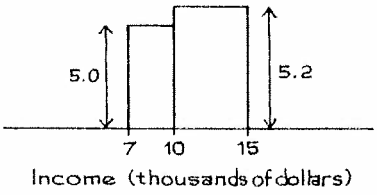


The source of the trouble is that some class intervals are longer than others, so the percents in table 1 are not on a par with one another. The 8% who earn $25,000 to $50,000, for instance, are spread over a much larger range of incomes than the 15% who earn $7,000 to $10,000. Plotting percents directly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

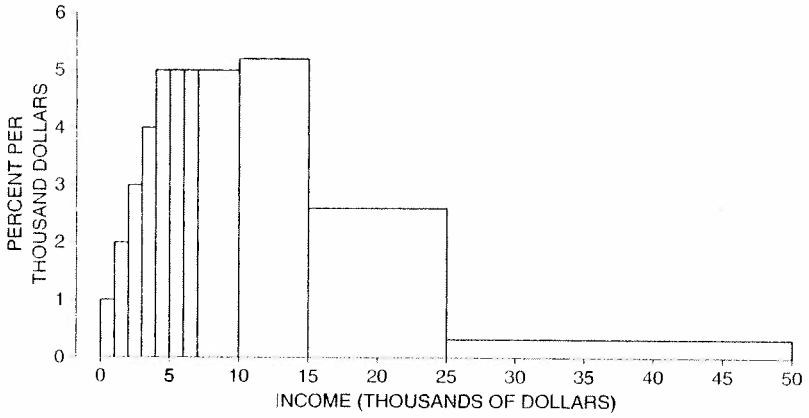


The work is done for two of the lines in table 1. To complete the histogram, do the same thing for the rest of the class intervals. Figure 4 (below) is the result.

> 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.

Exercise Set B

1. The table below gives the distribution of educational level for persons age 25 and over in the U.S. in 1960, 1970, and 1991. ("Educational level" means the number of years of schooling completed.) The class intervals include the left endpoint, but not the right; for example, from the second line of the table, in 1960 about 14% of the people had completed 5–8 years of schooling, 8 not included; in 1991, about 4% of the people were in this category. Draw a histogram for the 1991 data. You can interpret "16 or more" as 16–17 years of schooling; not many people completed more than 16 years of school, especially in 1960 and 1970. Why does your histogram have spikes at 8, 12, and 16 years of schooling?

Educational level (years of schooling)	*1960*	*1970*	*1991*
0–5	8	6	2
5–8	14	10	4
8–9	18	13	4
9–12	19	19	11
12–13	25	31	39
13–16	9	11	18
16 or more	8	11	21

Source: Statistical Abstract, 1988, Table 202; 1992, Table 220.

2. Redraw the histogram for the 1991 data, combining the first two class intervals into one (0–8 years, with 6% of the people). Does this change the histogram much?
3. Draw the histogram for the 1970 data, and compare it to the 1991 histogram. What happened to the educational level of the population between 1970 and 1991—did it go up, go down, or stay about the same?
4. What happened to the educational level from 1960 to 1970?

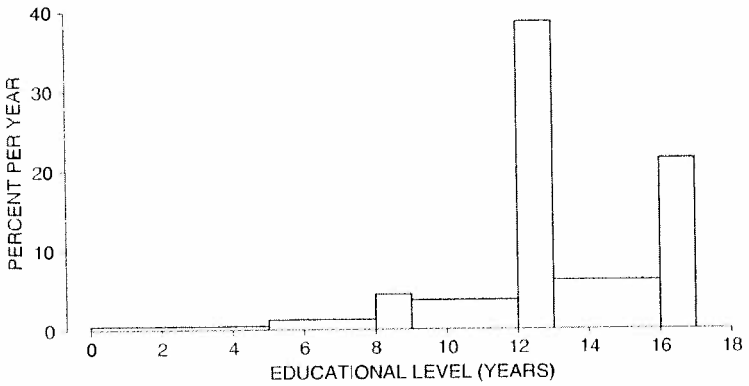
The answers to these exercises are on p. A46.

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. Imagine 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.

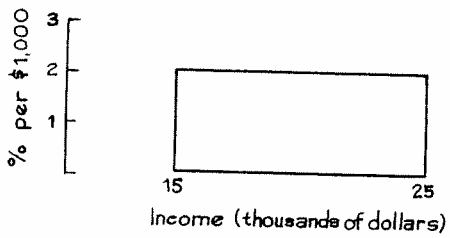
At first, it may be difficult to keep apart the notion of the crowding in an interval, represented by the height of the block, and the number in an interval, represented by the area of the block. An example will help. Look at the blocks over the intervals 8–9 years and 9–12 years in figure 5. The first 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?

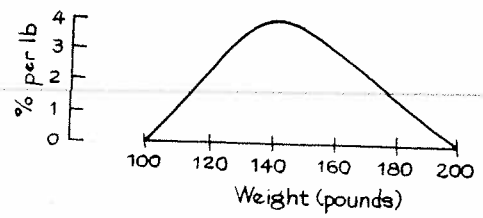


Solution. The height of the block is 2% per thousand dollars. Each thousand-dollar interval between \$15,000 and \$25,000 contains about 2% of the families in the city. There are 10 of these thousand-dollar intervals between \$15,000 and \$25,000. The answer is $10 \times 2\% = 20\%$. About 20% of the 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

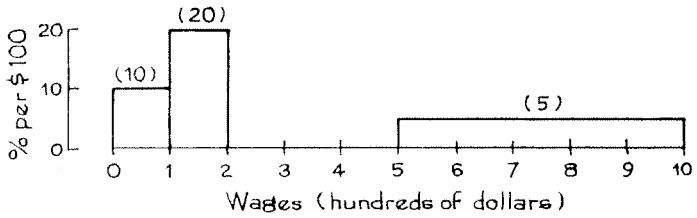
$$\frac{1}{2} \times \text{base} \times \text{height} = \frac{1}{2} \times 100 \text{ lb} \times 4\% \text{ per lb} = 200\%.$$

> 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%.

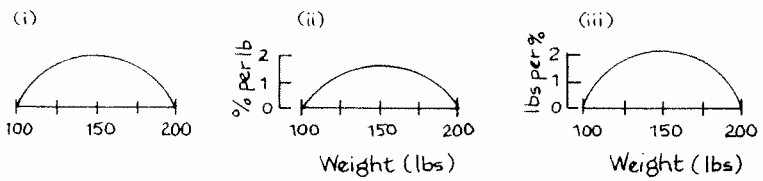
Since 1991, the educational level in the U.S. has continued to increase. Then, 21% of the population had a bachelor's degree or better (the "population" means people age 25 and over). In 2005, the corresponding figure was 28%.

Exercise Set C

1. A histogram of monthly wages for part-time employees is shown below (densities are marked in parentheses). Nobody earned more than $1,000 a month. The block over the class interval from $200 to $500 is missing. How tall must it be?



2. Three people plot histograms for the weights of subjects in a study, using the density scale. Only one is right. Which one, and why?



3. An investigator draws a histogram for some height data, using the metric system. She is working in centimeters (cm). The vertical axis shows density, and the top of the vertical axis is 10 percent per cm. Now she wants to convert to millimeters (mm). There are 10 millimeters to the centimeter. On the horizontal axis, she has to change 175 cm to _____ mm, and 200 cm to _____ mm. On the vertical axis, she has to change 10 percent per cm to _____ percent per mm, and 5 percent per cm to _____ percent per mm.

4. In a Public Health Service study, a histogram was plotted showing the number of cigarettes per day smoked by each subject (male current smokers), as shown below.[7] The density is marked in parentheses. The class intervals include the right endpoint, not the left.

   (a) The percentage who smoked 10 cigarettes or less per day is around

   1.5% 15% 30% 50%

   (b) The percentage who smoked more than a pack a day, but not more than 2 packs, is around

   1.5% 15% 30% 50%

   (There are 20 cigarettes in a pack.)

   (c) The percent who smoked more than a pack a day is around

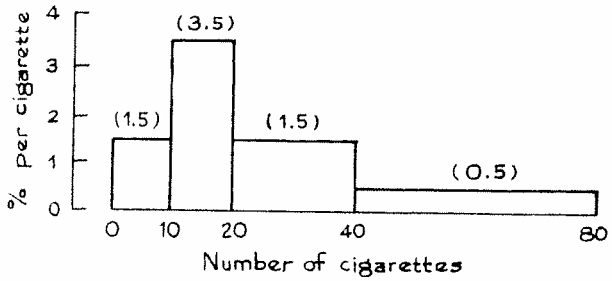
   1.5% 15% 30% 50%

   (d) The percent who smoked more than 3 packs a day is around

   0.25 of 1% 0.5 of 1% 10%

   (e) The percent who smoked 15 cigarettes per day is around

   0.35 of 1% 0.5 of 1% 1.5% 3.5% 10%



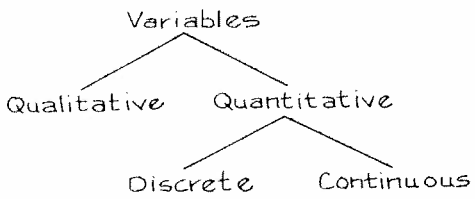
The answers to these exercises are on p. A46.

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, family 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.



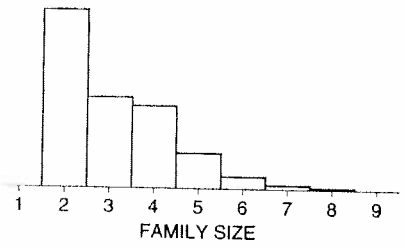
Section 2 showed how to plot a histogram starting with a distribution table. Often the starting point is the raw data—a list of cases (individuals, families, schools, etc.) and the corresponding values of the variable. In order to draw the histogram, a distribution table must be prepared. The first step 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.

Exercise Set D

1. Classify each of the following variables as qualitative or quantitative; if quantitative, as discrete or continuous.
   (a) occupation (b) region of residence (c) weight
   (d) height (e) number of automobiles owned

2. In the March Current Population Survey, women are asked how many children they have. Results are shown below for women age 25–39, by educational level.
   (a) Is the number of children discrete or continuous?
   (b) Draw histograms for these data. (You may take "5 or more" as 5—very few women had more than 5 children.)
   (c) What do you conclude?

Distribution of women age 25–39 by educational level and number of children (percent).

Number of children	*Women who are high-school graduates*	*Women with college degrees*
0	30.2	47.9
1	21.8	19.4
2	28.4	22.7
3	13.7	8.0
4	4.4	1.5
5 or more	1.5	0.5

Note: High-school graduates with no further education. College degrees at the level of a B.A. or B.Sc. Own, never-married children under the age of 18. Percents may not add to 100%, due to rounding.
Source: March 2005 Current Population Survey; CD-ROM supplied by the Bureau of the Census.

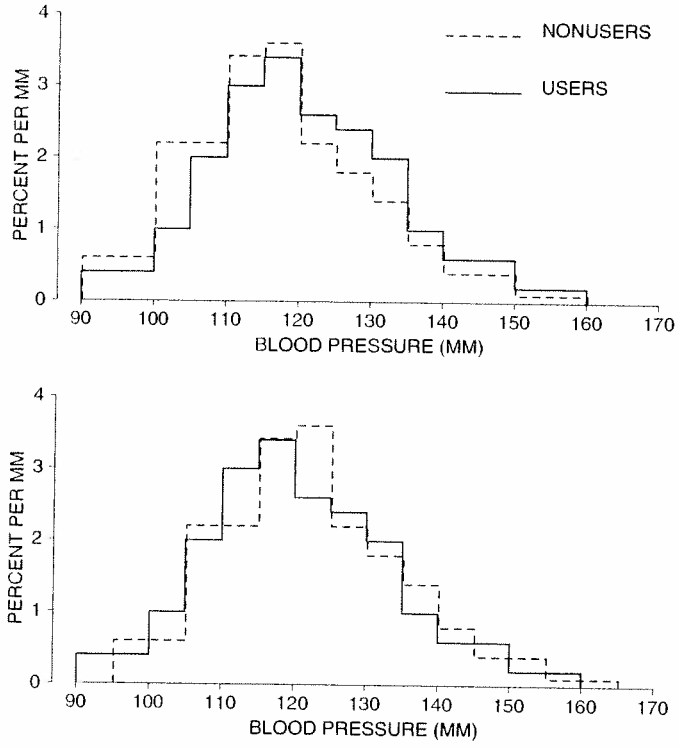
The answers to these exercises are on p. A47.

5. CONTROLLING FOR A VARIABLE

In the 1960s, many women began using oral contraceptives, "the pill." Since the pill alters the body's hormone balance, it is important to see what the side effects are. Research on this question is carried out by the Contraceptive Drug Study at the Kaiser Clinic in Walnut Creek, California. Over 20,000 women in the Walnut Creek area belong to the Kaiser Foundation Health Plan, paying a monthly insurance fee and getting medical services from Kaiser. One of these services is a routine checkup called the "multiphasic." During the period 1969–1971, about 17,500 women age 17–58 took the multiphasic and became subjects for the Drug Study. Investigators compared the multiphasic results for two different groups of women:

- "users" who take the pill (the treatment group);
- "non-users" who don't take the pill (the control group).

Figure 7. The effect of the pill. The top panel shows histograms for the systolic blood pressures of the 1,747 users and the 3,040 non-users age 25–34 in the Contraceptive Drug Study. The bottom panel shows the histogram for the non-users shifted to the right by 5 mm.



This is an observational study. It is the women who decided whether to take the pill or not. The investigators just watched what happens.

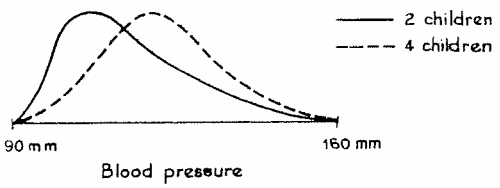
One issue was the effect of the pill on blood pressure. It might seem natural to compare the blood pressures for the users and non-users. However, this could be misleading. Blood pressure tends to go up with age, and the non-users were on the whole older than the users. For example, about 70% of the non-users were over 30, compared to 50% of the users. The effect of age is confounded with the effect of the pill. To make the full effect of the pill visible, it is necessary to make a separate comparison for each age group: this controls for age.[10] We will look only at the women age 25–34. Figure 7 shows the histograms for the users and non-users in this age group. (Blood pressure is measured relative to the length of a column of mercury; the units are "mm," that is, millimeters.)

The two histograms in the top panel of figure 7 have very similar shapes. However, the user histogram is higher to the right of 120 mm, lower to the left. High blood pressure (above 120 mm) is more prevalent among users, low blood pressure less prevalent. Now imagine that 5 mm were added to the blood pressure of each non-user. That would shift their histogram 5 mm to the right, as shown in the bottom panel of figure 7. In the bottom panel, the two histograms match up quite well. As far as the histograms are concerned, it is as if using the pill adds about 5 mm to the blood pressure of each woman.

This conclusion must be treated with caution. The results of the Contraceptive Drug Study suggest that if a woman goes on the pill, her blood pressure will go up by around 5 mm. But the proof is not complete. It cannot be, because of the design. The Drug Study is an observational study, not a controlled experiment. Part I showed that observational studies can be misleading about cause-and-effect relationships. There could be some factor other than the pill or age, as yet unidentified, which is affecting the blood pressures. For the Drug Study, this is a bit farfetched. The physiological mechanism by which the pill affects blood pressure is well established. The Drug Study data show the size of the effect.

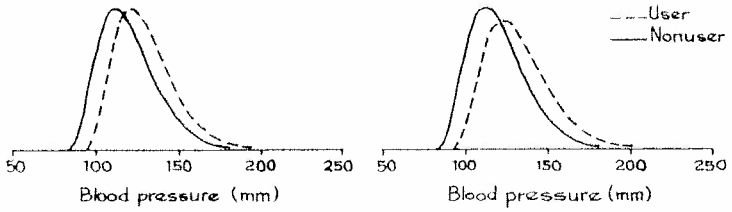
Exercise Set E

1. As a sideline, the Drug Study compared blood pressures for women having different numbers of children. Below are sketches of the histograms for women with 2 or 4 children. Which group has higher blood pressure? Does having children cause the blood pressures of the mothers to change? Or could the change be due to some other factor, whose effects are confounded with the effect of having children?



2. (Hypothetical.) The sketches on the next page show results from two other studies

of the pill, for women age 25–29. In one study, the pill adds about 10 mm to blood pressures; in the other, the pill adds about 10%. Which is which, and why?



The answers to these exercises are on p. A47.

6. CROSS-TABULATION

The previous section explained how to control for the effect of age: it was a matter of doing the comparison separately for each age group. The comparison was made graphically, through the histograms in figure 7. Some investigators prefer to make the comparison in tabular form, using what is called a *cross-tab* (short for *cross-tabulation*). A cross-tab for blood pressure by age and pill use is shown in table 2. Such tables are a bit imposing, and the eye naturally tends to skip over

Table 2. Systolic blood pressure by age and pill use, for women in the Contraceptive Drug Study, excluding those who were pregnant or taking hormonal medication other than the pill. Class intervals include the left endpoint, but not the right. – means negligible. Table entries are in percent; columns may not add to 100 due to rounding.

Blood pressure (millimeters)	*Age 17–24*		*Age 25–34*		*Age 35–44*		*Age 45–58*	
	Non-users	*Users*	*Non-users*	*Users*	*Non-users*	*Users*	*Non-users*	*Users*
	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
under 90	–	1	1	–	1	1	1	–
90–95	1	–	1	–	2	1	1	1
95–100	3	1	5	4	5	4	4	2
100–105	10	6	11	5	9	5	6	4
105–110	11	9	11	10	11	7	7	7
110–115	15	12	17	15	15	12	11	10
115–120	20	16	18	17	16	14	12	9
120–125	13	14	11	13	9	11	9	8
125–130	10	14	9	12	10	11	11	11
130–135	8	12	7	10	8	10	10	9
135–140	4	6	4	5	5	7	8	8
140–145	3	4	2	4	4	6	7	9
145–150	2	2	2	2	2	5	7	9
150–155	–	1	1	1	1	3	2	4
155–160	–	–	–	1	1	1	1	3
160 and over	–	–	–	–	1	2	2	5
Total percent	100	98	100	99	100	100	99	99
Total number	1,206	1,024	3,040	1,747	3,494	1,028	2,172	437

them until some of the numbers are needed. However, all the cross-tab amounts to is a distribution table for blood pressures, made separately for users and non-users in each age group.

Look at the columns for the age group 17–24. There were 1,206 non-users and 1,024 users. About 1% of the users had blood pressure below 90 mm; the corresponding percentage of non-users was negligible—that is what the dash means. To see the effect of the pill on the blood pressures of women age 17–24, it is a matter of looking at the percents in the columns for non-users and users in the age group 17–24. To see the effect of age, look first at the non-users column in each age group and see how the percents shift toward the high blood pressures as age goes up. Then do the same thing for the users.

Exercise Set F

1. Use table 2 to answer the following questions.
   (a) What percentage of users age 17–24 have blood pressures of 140 mm or more?
   (b) What percentage of non-users age 17–24 have blood pressures of 140 mm or more?
   (c) What do you conclude?
2. Draw histograms for the blood pressures of the users and non-users age 17–24. What do you conclude?
3. Compare the histograms of blood pressures for non-users age 17–24 and for non-users age 25–34. What do you conclude?

The answers to these exercises are on p. A47.

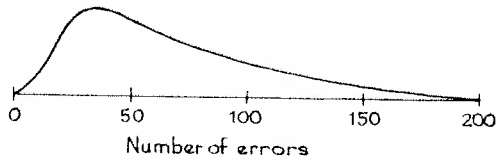
7. SELECTIVE BREEDING

In 1927, the psychologist Charles Spearman published *The Abilities of Man*, his theory of human intelligence. Briefly, Spearman held that test scores of intellectual abilities (like reading comprehension, arithmetic, or spatial perception) were weighted sums of two independent components: a general intelligence factor which Spearman called "g," and an ability factor specific to each test. This theory attracted a great deal of attention.

As part of his Ph.D. research in the psychology department at Berkeley, Robert Tryon decided to check the theory on an animal population, where it is simpler to control extraneous variables.[11] Tryon used rats, which are easy to breed in the laboratory. To test their intelligence, he put the rats into a maze. When they ran the maze, the rats made errors by going into blind alleys. The test consisted of 19 runs through the maze; the animal's "intelligence score" was the total number

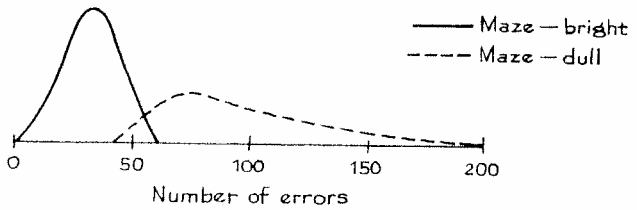
of errors it made. So the bright rats are the ones with low scores, the dulls are the ones with high scores. Tryon started out with 142 rats, and the distribution of their intelligence scores is sketched in figure 8.

Figure 8. Tryon's experiment. Distribution of intelligence in the original population.



The next step in the experiment was to breed for intelligence. In each generation, the "maze-bright" rats (the ones making only a small number of errors) were bred with each other. Similarly, the "maze-dull" animals (with high scores) were bred together. Seven generations later, Tryon had 85 rats in the maze-bright strain, and 68 in the maze-dull strain. There was a clear separation in scores. Figure 9 shows the distribution of intelligence for the two groups, and the histograms barely overlap. (In fact, Tryon went on with selective breeding past the seventh generation, but didn't get much more separation in scores.)

Figure 9. Tryon's experiment. After seven generations of selective breeding, there is a clear separation into "maze-bright" and "maze-dull" strains.

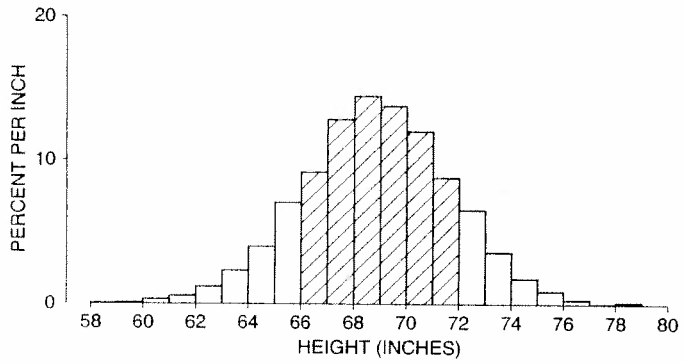


The two strains created by Tryon were used for many other experiments by the Berkeley psychology department. Generations later, rats from the maze-bright population continued to outperform the dulls at maze-running. So Tryon managed to breed for a mental ability—evidence that some mental abilities are at least in part genetically determined. What did the experiment say about Spearman's theory? Tryon found that the maze-bright rats did no better than the maze-dulls on other tests of animal intelligence, such as discriminating between geometric shapes, or between intensities of light. This was evidence against Spearman's theory of a general intelligence factor (at least for rats). On the other hand, Tryon did find intriguing psychological differences between the two rat populations. The "brights" seemed to be unsociable introverts, well adjusted to life in the maze, but neurotic in their relationships with other rats. The "dulls" were quite the opposite.

8. REVIEW EXERCISES

Review exercises may cover material from previous chapters.

1. The figure below shows a histogram for the heights of a representative sample of men. The shaded area represents the percentage of men whose heights were between _____ and _____. Fill in the blanks.



Source: Data tape supplied by the Inter-University Consortium for Political and Social Research.

2. The age distribution of people in the U.S. in 2004 is shown below. Draw the histogram. (The class intervals include the left endpoint, not the right; for instance, on the second line of the table, 14% of the people were age 5 years or more but had not yet turned 15. The interval for "75 and over" can be ended at 85. Men and women are combined in the data.) Use your histogram to answer the following questions.

   (a) Are there more children age 1, or elders age 71?
   (b) Are there more 21-year-olds, or 61-year-olds?
   (c) Are there more people age 0–4, or 65–69?
   (d) The percentage of people age 35 and over is around 25%, 50%, or 75%?

Age	*Percent of population*	*Age*	*Percent of population*
0–5	7	35–45	15
5–15	14	45–55	14
15–20	7	55–65	10
20–25	7	65–75	6
25–30	7	75 and over	6
30–35	7		

Source: *Statistical Abstract*, 2006, Table 11.

3. The American Housing Survey is done every year by the Bureau of the Census. Data from the 2003 survey can be used to find the distribution of occupied housing units (this includes apartments) by number of rooms. Results for the whole U.S. are shown below, separately for "owner-occupied" and "renter-

occupied" units. Draw a histogram for each of the two distributions. (You may assume that "10 or more" means 10 or 11; very few units have more than 11 rooms.)

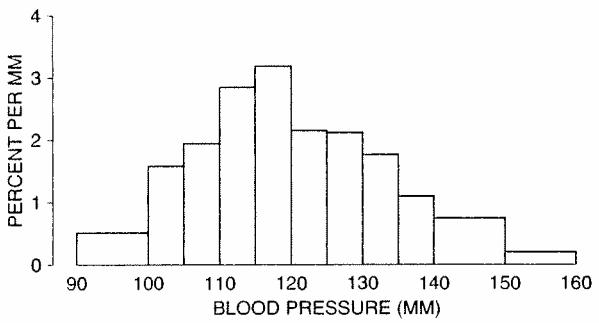
(a) The owner-occupied percents add up to 100.2% while the renter-occupied percents add up to 100.0%. Why?
(b) The percentage of one-room units is much smaller for owner-occupied housing. Is that because there are so many more owner-occupied units in total? Answer yes or no, and explain briefly.
(c) Which are larger, on the whole: the owner-occupied units or the renter-occupied units?

Number of rooms in unit	*Owner-occupied (percent)*	*Renter-occupied (percent)*
1	0.0	1.0
2	0.1	2.8
3	1.4	22.7
4	9.7	34.5
5	23.3	22.6
6	26.4	10.4
7	17.5	3.6
8	10.4	1.2
9	5.0	0.5
10 or more	6.4	0.7
Total	100.2	100.0
Number	72.2 million	33.6 million

Source: www.census.gov/hhres/www/housing/ahs/nationaldata.html

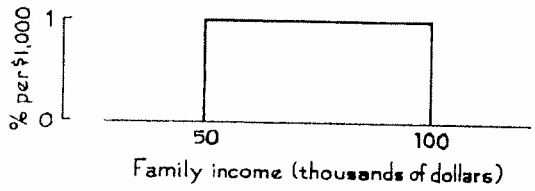
4. The figure below is a histogram showing the distribution of blood pressure for all 14,148 women in the Drug Study (section 5). Use the histogram to answer the following questions:

(a) Is the percentage of women with blood pressures above 130 mm around 25%, 50%, or 75%?
(b) Is the percentage of women with blood pressures between 90 mm and 160 mm around 1%, 50%, or 99%?
(c) In which interval are there more women: 135–140 mm or 140–150 mm?



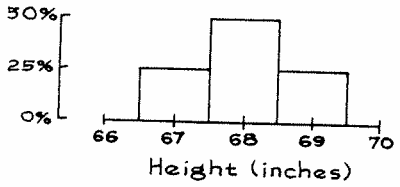
(d) Which interval is more crowded: 135–140 mm or 140–150 mm?
(e) On the interval 125–130 mm, the height of the histogram is about 2.1% per mm. What percentage of the women had blood pressures in this class interval?
(f) Which interval has more women: 97–98 mm or 102–103 mm?
(g) Which is the most crowded millimeter of all?

5. Someone has sketched one block of a family-income histogram for a wealthy suburb. About what percentage of the families in this suburb had incomes between $90,000 and $100,000 a year?

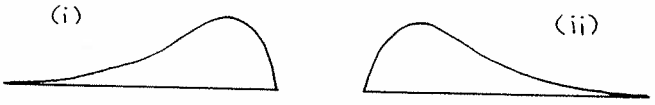


6. (Hypothetical.) In one study, 100 people had their heights measured to the nearest eighth of an inch. A histogram for the results is shown below. Two of the following lists have this histogram. Which ones, and why?

(i) 25 people, 67 inches tall; 50 people, 68 inches tall; 25 people, 69 inches tall.
(ii) 10 people, $66\frac{3}{4}$ inches tall; 15 people, $67\frac{1}{4}$ inches tall; 50 people, 68 inches tall; 25 people, 69 inches tall.
(iii) 30 people, 67 inches tall; 40 people, 68 inches tall; 30 people, 69 inches tall.



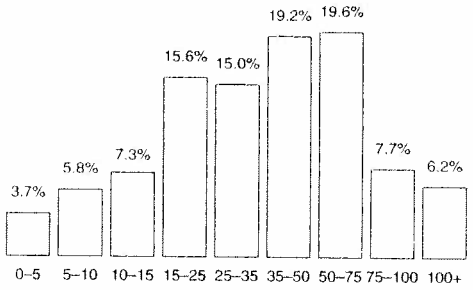
7. Two histograms are sketched below. One shows the distribution of age at death from natural causes (heart disease, cancer, and so forth). The other shows age at death from trauma (accident, murder, suicide). Which is which, and why?



8. The figure on the next page (adapted from the *San Francisco Chronicle*, May 18, 1992) shows the distribution of American families by income. Ranges include

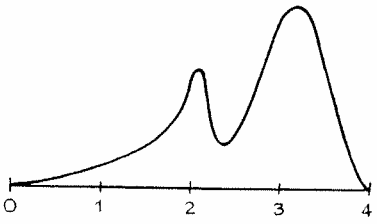
the left endpoint but not the right. For example, 3.7% of the families had incomes in the range \$0–\$4,999, 5.8% had incomes in the range \$5,000–\$9,999, and so forth. True or false, and explain:

(a) Although American families are not spread evenly over the whole income range, the families that earn between \$10,000 and \$35,000 are spread fairly evenly over that range.
(b) The families that earn between \$35,000 and \$75,000 are spread fairly evenly over that range.
(c) The graph is a histogram.



9. In a survey carried out at the University of California, Berkeley, a sample of students were interviewed and asked what their grade-point average was. A histogram of the results is shown below. (GPA ranges from 0 to 4, and 2 is a bare pass.)

(a) True or false: more students reported a GPA in the range 2.0 to 2.1 than in the range 1.5 to 1.6.
(b) True or false: more students reported a GPA in the range 2.0 to 2.1 than in the range 2.5 to 2.6.
(c) What accounts for the spike at 2?



10. The table on the next page shows the distribution of adults by the last digit of their age, as reported in the Census of 1880 and the Census of 1970.[12] You might expect each of the ten possible digits to turn up for 10% of the people, but this is not the case. For example, in 1880, 16.8% of all persons reported an age ending in 0—like 30 or 40 or 50. In 1970, this percentage was only 10.6%.

(a) Draw histograms for these two distributions.

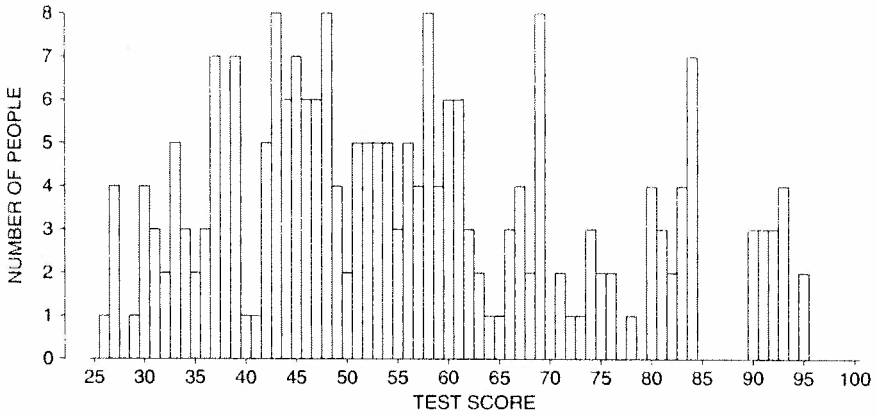
(b) In 1880, there was a strong preference for the digits 0 and 5. How can this be explained?
(c) In 1970, the preference was much weaker. How can this be explained?
(d) Are even digits more popular, or odd ones, in 1880? 1970?

Digit	*1880*	*1970*
0	16.8	10.6
1	6.7	9.9
2	9.4	10.0
3	8.6	9.6
4	8.8	9.8
5	13.4	10.0
6	9.4	9.9
7	8.5	10.2
8	10.2	10.0
9	8.2	10.1

Source: United States Census.

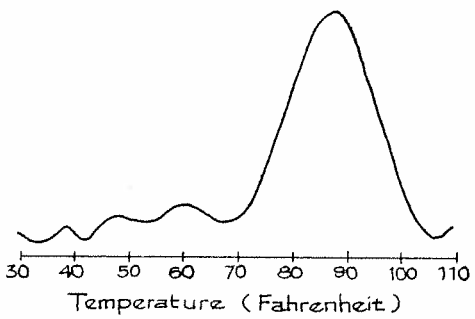
11. In the Sanitary District of Chicago, operating engineers are hired on the basis of a competitive civil-service examination. In 1966, there were 223 applicants for 15 jobs. The exam was held on March 12; the test scores are shown below, arranged in increasing order. The height of each bar in the histogram (top of next page) shows the number of people with the corresponding score. The examiners were charged with rigging the exam.[13] Why?

26	27	27	27	27	29	30	30	30	30	31	31	31	32	32
33	33	33	33	33	34	34	34	35	35	36	36	36	37	37
37	37	37	37	37	39	39	39	39	39	39	39	40	41	42
42	42	42	42	43	43	43	43	43	43	43	43	44	44	44
44	44	44	45	45	45	45	45	45	45	46	46	46	46	46
46	47	47	47	47	47	47	48	48	48	48	48	48	48	48
49	49	49	49	50	50	51	51	51	51	51	52	52	52	52
52	53	53	53	53	53	54	54	54	54	54	55	55	55	56
56	56	56	56	57	57	57	57	58	58	58	58	58	58	58
58	59	59	59	59	60	60	60	60	60	60	61	61	61	61
61	61	62	62	62	63	63	64	65	66	66	66	67	67	67
67	68	68	69	69	69	69	69	69	69	69	71	71	72	73
74	74	74	75	75	76	76	78	80	80	80	80	81	81	81
82	82	83	83	83	83	84	84	84	84	84	84	84	90	90
90	91	91	91	92	92	92	93	93	93	93	95	95		



12. The late 1960s and early 1970s were years of turmoil in the U.S. Psychologists thought that rioting was related (among other things) to temperature, with hotter weather making people more aggressive.[14] Two investigators, however, argued that "the frequency of riots should increase with temperature through the mid-80s but then go down sharply with increases in temperature beyond this level."

To support their theory, they collected data on 102 riots over the period 1967–71, including the temperature in the city where the riot took place. They plotted a histogram for the distribution of riots by temperature (a sketch is shown below). There is a definite peak around 85°. True or false, and explain: the histogram shows that higher temperatures prevent riots.



9. 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*.