Data and Numbers

2

Chapter Outline
2.1 Data: Numerical Representation
2.2 Observations and Variables
2.3 Scales Used with Variables
2.4 Reliability and Validity
2.5 Randomized Response Technique
2.6 Common Data Problems

Appropriate use of statistical procedures requires that we understand the data and the process that generated them. This chapter focuses on data, specifically: (1) the link between numbers and phenomena, (2) types of variables, (3) data reliability and validity, and (4) ways data quality can be compromised.

2.1 Data: Numerical Representation

Any record, descriptive accounts, or symbolic representation of an attribute, event, or process may constitute a data point. Data are usually measured on a numerical scale or classified into categories that are numerically coded. Here are three examples:

1. Blood pressure (diastolic) is measured for all middle and high school students in a school district to learn what percent of students have a diastolic blood pressure reading over 90 mm Hg. [data = blood pressure reading]
2. All employees of a large company are asked to report their weight every month to evaluate the effects of a weight control program. [data = self-reported weight measurement]
3. The question "Have you ever driven a car while intoxicated?" was asked of all licensed drivers in a large university to build the case for an educational program. [data = yes (coded as 1) or no (coded as 0)]

We try to understand the real world — for example, blood pressure, weight, and the prevalence of drunken driving — through data recorded as or converted to numbers. This numerical representation and the understanding of the reality, however, do not occur automatically. It is easy for problems to occur in the conceptualization and measurement processes that make the data irrelevant or imprecise. Referring to the preceding examples, inexperienced school teachers may measure blood pressure inaccurately; those employees who do not measure their weight regularly each month may report inaccurate values; and some drivers may be hesitant to report drunken driving. Therefore, we must not draw any conclusions from the data before we determine whether or not any problems exist in the data and, if so, their possible effects. Guarding against

misuse of data is as important as learning how to make effective use of data. Repeated exposure to misuses of data may lead people to distrust data altogether. A century ago, George Bernard Shaw (1909) described people's attitudes toward statistical data this way:

> The man in the street. . . . All he knows is that "you can prove anything by figures," though he forgets this the moment figures are used to prove anything he wants to believe.

The situation is certainly far worse today as we are constantly exposed to numbers purported to be important in advertisements, news reporting, and election campaigns. We need to learn to use numbers carefully and to examine critically the meaning of the numbers in order to distinguish fact from fiction.

2.2 Observations and Variables

In statistics, we observe or measure characteristics, called *variables*, of study subjects, called *observational units*. For each study subject, the numerical values assigned to the variables are called *observations*. For example, in a study of hypertension among schoolchildren, the investigator measures systolic and diastolic blood pressures for each pupil. *Systolic and diastolic blood pressure are the variables, the blood pressure readings are the observations, and the pupils are the observational units.* We usually observe more than one variable on each unit. For example, in a study of hypertension among 500 school children, we may record each pupil's age, height, and weight in addition to the two kinds of blood pressure readings. In this case we have a data set of 500 students with observations recorded on each of five variables for each student or observational unit.

2.3 Scales Used with Variables

There are four scales used with variables: *nominal*, *ordinal*, *interval*, and *ratio*. The scales are defined in terms of the information conveyed by the numeric values assigned to the variable. The distinction between the scales is not terribly important. These scale types have frequently been used in the literature, so we are presenting them to be sure the reader understands them.

In some cases the numbers are simply indicators of a category. For example, when considering gender, 1 may be used to indicate that the person is female and 2 to indicate that the person is male. When the numbers merely indicate to which category a person belongs, a *nominal* scale is being used. Hence, gender is measured on a nominal scale, and it makes no difference what numeric values are used to represent females and males.

In other cases the numbers represent an ordering or ranking of the observational units on some variable. For example, from a worker's job description or work location, it may be possible to estimate the exposure to asbestos in the workplace, with 1 representing low, 2 representing medium, and 3 representing high exposure. In this case, the exposure to asbestos variable is measured on the *ordinal* scale. Values of 10, 50, and 100 could have been used instead of 1, 2, and 3 for representing the categories of low, medium, and high. The only requirement is that the order is maintained.

Other variables are measured on a scale of equal units — for example, temperature in degrees Celsius (*interval* scale) or height in centimeters (*ratio* scale). There is a subtle distinction between interval and ratio scales: A ratio scale has a zero value, which means there is none of the quantity being measured. For example, zero height means there is no height, whereas zero degrees Celsius does not mean there is no temperature. When a variable is measured on a ratio scale, the ratio of two numbers is meaningful. For example, a boy 140 centimeters tall is 70 centimeters taller and also twice as tall as a boy 70 centimeters tall. However, temperature in degrees Celsius is an interval variable but not a ratio variable because an oven at 300º is not twice as hot as one at 150º. This distinction between interval and ratio scales is of little importance in statistics, and both are measured on a scale continuously marked off in units.

These different scales measure three types of data: *nominal* (categorical), *ordinal* (ordered), and *continuous* (interval or ratio). The scale used often depends more on the method of measurement or the use made of it than on the property measured. The same property can be measured on different scales; for example, age can be measured in years (ratio scale), placed into young, middle-aged, and elderly age groups (ordinal scale), or classified as economically productive (ages 16 to 64) and dependent (under 16 and over 64) age groups (nominal scale). It is possible to convert a higher-level scale (ratio or interval) into a lower-level scale (ordinal and nominal scales) but not to convert from a lower level to a higher level. One final point is that all recorded measurements themselves are discrete. Age, for example, can be measured in years, months, or even in hours, but it is still measured in discrete steps. It is possible to talk about a continuous variable, yet actual measurements are limited by the measuring instruments.

2.4 Reliability and Validity

Data are collected by direct observation or measurement and from responses to questions. For example, height, weight, and blood pressure of school children are directly measured in a health examination. The investigator is concerned about accurate measurement. The measurement of height and weight sounds easy, but the measurement process must be well defined and used consistently. For example, we measure an individual's height without shoes on. Therefore, to understand any measurement, we need to know the operational definition — that is, the actual procedures used in the measurement. In measuring blood pressure, the investigator must specify what instrument is to be used, how much training will be given to the measurers, at what time of the day the blood pressure should be measured, what position the person should be in (sitting or standing), and how many times it should be measured.

There are two issues in specifying operational definitions: reliability and validity. *Reliability* requires that the operational definition should be sufficiently precise so that all persons using the procedure or repeated use of the procedure by the same person will have the same or approximately the same results. If the procedures for measuring height and weight of students are reliable, then the values measured by two observers — say, the teacher and the nurse — will be the same. If the person reading the blood pressure is hard of hearing, the diastolic blood pressure values, recorded at the point of complete cessation of the Korotkoff's sounds, or, if no cessation, at the point of muffling, may not be reliable.

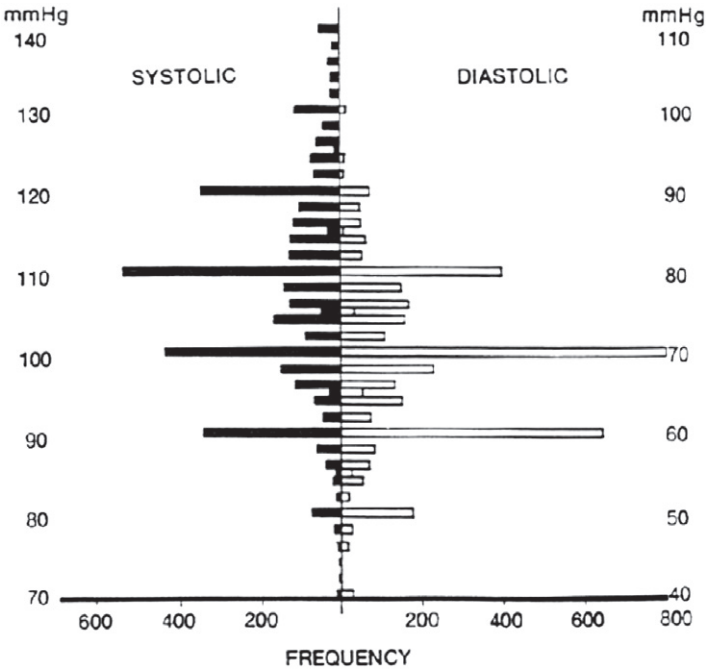

Figure 2.1 Blood pressure values (first reading) for 4053 children and adolescents in NHANES II. From Forthofer (1991).

Validity is concerned with the appropriateness of the operational definition — that is, whether or not the procedure measures what it is supposed to measure. For example, if a biased scale is used, the measured weight is not valid, even though the repeated measurements give the same results. Another example of a measurement that may not be valid is the blood pressure reading obtained when the wrong size cuff is used. In addition, the person reading the blood pressures may have a digit preference that also threatens validity. The data shown in Figure 2.1, from Forthofer (1991), suggests that there may have been a digit preference in the blood pressure data for children and adolescents in the second National Health and Nutrition Examination Survey (NHANES II). This survey, conducted by the NCHS from 1976 to 1980, provides representative health and nutrition data for the noninstitutionalized U.S. population. In this survey, the blood pressure values ending in zero have a much greater frequency of occurrence than the other values.

The reliability and validity issues are not only of concern for data obtained from measurements but also for data obtained from questionnaires. In fact, the concern may be greater because of the larger number of ways that problems threatening data accuracy can be introduced with questionnaires (Juster 1986; Marquis, Marquis, and Polich 1986; Suchman and Jordan 1990). One problem is that the question may be misinterpreted, and thus a wrong or irrelevant response may be elicited. For example, in a mail survey, a question used the phrase "place of death" instead of instructing the respondent to provide the county and state where a relative had died. One person responded that the deceased had died in bed. Problems like this one can be avoided or greatly reduced if careful thought goes into the design of questionnaires and into the preparation of instructions for the interviewers and the respondents. However, even when there are no obvious faults in the question, a different phrasing may have obtained a different

response. For example, age can be ascertained by asking age at the last birthday or date of birth. It is known that the question about the date of birth tends to obtain the more accurate age.

Another problem often encountered is that many people are uncomfortable in appearing to be out of step with society. As a result, these people may provide a socially acceptable but false answer about their feelings on an issue. A similar problem is that many people are reluctant to provide accurate information regarding personal matters, and often the respondent refuses to answer or intentionally distorts the response. Some issues are particularly sensitive — for example, questions about whether a woman has had an abortion or if a person has ever attempted suicide. The responses, if any are obtained, to these sensitive questions are of questionable accuracy. Now we'll look at some ways to obtain accurate data on sensitive issues.

2.5 Randomized Response Technique

There is a statistical technique that allows investigators to ask sensitive questions, for example, about drug use or driving under the influence of alcohol, in a way that should elicit an honest response. It is designed to protect the privacy of individuals and yet provide valid information. This technique is called randomized response (Campbell and Joiner 1973; Warner 1965) and has been used in surveys about abortions, drinking and driving, drug use, and cheating on examinations.

In this technique, a sensitive question is paired with a nonthreatening question, and the respondent is told to answer only one of the questions. The respondent uses a chance mechanism — for example, the toss of a coin — to determine which question is to be answered, and only the respondent knows which question was answered. The interviewer records the response without knowing which question was answered. It may appear that these answers are of little value, but the following example demonstrates how they can be useful.

In the drinking and driving situation, the sensitive question is "Have you driven a car while intoxicated during the last six months?" This question is paired with an unrelated, nonthreatening question, such as "Were you born in either September or October?" Each respondent is asked to toss a coin and not to reveal the outcome; those with heads are asked to answer the sensitive question and those with tails answer the nonthreatening question. The interviewer records the yes or no response without knowing which question is being answered. Since only the respondent knows which question has been answered, there is less reason to answer dishonestly.

Suppose 36 people were questioned and 12 gave "yes" answers. At first glance, this information does not seem very useful, since we do not know which question was answered. However, Figure 2.2 shows how we can use this information to estimate the proportion of the respondents who had been driving while intoxicated during the past six months.

Since each respondent tossed a fair coin, we expect that half the respondents answered the question about drunk driving and half answered the birthday question. We also expect that 1/6 (2 months out of 12) of those who answered the birthday question will give a yes response. Hence, the number of yes responses from the birthday question

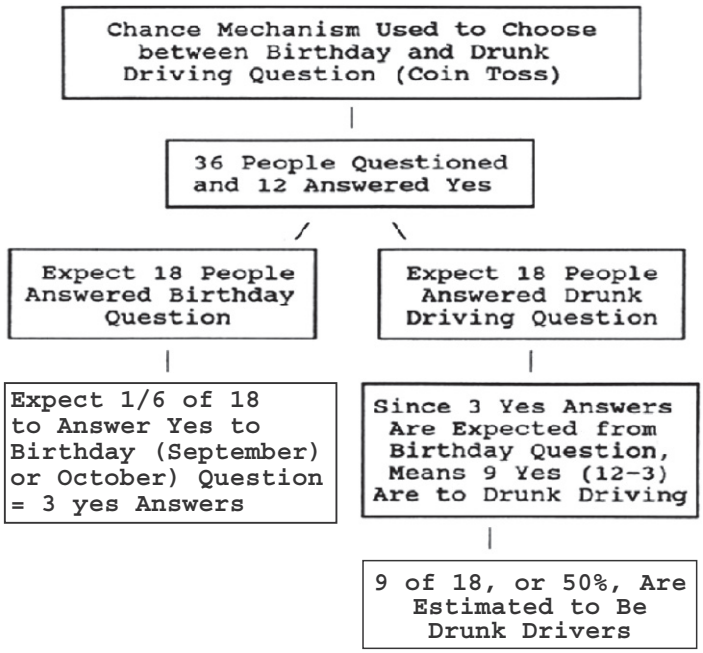

Figure 2.2 Use of randomized response information.

should be 3 [(36/2) ∗ (1/6)]; the expected number of yes responses to the drinking and driving question then is 9 (the 12 yes answers minus the 3 yes answers from the birthday question). Then the estimated proportion of drunk drivers is 50 percent (= 9/18).

There is no way to prove that the respondents answered honestly, but they are more likely to tell the truth when the randomized response method was used rather than the conventional direct question. Note that the data gathered by the randomized response technique cannot be used without understanding the process by which the data were obtained. Individual responses are not informative, but the aggregated responses can provide useful information at the group level. Of course, we need to include a sufficiently large number of respondents in the survey to make the estimate reliable.

2.6 Common Data Problems

Examination of data can sometimes provide evidence of poor quality. Some clues to poor quality include many missing values, impossible or unlikely values, inconsistencies, irregular patterns, and suspicious regularity. Data with too many missing values will be less useful in the analysis and may indicate that something went wrong with the data collection process. Sometimes data contain extreme values that are seemingly unreasonable. For example, a person's age of 120 would be suspicious, and 200 would be impossible. Missing values are often coded as 99 or 999 in the data file, and these may be mistakenly interpreted as valid ages. The detection of numerous extreme ages in a data set would cast doubt on the process by which the data were collected and recorded, and hence on all other observations, even if they appear reasonable. Also, inconsistencies are often present in the data set. For example, a college graduate's age of 15 may appear inconsistent with the usual progress in school, but it is difficult to attribute this to an error. Some inconsistencies are obvious errors. The following examples illustrate various problems with data.

Example 2.1

As described in Example 1.5, Edward Jarvis (1803–1884) discovered that there were numerous inconsistencies in the 1840 Population Census reports; for example, in many towns in the North, the numbers of black "insane and idiots" were larger than the total numbers of blacks in those towns. He published the results in medical journals and demanded that the federal government take remedial action. This demand led to a series of statistical reforms in the 1850 Population Census (Regan 1973).

Example 2.2

A careful inspection of data sometimes reveals irregular patterns. For example, ages reported in the 1945 census of Turkey have a much greater frequency of multiples of 5 than numbers ending in 4 or 6 and more even-numbered ages than odd-numbered ages (United Nations 1955), as shown in Figure 2.3. This tendency of digit preference in age reporting is quite common. Even in the U.S. Census we can find a slight clumping or heaping at age 65, when most of the social benefit programs for the elderly begin. The same phenomenon of digit preference is often found in laboratory measurements as we just saw with the blood pressure measurements in NHANES II.

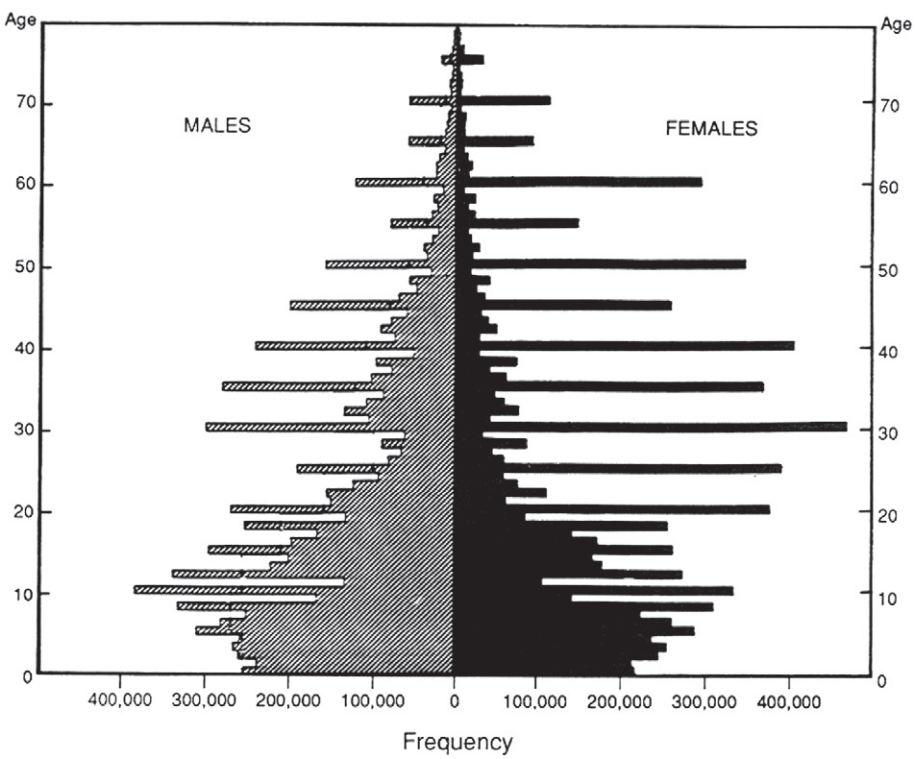

Figure 2.3 Population of Turkey, 1945, by sex and by single years of age.

Example 2.3

Large and consistent differences in the values of a variable may indicate that there was a change in the measurement process that should be investigated. An example of large differences is found in data used in the *Report of the Second Task Force on Blood Pressure Control in Children* (NHLBI Task Force 1987). Systolic blood pressure values for 5-year-old boys averaged 103.5 mmHg in a Pittsburgh study compared to 85.6 mmHg in a Houston study. These averages were based on 61 and 181 boys aged 5 in the Pittsburgh and Houston studies, respectively. Hence, these differences were not due to small sample sizes. Similar differences were seen for 5-year-old girls and for 3- and 4-year-old boys and girls as well. There are large differences between other studies also used by this Task Force, but the differences are smaller for older children. These incredibly large differences between the Pittsburgh and Houston studies were likely due to a difference in the measurement process. In the Houston study, the children were at the clinic at least 30 minutes before the blood pressure was measured compared to a much shorter wait in the Pittsburgh study. Since the measurement processes differed, the values obtained do not reflect the same variable across these two studies. The use of data from these two studies without any adjustment for the difference in the measurement process is questionable.

Example 2.4

The use of data from laboratories is another area in which it is crucial to monitor constantly the measurement process — in other words, the equipment and the personnel who use the equipment. In large multicenter trials that use different laboratories, or even with a single laboratory, referent samples are routinely sent to the laboratories to determine if the measurement processes are under control. This enables any problems to be detected quickly and prevents subjects from being either unduly alarmed or falsely comforted. It also prevents erroneous values from being entered into the data set. The Centers for Disease Control (CDC) has an interlaboratory program, and data from it demonstrate the need for monitoring. The CDC distributes samples to about 100 laboratories throughout the United States. The April 1980 results of measuring lead concentration in blood are shown in Figure 2.4

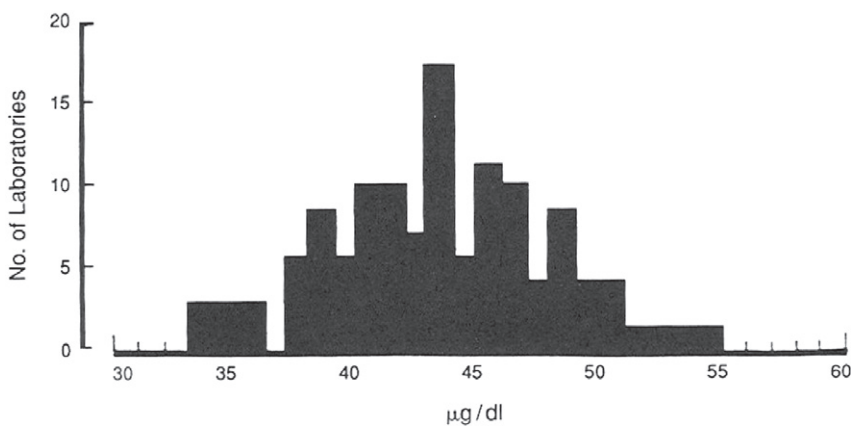

Figure 2.4 Distribution of measurements of blood lead concentration by separate laboratories, Centers for Disease Control.

(Hunter 1980). The best estimate of the blood lead concentration in the distributed sample was 41 micrograms per deciliter (μg/dL), but the average reported by all participating laboratories was 44 μg/dL. The large variability from the value of 41 shown in Figure 2.4 is a reason for concern, particularly since the usual value in human blood lies between 15 and 20 μg/dL.

Example 2.5

Of course, the lack of inconsistencies and irregularities does not mean that there are no problems with the data. Too much consistency and regularity sometimes is grounds for a special inquiry into its causes. Scientific frauds have been uncovered in some investigations in which the investigator discarded data that did not conform to theory. Abbe Gregor Mendel, the 19th-century monk who pioneered modern gene theory by breeding and crossbreeding pea plants, came up with such perfect results that later investigators concluded he had tailored his data to fit predetermined theories. Another well-known fabrication of data in science is the case of Sir Cyril Burt, a British pioneer of applied psychology. In his frequently cited studies of intelligence and its relation to heredity, he reported the same correlation in three studies of twins with different sample sizes (0.771 for twins reared apart and 0.944 for twins reared together). The consistency of his results eventually raised concern as it is highly unlikely that the exact same correlations would be found in studies of humans with different sample sizes. Science historians generally agree that his analyses were creations of his imagination with little or no data to support them (Gould 1981).

Example 2.6

Fabrication of data presents a real threat to the integrity of scientific investigation. The *San Francisco Chronicle* reported a case of data fabrication under the headline "Berkeley lab found research fabricated: Scientist accused of misconduct fired" (SFC 2002). A physicist at Lawrence Berkeley National Laboratory claimed the discovery of two new elements in the structure of the atomic nucleus in May 1998. Energy Secretary Bill Richardson called it "a stunning discovery which opens the door to further insights into the structure of the atomic nucleus." However, in follow-up experiments, outside labs were unable to replicate the results. The Lawrence Berkeley Laboratory retracted the finding after independent scientists were unable to duplicate the results. In announcing the laboratory's decision, director Charles V. Shank acknowledged that the false claim was "a result of fabricated research data and misconduct by one individual." He further reported, "The most elementary checks and data archiving were not done."

Conclusion

Data are a numerical representation of a phenomenon. By assigning numerical values to occurrences of the phenomenon, we are thus able to describe and analyze it. The assignment of the numerical values requires an understanding of the phenomenon and

careful measurement. In the measurement process, some unexpected problems may be introduced, and the data then contain the intended numerical facts as well as the unintended fictions. Therefore, we cannot use data blindly. The meanings of data and their implications have been explored in a number of examples in this chapter.

EXERCISES

2.1 Identify the scale used for each of the following variables:
   a. Calories consumed during the day
   b. Marital status
   c. Perceived health status reported as poor, fair, good, or excellent
   d. Blood type
   e. IQ score

2.2 A person's level of education can be measured in several ways. It could be recorded as years of education, or it could be treated as an ordinal variable — for example, less than high school, high school graduate, and so on. Is it always better to use years of education than the ordinal variable measurement of education? Explain your answer.

2.3 In a health interview survey, a large number of questions are asked. For the following items, discuss (1) how the variable should be defined operationally, (2) whether nonresponse is likely to be high or low, and (3) whether reliability is likely to be high or low. Explain your answers.
   a. Weight
   b. Height
   c. Family income
   d. Unemployment
   e. Number of stays in mental hospitals

2.4 The pulse is usually reported as the number of heartbeats per minute, but the actual measurement can be done in several different ways — for example:
   a. Count the beats for 60 seconds
   b. Count for 30 seconds and multiply the count by 2
   c. Count for 20 seconds and multiply the count by 3
   d. Count for 15 seconds and multiply the count by 4

   Which procedure would you recommend to be used in clinics, considering accuracy and practicality?

2.5 Two researchers coded five response categories to a question differently as follows:

Response Category	Researcher A	Researcher B
Strongly agree	1	2
Agree	2	1
Undecided	3	0
Disagree	4	–1
Strongly disagree	5	–2

   a. What type of scale is illustrated by Researcher A?
   b. What type of scale is illustrated by Researcher B?
   c. Which coding scheme would you use and why?

2.6 The first U.S. Census was taken in 1790 under the direction of Thomas Jefferson. The task of counting the people was given to 16 federal marshals, who in turn hired enumerators to complete the task in nine months. In October 1791, all of the census reports had been turned in except the one from South Carolina, which was not received until March 3, 1792. As can be expected, the marshalls encountered many obstacles and the counting was incomplete. The first census revealed a population of 3,929,326. This result was viewed as an undercount, as is indicated in the following excerpt from a letter written by Jefferson:

> I enclose you also a copy of our census, written in black ink so far as we have actual returns, and supplied by conjecture in red ink, where we have no returns; but the conjectures are known to be very near the truth. Making very small allowance for omissions, we are certainly above four millions. . . . (Washington 1853)

Discuss what types of obstacles they might have encountered and what might have led Jefferson to believe there was an undercounting of the people.

2.7 The National Center for Health Statistics matched a sample of death certificates in 1960 with the 1960 population census records to assess the quality of data and reported the following results (NCHS 1968):

Agreement and Disagreement in Age Reporting, 1960

		White		Nonwhite	
	Total	**Male**	**Female**	**Male**	**Female**
Agreement	68.8%	74.5%	67.9%	44.7%	36.9%
Disagreement					
1 year difference	17.8	16.6	18.8	20.8	20.2
2+ year difference	13.4	8.9	13.3	34.5	42.9

Do you think that the age reported in the death certificate is more accurate than that reported in the census? How do you explain the differential agreement by gender and race? How do you think these disagreements affect the age-specific death rates calculated by single years and those computed by five-year age groups?

2.7 Discuss possible reasons for the digit preference in the 1945 population census of Turkey that is shown in Figure 2.3. Why was the digit preference problem more prominent among females than among males? How would you improve the quality of age reporting in census or surveys? How do you think the digit preference affects the age-specific rates calculated by single years of age and those computed by five-year age groups?

2.8 Get the latest vital statistics report for your state from a library and find out the following:

a. Are residents of your state who died in a foreign country included in the report?
b. Are the data from your state report consistent with the data from the *National Vital Statistics Report* from the National Center for Health Statistics?
c. Is an infant born to a foreign student couple in your state included in the report?

REFERENCES

Campbell, C., and B. L. Joiner. "How to Get the Answer without Being Sure You've Asked the Question." *The American Statistician* 27:229–231, 1973.

Forthofer, R. N. "Blood Pressure Standards in Children." Paper presented at the American Statistical Association Meeting, August 1991. See Appendix C for details of the National Health and Nutrition Examination Survey.

Gould, S. J. *The Mismeasure of Man*. New York W. W. Norton, 1981.

Hunter, J. S. "The National System of Scientific Measurement." *Science* 210:869–874, 1980.

Juster, F. T. "Response Errors in the Measurement of Time Use." *Journal of the American Statistical Association* 81:390–402, 1986.

Marquis, K. H., M. S. Marquis, and J. M. Polich. "Response Bias and Reliability in Sensitive Topic Surveys." *Journal of the American Statistical Association* 81:381–389, 1986.

National Center for Health Statistics, *Vital and Health Statistics, Series 2*, November 29, 1968.

The NHLBI Task Force on Blood Pressure Control in Children. "The Report of the Second Task Force on Blood Pressure Control in Children, 1987." *Pediatrics* 79:1–25, 1987.

Regan, O. G. "Statistical Reforms Accelerated by Sixth Census Errors." *Journal of the American Statistical Association* 68:540–546, 1973. See Appendix D for details of the population census.

San Francisco Chronicle. July 13, 2002, p. 1.

Shaw, S. "Preface on Doctors." In *The Doctor's Dilemma*. New York: Brentano's, 1909.

Suchman, L., and B. Jordan. "Interactional Troubles in Face-to-Face Survey Interviews." *Journal of the American Statistical Association* 85:232–241, 1990.

United Nations. *Methods of Appraisal of Quality of Basic Data for Population Estimates, Population Studies, No. 23.* New York: United Nations Dept. of Economic and Social Affairs, 1955, p. 34.

Warner, S. L. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60:63–69, 1965.

Washington, H. A., ed. *The Writings of Thomas Jefferson*, III, 287. Washington, DC: Taylor & Maury, 1853.