PART I

Organization and description of data

CHAPTER 1

The organization of data

CHAPTER OUTLINE

A. THE MEANING OF DATA

The word "data" appears in many contexts and frequently is used in ordinary conversation. Although the word carries something of an aura of scientific mystique, its meaning is quite simple and mundane. It is Latin for "those that are given" (the singular form is "datum"). Data may therefore be thought of as the *results of observation*. Data are collected in many aspects of everyday life. Statements given to a police officer or physician or psychologist during an interview are data. So are the correct and incorrect answers given by a student on a final examination. Almost any athletic event produces data: the time required by a runner to complete a marathon, the number of errors committed by a baseball team in nine innings of play, the number of shots on goal during a period of hockey. And, of course, data are obtained in the course of scientific inquiry: the positions of artifacts and fossils in an archaeological site, the number of interactions between two members of an animal colony during a period of observation, or the spectral composition of light emitted by a star.

1. Quantitative data

Data may be used in a variety of ways to reach conclusions or generate interpretations. For example, the symptoms reported to a physician or clinical psychologist may lead to a diagnosis. The discovery of a particular tool or other artifact in close proximity to fossilized skeletal remains may lead an anthropologist to conclude that the remains are those of a particular species of hominid, e.g., Neanderthal man or Modern man. However, if data are to be treated *statistically*, the observations must be expressed in *numerical* form. That is, they must be *quantitative*. Data appropriate to statistical analysis would therefore include *scores* on the schizophrenia scale of the Minnesota Multiphasic Personality Inventory, the *number* of correct answers on a multiple-choice final examination, the *vertical distance* between a fossilized bone and a flint cutting tool, the *number* of hours, minutes, and seconds required to run a marathon. Consequently, *statistical* data always consist of a *collection of numbers*.

2. Types of quantitative data

The numbers with which a statistician begins are ordinarily of two types, *measurements*, or *scores*, and *frequencies*. When the phenomena that a scientist observes are expressed as or translated into numerical values, these data are called *measurements*. Height expressed in feet and inches is a *measurement* of linear body size. Weight expressed in kilograms is a *measurement* of body mass. An IQ *score* is a *measurement* of academic aptitude. Employee ratings are *measurements* of work performance. In the general case, measurements are represented by letters falling toward the end of the alphabet, usually x, y, or z. When events are expressed as

numbers, the term *observations* is understood to mean *numerical* observations, i.e., *measurements*.

Some phenomena cannot be translated into numbers. That is, some observations are *qualitative*, or *categorical*, rather than *quantitative*. Gender (male versus female), political party affiliation (Democrat versus Republican), and diagnostic classification (schizophrenia, depression, psychopathy) are examples of qualitative observations. Whether the events one observes are quantitative (and expressed as numbers) or qualitative, one can record the *number of times* each event occurs. Data of this sort are called *frequencies*. Examples of frequencies would include the *number of students* who obtained a particular score of an examination, the *number of women* who favor a particular candidate or ballot proposal, the *number of students* who miss a particular examination item, the *number of times* a particular member of a baboon troop displays dominance or threat toward other members of the troop, and so forth. In the tables that follow, and subsequently throughout this text, a lowercase f will be use to indicate the frequency with which an event occurs.

B. REPRESENTATIONS OF DATA

Despite the fact that an 18-month-old baby is capable of mental activity that challenges the ability of even the most sophisticated computer, the human mind is very limited in some ways. Unlike a computer, for example, it can keep track of only a very small number of unorganized bits of information. Try this experiment on yourself: *Without putting them in any particular geometric pattern*, imagine first a single point in space, then two points, then three points, four points, and so on. Chances are that unless you "cheated" and organized the points systematically (e.g., located them at the corners of imaginary squares or arranged them in evenly spaced rows and columns) you lost track somewhere between five and nine.

Even a small-scale piece of research may involve dozens of observations, and studies that yield several thousand observations are not uncommon. Before a scientist can even contemplate the task of interpretation, therefore, data must be organized so that they are reduced to manageable proportions. The most common organizational schemes are *tabular* and *graphic* representations.

1. Tabular representation: the frequency table

Example 1.1. Let us suppose that a statistics teacher has 25 students each toss four coins and record the number of heads. These experiments might yield the following 25 results:

2 2 0 1 1 1 2 3 1 2 2 3 3 4 1 3 2 3 3 1 2 2 1 2 2

It is obvious that even with only 25 observations the data call for a more comprehensible display. One very simple expedient would be to array

them from the smallest value to the largest value:

0 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 4

This sort of presentation *organizes* data, but it does not *reduce* the quantity of data. However, it does reveal something interesting: Although the experiments produced 25 *observations*, these data comprise only 5 different, or distinguishable, numerical *values*. Every value is equal to 0, 1, 2, 3, or 4. Consequently, the display of data is condensed significantly with no loss of information if each observed value is paired with its *frequency*, as given in Table 1.1.

Table 1.1. Frequency distribution for number of heads obtained in toss of four coins with experiment repeated 25 times

Observation (x)	0	1	2	3	4
Frequency (f)	1	7	10	6	1

Because it indicates how frequencies are distributed among the outcomes of an experiment, a representation of data in which every observed value is paired with its frequency is called a *frequency distribution*.

In Example 1.1 the numerical value 2 appears 10 times, which constitutes 40 percent of the total number of observations (25). If, however, the value 2 had appeared 10 times in 100 observations, this would amount to only 10 percent of the observations. Sometimes, then, the *frequency* with which a particular value is observed is less revealing than is its *relative* frequency, that is, the frequency divided by the number of observations. A representation of data in which every observed value is paired with its relative frequency is called a *relative frequency distribution*. Customarily, the number of observations is denoted by the capital letter N, so relative frequency becomes f/N. The relative frequency distribution for the data in our imaginary coin-toss experiment is given in Table 1.2.

Another useful presentation of data is the *cumulative* frequency distribution. The cumulative frequency of a value is its frequency *plus the frequencies of all smaller values*. Accordingly, the cumulative frequencies given in Table 1.3 indicate that 18 students obtained two or *fewer* heads

Table 1.2. Relative frequency distribution for number of heads obtained in toss of four coins with experiment repeated 25 times

Observation (x)	0	1	2	3	4
Relative Frequency (f/N)	.04	.28	.40	.24	.04

Table 1.3. Cumulative frequency distribution and relative frequency distribution for number of heads obtained in toss of four coins with experiment repeated 25 times

Observation (x)	0	1	2	3	4
Cumulative f	1	8	18	24	25
Cumulative f/N	.04	.32	.72	.96	1.00

in their four tosses, 24 students obtained three or *fewer* heads, and all of the students obtained four or fewer heads.

The cumulative *relative* frequency of any value is similarly defined as the relative frequency of the value plus the relative frequencies of all smaller values. From Table 1.2 we know that the relative frequency of 0 heads is .04 and the relative frequency of 1 head is .28. As shown in the bottom row of Table 1.3, the cumulative relative frequency of 1 head is .04 + .28 = .32. It is also apparent from inspection of Table 1.3 that the cumulative relative frequency of any value is equal to its cumulative frequency divided by the total number of observations, N. The cumulative relative frequency of the largest observed value (e.g., 4) is, therefore, always equal to 1.0.

a. Tabular representation of grouped data. In the preceding example only five possible numbers could be observed. In some experiments, the number of values that it is possible to observe may be so large that even the sorts of frequency distributions discussed above become unwieldy.

Example 1.2.1. Suppose 1,000 persons each toss 100 coins and count the number of heads. In this experiment 101 different values might be observed (0 heads, 1 head, . . . , 100 heads), which means that a distribution that represented *every* value would be very cumbersome indeed. Furthermore, even with a large number of observations, it is likely that *some* of the values (e.g., 0 heads or 100 heads) would not turn up at all.

Under these conditions it is customary to *group* the values into intervals, or classes, and to tabulate the frequency (or relative frequency, or cumulative frequency, etc.) of observations in each class. Table 1.4 is a

Table 1.4. Frequency distribution for number of heads obtained in toss of 100 coins with experiment repeated 1,000 times

x	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89	90–99
f	13	41	93	147	240	200	160	67	13	26

grouped frequency distribution for 1,000 people tossing 100 coins and counting the number of heads.

A grouped distribution organizes *and* reduces data, but there is a loss of information. In this example we know that 240 persons obtained some number between 40 and 49, but we cannot tell from the table how many persons observed *exactly* 40 heads, how many times *exactly* 41 heads turned up, and so forth. Nevertheless, the benefits of comprehensibility are generally assumed to offset the disadvantage of information loss, especially if some of the x-values account for only a small percentage of the total observations.

In setting up a grouped distribution, one must first decide how many intervals to use and how wide the intervals will be. The *fewer* intervals one uses, the wider each interval and the *greater the information loss*; the *more* intervals one uses, the narrower each interval and the *less comprehensible* the display. Comprehensibility is largely in the eye of the beholder, but 6 to 20 intervals is usually workable. In addition, the number of intervals n should be small enough so that the *average* interval frequency (N/n) is greater than 5.

To calculate interval width, call the smallest value that will appear in the table x_a and the largest value x_b. If all n intervals are to be of the same width, then for integer data

$$\frac{x_b - x_a + 1}{n} = \text{interval width} \qquad \textbf{[1.1]}$$

The table is easier to construct and interpret if the interval width **[1.1]** is a whole number. This usually can be achieved either by adjusting n or by letting x_a be smaller than the smallest observed value or letting x_b be larger than the largest observed value.

Example 1.2.2. To see how this works, let us suppose that 1 and 99 were the smallest and largest values actually obtained in Example 1.2.1 and that we wanted to organize the data into 10 intervals ($n = 10$). Then $x_a = 1$ and $x_b = 99$, and by equation **[1.1]** the interval width is 9.9. This is not a whole number, so we have two choices. We could have used 11 intervals instead of 10. By equation **[1.1]** this results in an interval width of 9. The other option was either to let $x_b = 100$ or to let $x_a = 0$ (as illustrated in Table 1.4). Either adjustment yields an interval width of 10.

Then one must determine the lower and upper *class limits* of each interval, that is, the smallest and largest value, respectively, to be included in each class. Interval width may be thought of as the difference between the *lower* class limit of any interval and the *lower* class limit of the *next* interval and is therefore the most obvious consideration in calculating class limits. Once it was decided in Example 1.2.2 that the interval width was to be 10 and that $x_a = 0$, the lower class limit of the second interval was determined: $x_a + 10 = 10$.

A less obvious consideration is the precision of one's measurements. In Example 1.2.1, the numerical observations could assume only integer values. That is, they could be expressed only as whole numbers. One

cannot, after all, obtain 9.5 heads in 100 tosses of a coin. Since it is impossible to obtain any value *between* 9 and 10, there is no potential ambiguity created by setting the upper class limit of the first interval equal to 9 and the lower class limit of the second interval equal to 10. The situation is different if observations can take fractional values.

Example 1.3.1. Suppose that the x-values in Table 1.4 are *lengths* in inches. Depending on the precision of the measuring instrument, it is perfectly feasible to expect a measurement of, say, 9.5 in. or 9.75 in. or 9.0000001 in. all of which fall "between" the first and second classes defined in Table 1.4. If, on the other hand, we defined our classes as 0 to 10, 10 to 20, and so forth, an observation of exactly 10 inches would fall in *two* classes.

To avoid this dilemma statisticians have adopted the convention of extending class limits *half a measurement unit above the largest* observable value in the class and *half a measurement unit below the smallest* observable value in the class. The intervals in Table 1.4 would thus become $-.5$ to 9.5, 9.5 to 19.5, 19.5 to 29.5, and so forth. Whether the x-values represent measurements taken to the nearest inch or the number of heads obtained in 100 tosses of a coin, every possible observation can fall in one and *only one* interval. The principle is simple, but it can be difficult to put the principle into practice if observations are recorded in *multiples* of the units in which measurements were taken.

Example 1.3.2. Suppose once more that the x-values in Table 1.4 represent lengths in inches, but let us now imagine that our instrument is accurate to the nearest *tenth* of an inch and that our smallest measurement is 0.1 inches and our largest is 99.9 inches.

One immediate difficulty is that 0.1 and 9.9 are *decimal* values, and we said earlier that equation **[1.1]** is for *integer* data. The requirements for equation **[1.1]** are actually more specific and less restrictive: The 1 in the numerator represents 1 *unit of measurement*, so x_a and x_b must be *whole numbers* of units of measurement. By *units of measurement*, we mean the *most precise* units that the measurement instrument records. In this example, the unit of measurement is the *tenth* of an inch. Since the smallest and largest values are given in inches, we multiply them by 10 to express them in *tenths* of inches. That is, $x_a = 1$ *tenth* and $x_b = 999$ *tenths*. Then for $n = 10$ intervals, equation **[1.1]** gives an interval width of 99.9 *tenths*. If we set $x_a = 0$, the interval width calculated from equation **[1.1]** is once again a whole number:

$$\frac{x_b = x_a + 1}{n} = \frac{999 - 0 + 1}{10} = 100$$

The first interval therefore includes all measured lengths from 0 to 99 *tenths* of an inch, and the class limits are $-.5$ to 99.5 tenths. The second

interval includes all measured lengths from 100 to 199 *tenths* of an inch, and the class limits are 99.5 to 199.5 tenths, and so on. For a table with lengths expressed in *inches*, simply divide the class limits by 10, which gives us −.05 to 9.95, 9.95 to 19.95, 19.95 to 29.95, and so on.

If the observations are in *hundredths* of a unit, one multiplies x_a and x_b by 100, if in *thousandths*, by 1000, and so on. This approach also works if data are in *integer* (rather than fractional) multiples of the unit of measurement.

BOX 1.1

Calculating class limits for grouped data

Step 1. Find the smallest observation and the largest observation and express these values in *units* of measurement. That is, if observations are recorded in *tenths* of units, multiply by 10; if observations are recorded in *tens* of units, multiply by .1, and so on.

Step 2. Decide how many intervals you want. Call this number n.

Step 3. Calculate the interval width:

$$\frac{x_b - x_a + 1}{n}$$

where x_a is the smallest and x_b is the largest observable value represented in the table. As a first approximation, let x_a equal the smallest observation and let x_b equal the largest observation.

Step 4. If the interval width calculated in Step 3 is a whole number, go to Step 7. If the interval width is not a whole number, round it *up* to the nearest integer.

Step 5. Multiply the *rounded-up* interval width by n.

Step 6. Change n or adjust x_a or x_b (or both) so that $x_b - x_a + 1$ equals the value calculated in Step 5. Make x_a *less than* or equal to the smallest observation and x_b equal to or *greater than* the largest observation.

Step 7. Calculate *lower* class limits. The lower limit of the first interval is $x_a - .5$. The lower limit of the second interval is obtained by adding the interval width to the lower limit of the first interval, etc.

Step 8. Calculate *upper* class limits. The upper limit of an interval is equal to the lower limit of the next interval. The upper limit of the last interval is its lower limit plus the interval width.

Step 9. Express class limits in the same scale as the observations.

Example 1.4. Suppose that the data in Table 1.4 are annual incomes measured to the nearest thousand dollars and that the smallest observed value is \$1,000 and the largest is \$99,000. The 1 in equation **[1.1]** therefore represents 1 thousand dollars, and we multiply 1,000 and 99,000 by .001 to obtain x_b and x_b.

The procedure discussed above (and outlined in Box 1.1) assumes that intervals are to be of equal width. Although this is the most common practice, it is not unusual to leave the first or last interval "open ended." For example, if only one or two persons in Example 1.2.1 had obtained 0 to 9 heads (or 90 to 99 heads), the first interval might have been defined as all values below 20 (or the last interval as all values above 79).

2. Graphic representation: polygons and histograms

Tabular distributions provide both economy of representation and organization. The adage that "One picture is worth a thousand words," however, has a certain measure of truth in the representation of quantitative data. That is, data are often better understood if represented in graphic form than in tabular form. If we represent the values of x on the horizontal axis of a graph and the frequencies associated with these values on the vertical axis, then every point represents a numerical value of x and the frequency with which it is observed, f_x. The five points in Figure 1.1, for example, represent the number of heads observed in four tosses of a fair coin, each paired with its frequency of observation, as given in Table 1.1.

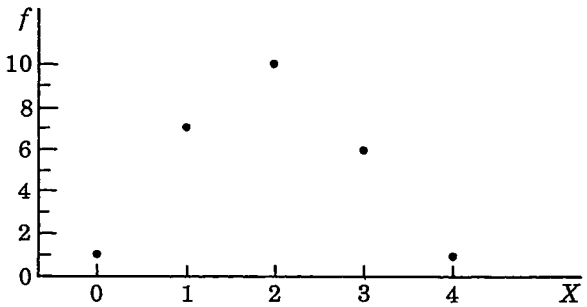

Figure 1.1. Point graph representing frequency distribution in Table 1.1.

When many points are involved, it is sometimes difficult to identify each point with its appropriate x-value. A point graph can be made more readable by connecting the points and creating a *frequency polygon*, as shown in Figure 1.2.