CHAPTER 1

STATISTICS AND EXCEL

Statistics is a subject that for many people is pure tedium. For others, it is more likely to be anathema. Still others find statistics interesting, even stimulating, but they are usually in the minority in any group.

This book is premised on the recognition that in the health care industry, as indeed among people in any industry or discipline, there are at least these three different views of statistics, and that any statistics class is likely to be made up more of the first two groups than the last one. This book provides an introduction to statistics in health policy and administration that is relevant, useful, challenging, and informative.

1.1 How This Book Differs from Other Statistics Texts

The primary difference between this statistics text and most others is that this text uses Microsoft Excel as the tool for carrying out statistical operations and understanding statistical concepts as they relate to health policy and health administration issues. This is not to say that no other statistics texts use Excel. Levine, Stephan, Szabat (2013) have produced a very useable text, *Statistics for Managers Using Microsoft Excel*. But their book focuses almost exclusively on non–health-related topics. We agree that the closer the applications of statistics are to students' real-life interests and experiences, the more effective students will be in understanding and using statistics. Consequently, this book focuses its examples entirely on subjects that should be immediately familiar to people in the health care industry.

LEARNING OBJECTIVES

- Understand how this book differs from other statistics texts
- Understand how knowledge of statistics may be beneficial to health policy or health administration professionals
- Understand the "big picture" with regard to the use of statistics for health policy and administration
- Understand the definitions of the following terms:
  - Populations and samples
  - Random and nonrandom samples
  - Types of random samples
  - Variables, independent and dependent
- Identify the five separate statistical tests: chi-square test, the *t* test, analysis of variance (ANOVA), regression analysis, and *Logit*

Excel, which most people know as a *spreadsheet* program for creating budgets, comparing budgeted and expended amounts, and generally fulfilling accounting needs, is also a very powerful statistical tool. Books that do not use Excel for teaching statistics generally leave the question of how to carry out the actual statistical operations in the hands of the student or the instructor. It is often assumed that relatively simple calculations, such as means, standard deviations, and *t* tests, will be carried out on paper or with a calculator. For more complicated calculations, the assumption is usually that a dedicated statistical package, such as SAS, SPSS, STATA, or SYSTAT, will be used. There are at least two problems with this approach that we hope to overcome in this book. First, calculations done on paper, or even those done with a calculator, can make even simple statistical operations overly tedious and prone to errors in arithmetic. Second, because dedicated statistical packages are designed for use rather than for teaching, they often obscure the actual process of calculating the statistical results, thereby hindering students' understanding of both how the statistic is calculated and what the statistic means.

In general, this is not true of Excel. It is true that when using this book, a certain amount of time must be devoted to the understanding of how to use Excel as a statistical tool. But once that has been done, Excel makes the process of carrying out the statistical procedures under consideration relatively clear and transparent. The student should end up with a better understanding of what the statistic means, through an understanding of how it is calculated, and not simply come away with the ability to get a result by entering a few commands into a statistical package. This is not to say that Excel cannot be used to eliminate many of the steps needed to get particular statistical results. A number of statistical tests and procedures are available as add-ins to Excel. However, using Excel as a relatively powerful—yet transparent—calculator can lead to a much clearer understanding of what a statistic means and how it may be used.

1.2 Statistical Applications in Health Policy and Health Administration

When teaching statistics to health policy and health administration students, we often encounter the same question: "All these statistics are fine, but how do they apply to anything I am concerned with?" The question not only is a reasonable one, but also points directly to one of the most important and difficult challenges for a statistics teacher, a statistics class, or a statistics text. How can it be demonstrated that these statistics have

any real relevance to anything that the average person working in the health care industry ever needs to know or do?

To work toward a better understanding of why and when the knowledge of statistics may be useful to someone working in health policy or health administration, we've selected six examples of situations in which statistical applications can play a role. All six of these examples were inspired by real problems faced by students in statistics classes, and they represent real statistical challenges that students have faced and hoped to solve. In virtually every case, the person who presented the problem recognized it as one that could probably be dealt with using some statistical tool. But also in every case, the solution to the problem was not obvious in the absence of some understanding of statistics. Although these case examples are not likely to resonate with every reader, perhaps they will give many readers a little better insight into why knowledge of statistics can be useful.

Documentation of Medicare Reimbursement Claims

The Pentad Home Health Agency provides home health services in five counties of an eastern state. The agency must be certain that its *Medicare* reimbursement claims are appropriately and correctly documented in order to ensure that Medicare will process these claims and issue benefits in a timely manner. All physician orders, including medications, home visits for physical therapy, home visits of skilled nursing staff, and any other orders for service, must be correctly documented on a Form CMS-485. Poorly or otherwise inadequately prepared documentation can lead to rejection or delay in processing of the claim for reimbursement by the Centers for Medicare and Medicaid Services (CMS).

Pentad serves about 800 clients in the five-county region. In order to assure themselves that all records are properly documented, the administration runs a chart audit of 1 in 10 charts each quarter. The audit seeks to determine (1) whether all orders indicated in the chart have been carried out and (2) if the orders have been correctly documented in the Form CMS-485. Orders that have not been carried out, or orders incorrectly documented, lead to follow-up training and intervention to address these issues and ensure that the orders and documentation are properly prepared going forward.

Historically, the chart audit has been done by selecting each tenth chart, commencing at the beginning or at the end of the chart list. Typically, the chart audit determines that the majority of charts, usually 85 to 95 percent, have been correctly documented. But there are occasionally areas, such

as in skilled nursing care, where the percentage of correct documentation may fall below that level. When this happens, the administration initiates appropriate corrective action.

Sampling, Data Display, and Probability

One of the questions of the audit has been the selection of the sample. Because the list of clients changes relatively slowly, the selection of every tenth chart often results in the same charts being selected for audit from one quarter to the next. That being the case, a different strategy for chart selection is desirable. It has been suggested by statisticians that using a strictly random sample of the charts might be a better way to select them for quarterly review, as this selection would have a lesser likelihood of resulting in a review of the same charts from quarter to quarter. But how does one go about drawing a strictly random sample from any population? Or, for that matter, what does "strictly random" actually mean and why is it important beyond the likelihood that the same files may not be picked from quarter to quarter? These questions are addressed by statistics, specifically the statistics associated with sample selection and data collection.

Another question related to the audit concerns when to initiate corrective action. Suppose a sample of 1 in 10 records is drawn (for 800 clients that would be 80 records) and it is discovered that 20 of the records have been incorrectly documented. Twenty of 80 records incorrectly documented would mean that only 75 percent of the records were correctly documented. This would suggest that an intervention should be initiated to correct the documentation problem. But it was a *sample* of the 800 records that was examined, not the entire 800. Suppose that the 20 incorrectly documented records were, by the luck of the draw, so to speak, the only incorrectly documented records in the entire 800. That would mean that only 2.5 percent of the cases were incorrectly documented.

If the required corrective action were an expensive five-day workshop on correct documentation, the agency might not want to incur that expense when 97.5 percent of all cases are correctly documented. But how would the agency know from a sample what proportion of the total 800 cases might be incorrectly documented, and how would they know the likelihood that fewer than, say, 85 percent of all cases were correctly documented if 75 percent of a sample were correctly documented? This, again, is a subject of statistics.

Emergency Trauma Color Code

The emergency department (ED) of a university hospital was the site of difficulties arising from poor response time to serious trauma. Guidelines

indicate that a trauma surgeon must attend for a certain level of trauma severity within 20 minutes and that other trauma, still severe but less so, should be attended by a trauma nurse within a comparable time. In general, it had been found that the response time for the ED in the university hospital was more or less the same for all levels of severity of trauma—too long for severe cases and often quicker than necessary, given competing priorities, for less severe cases.

The ED director knew that when a trauma case was en route to the hospital, the ambulance attendants called the ED to advise that an emergency was on its way. Part of the problem as perceived by the director of the ED was that the call-in did not differentiate the trauma according to severity. The ED director decided to institute a system whereby the ambulance attendants would assign a code red to the most severe trauma cases, a code yellow to less severe trauma cases, and no color code to the least severe trauma cases. The color code of the trauma would be made known to the ED as the patient was being transported to the facility. The intent of this coding was to ensure that the most severe traumas were attended within the 20-minute window. This in turn was expected to reduce the overall time from admission to the ED to discharge of the patient to the appropriate hospital department. (All trauma cases at the red or yellow level of severity are transferred from the ED to a hospital department.)

Descriptive Statistics, Confidence Limits, and Categorical Data

A major concern of the director of the ED was whether the new system actually reduced the overall time between admission to the ED, treatment of the patient in the ED, and discharge to the appropriate hospital department. The director of the ED had considerable information about each ED admission going back a period of several months before the implementation of the new color-coding system and six months of experience with the system after it was implemented. This information includes the precise time that each trauma patient was admitted to the ED and the time that the patient was discharged to the appropriate hospital department.

The information also includes the severity of the trauma at admission to the ED on a scale of 0 to 75, as well as information related to gender, age, and whether the admission occurred before or after the color-coding system was implemented. The ED director also has information about the color code assigned after the system was initiated that can generally be equated to the severity score assigned at admission to the ED. Trauma scoring 20 or more on the scale would be assigned code red; below 20, code yellow; and not on the scale, no color.

Inferential Statistics, Analysis of Variance, and Regression

The question the ED director wishes to address is how she can use her data to determine whether the color-coding system has reduced the time that trauma victims spend in the ED before being discharged to the appropriate hospital department. At the simplest level, this is a question that can be addressed by using a statistic called the *t* test for comparing two different groups. At a more complex level, the ED director can address the question of whether any difference in waiting time in the ED can be seen as at all related to changes in severity levels of patients before or after the color-coding scheme was introduced. She can also examine whether other changes in the nature of the patients who arrived as trauma victims before and after the introduction of the color-coding scheme might be the cause of possible differences in waiting time. These questions can be addressed by using regression analysis.

Two Caveats of Statistics: Establishing a Significant Difference and Causality

Two caveats regarding the use of statistics apply directly to this example. The first is that no statistical analysis may be needed at all if the waiting time after the initiation of the color-coding scheme is clearly shorter than the waiting time before. Suppose, for example, that the average waiting time before the color-coding scheme was three hours from admission to the ED to discharge to hospital department, and that after the initiation of the scheme, the average waiting time was 45 minutes. In this scenario, no statistical significance tests would be required to show that the color-coding scheme was associated with a clear improvement in waiting time. Furthermore, it is likely that the color-coding scheme would not only become a permanent part of the ED armamentarium of the university hospital but also be adopted widely by other hospitals.

However, suppose that after the initiation of the color-coding scheme the average waiting time in the ED was reduced from 3 hours to 2 hours and 50 minutes. A statistical test (probably the *t* test) would show whether 170 minutes of waiting was actually less, statistically, than 180 minutes. Although such a small difference may seem to have little practical significance, it may represent a statistically significant difference. In such a case, the administrator would have to decide whether to retain an intervention (the color-coding scheme) that had a statistical, but not a practical, effect.

The second caveat to the use of statistics is the importance of understanding that a statistical test cannot establish causality. It might be possible,

statistically, to show that the color-coding scheme was associated with a statistical reduction in waiting time. But in the absence of a more rigorous study design, it is not possible to say that the color-coding scheme actually *caused* the reduction in waiting time. In a setting such as this, where measurements are taken before and after some intervention (in this case, the color-coding scheme), a number of factors other than the color-coding scheme might have accounted for an improvement in waiting time.

The very recognition of the problem and consequent concern by ED physicians and nurses may have had more effect on waiting time than the color-coding scheme itself. But this is not a question that statistics, per se, can resolve. Such questions may be resolved in whole or in part by the nature of a study design. A double-blind, random clinical trial, for example, is a very powerful design for resolving the question of causality. But, in general, statistical analysis alone cannot determine whether an observed result has occurred *because of* a particular intervention. All that statistical analysis can do is establish whether two events (in this case, the color-coding scheme and the improvement in waiting time) are or are not independent of each other. This notion of independence will come up many more times, and, in many ways, it is the focus of much of this book.

Length of Stay, Readmission Rates, and Cost per Case in a Hospital Alliance

The ever-increasing costs of providing hospital services have sparked a keen interest on the part of hospital administrators in practical mechanisms that can account for—and control or mitigate—those costs. The administrators for the Sea Coast Alliance, a system of eight hospitals, want to be able to use the previous case data to provide guidance on how to control costs. Because Sea Coast is associated with eight hospitals, it has a substantial volume of case data that the administrators believe can be useful in achieving their goal.

There are, in particular, three measures of hospital performance related to costs that need to be evaluated in this case: length of stay (LOS), readmission rates, and cost per case. One of the initial questions is whether there are real differences among the eight hospitals in these three cost-related measures of hospital performance. The question of what is a real difference is, of course, paramount. If the average LOS in one of Sea Coast's hospitals is five days for all hospital stays over the past year while the average LOS for another Sea Coast hospital is six days, is this a real difference? Given certain assumptions about what the average LOS for a year in these two hospitals represents, this is a question that can be answered with statistics.

Establishing a Statistical Difference between Two Groups Using *t* Tests and Analysis of Variance

If the interest is in comparing two hospitals with each other, the statistic that could be used would be a *t* test. In general, though, the real interest would be in deciding if there was any difference among all eight hospitals, taken simultaneously. This question can be examined in a couple different ways. One would be to use analysis of variance (ANOVA). Another would be to use multiple regression. If by using any of these statistical techniques it is determined that LOS is different from hospital to hospital, efforts could be directed toward determining whether lessons could be learned from the better performers about how to control costs that might be applied to the poorer performers. The same approach could be applied to understanding readmission rates and cost per case.

One particular focus of the Sea Coast administrators is diagnostic-related groupings (DRG) that have especially high costs. In addition to looking at the performance across the eight hospitals on high-cost DRGs, Sea Coast would like to be able to examine the question of whether individual physicians seemed to stand out in LOS, readmission rates, or cost per case. Identification of individual physicians who have unusually high LOS, readmission rates, or cost per case can allow Sea Coast to engage in selective educational efforts toward reduced costs. But an important question in looking at individual physician differences is whether what may appear to be unusually high values for LOS, readmission rates, or cost per case actually are unusual. Again, this question can be answered with statistics. In particular, predicted values for LOS, readmission rates, and cost per case can be determined by using regression analysis.

Establishing a Statistical Difference between Two Groups Using Regression

Regression can also be used to assess whether the differences that may exist across hospitals or across individual physicians could be attributed to differences in the mix of cases or patients whom the hospitals accept or the physicians see. Such differences may be attributable statistically to such characteristics of patients as sex, age, and payer, which may differ across the eight hospitals or the numerous physicians. There might also be differences across cases related to severity or diagnoses. If these were differentially distributed among hospitals or physicians, they could account for visible differences. All of these questions can be addressed (although not necessarily answered in full) by using multiple regression analysis.

At the Carteret Falls regional hospital, the emergency department has instituted a major change in how physicians are contracted to provide services and consequently how services are billed in the ED. Prior to January 1 of a recent year, emergency department physicians were employed by the regional hospital, and the hospital billed for their services. Beginning January 1, the physicians became private contractors working within the emergency department, essentially working on their own time and billing for that time directly. Bills are submitted to Medicare under five different coding levels that correspond to the level of reimbursement Medicare provides. The higher the coding level, the more Medicare actually reimburses for the service.

The practice manager for the physicians is concerned that she will begin to see the billing level creep upward as physicians begin billing for their own services. As the distinction between levels is frequently a matter of judgment, the practice manager is concerned that physicians may begin, even unconsciously, upgrading the level of the coding because it is directly tied to their reimbursement. The question the practice manager faces is how to decide if the physicians are upgrading their codes, consciously or not, after the initiation of the new system. If they are, the practice needs to take steps, either to ensure that the coding remains constant before and after the change in billing, or to have very good justification for CMS as to why it should be different.

Establishing a Difference Using Statistical Tests

The first problem is to determine if the billing levels have changed from before the change in billing to after the change. But it is simply not enough to say that there is a change, if one is seen to have occurred. It is critical to be able to say that this change is or is not a change that would have been expected, given the pattern of billing in the past. In other words, is any change seen as large enough to be deemed a real change and not just a chance occurrence? If a change has occurred, and if it is large enough to be viewed as a real change, then the second problem arises: determining whether anything in the nature of the ED cases before and after the billing change might account for the difference and thus be the explanation of the difference that will satisfy the Medicare administration.

Both of these problems can be examined by using statistics. In regard to the first problem, a difference between the distribution of billings across the five categories before and after the change in billing source can be assessed by using the *chi-square statistic*. Or, because the amount of a bill is constant within the five categories, it is also possible to compare the two

groups, before and after, using the *t* test. The second problem, of whether any difference can be attributed to changed characteristics of the cases seen in the ED, can be assessed by using regression—when the cost of the bills before and after is the measure of change.

A Study of the Effectiveness of Breast Cancer Education

A resident at a local hospital has been asked by the senior physician to develop a pilot study on the effectiveness of two alternative approaches to breast cancer education, both aimed at women coming to a women's health center. The first alternative is the distribution of a brochure on breast cancer to the women when they arrive at the clinic. The second alternative is time specifically allocated during a clinic visit wherein the physician spends 5 to 10 minutes with each woman, giving direct information and answering questions on the same topics covered in the brochure.

The resident recognizes that a study can be designed in which one group of women would receive the brochure and a second group would participate in a session with a physician. She also believes that a questionnaire can be developed to measure the knowledge women have about breast cancer before the distribution of the brochure or session with the physician and after either event, to assess any difference in knowledge. She also is concerned about possibly needing a control group of women to determine whether either method of information dissemination is better than no intervention at all. And perhaps she is interested in whether the brochure and discussion with the physician together would be better than either alternative singly.

Although she has been asked to design a pilot study only, the student-resident wishes to be as careful and as thoughtful as possible in developing her study. She might consider a number of different alternatives. One would be a simple *t* test of the difference between a group of women who received the brochure and a group of women who participated in the sessions with a physician. The measurement for this comparison could be the knowledge assessment administered either after the distribution of the brochure or after the physician encounter.

Using Analysis of Variance versus *t* Tests

But the resident may very well be dissatisfied with the simple *t* test. One problem is that she wants to include a control group of women who received no intervention at all. She may also wish to include another group of women—those who received the brochure *and* spoke to a physician. Again, the effect of any intervention (or of none) could be measured using

her previously developed knowledge assessment, administered after the fact. This assessment could be carried out using a one-way analysis of variance (ANOVA).

Again, however, the resident may not be entirely satisfied with either the *t* test or the one-way analysis of variance. She might wish to be sure that in her comparison she is not simply measuring a difference between women that existed prior to the receipt of the brochures or the physician sessions. To ensure this, she might wish to measure women's knowledge both before and after the interventions, at both times using her knowledge assessment questionnaire. This assessment could be carried out using a two-way analysis of variance.

Regardless of whether the resident decides to go with a *t* test, a one-way ANOVA, or a two-way ANOVA, one of the more important aspects of the study will be to randomly allocate women to the experimental or control group. When measuring knowledge only after the intervention, the resident will be able to ensure that prior knowledge is not responsible for any differences she might find only if she is certain that there is only a small chance that the groups of women receiving different interventions were not different to begin with. The only effective way to ensure this is through random assignment to the groups.

Calculating a Standard Hourly Rate for Health Care Personnel

In an article published in *Healthcare Financial Management*, Richard McDermott (2001) discusses the problem and importance of establishing standard hourly labor rates for employee reimbursement. He points out that many compensation systems have been worked out over a number of years by different human resources directors, each with his own compensation philosophy. As a result, these systems may fail to reflect market conditions and may be inconsistent in their treatment of differing categories of labor. McDermott suggests a regression approach to calculating labor rates that have both internal consistency and external validity.

The approach McDermott suggests for establishing labor rates is based on an example in which he provides data for 10 different positions. Each position is assigned a score from 0 to 5 based on the degree of complexity in the job in five separate categories, such as level of decision making, amount of planning required, educational requirements, and so on. He does not indicate specifically which five characteristics are employed in the example. The assigned scores in each category would have been developed through an examination of the requirements of the job by a compensation

consultant after interviews with the incumbent of each position. Each of the 10 positions also includes an actual hourly wage.

Relating Variables via Regression Analysis: Some Issues

Regression analysis was used by McDermott to assess the relationship between each of the five characteristics of the job and the actual hourly compensation. The regression analysis indicates both the relationship between any one of the five characteristics (when all characteristics are considered simultaneously) and hourly compensation, and it provides a set of coefficients by which to translate assigned values on any set of characteristics into a predicted hourly compensation. This, then, becomes a relatively objective means of determining hourly compensation for a person in any position.

There are purely statistical problems in using this regression approach, at least as discussed by McDermott, to propose hourly compensation. Particularly, 10 observations (the jobs assessed) are rarely considered by statisticians to be an adequate number with which to assess the relationship between five predictor variables (the characteristics) and a sixth predicted variable (the hourly compensation). While there are no absolute rules for the number of observations needed relative to the number of variables assessed, it is often accepted that there should be at least three times as many observations as variables, and some statisticians suggest a ratio of as many as 10 observations to each variable.

A second problem with this approach to assigning hourly compensation is inherent in the fact that many jobs are essentially the same, with similar job titles and expectations. If such jobs are included in an analysis of the type discussed here, one of the basic premises of regression analysis, that there is no correlation between observations, will be violated. This can be overcome, in part, by the use of *dummy variables*.

EXERCISES FOR SECTION 1.2

1. Look in magazines or journals that might deal with subjects relevant to your current work situation or your planned area of work. Can you find discussions that involve statistics? If so, briefly describe these and how the statistics are applied.

2. Consider experience you have had or a situation that you are familiar with in your work or planned area of work. Can you imagine any way that this experience or situation could benefit from the application of statistics? Briefly describe this experience or situation.

1.3 What Is the "Big Picture"?

Having discussed several specific examples of why a health care worker might be interested in knowing statistics, and having suggested some ways in which this book can provide that knowledge, we now want to step back and ask, what are we actually trying to do? To put it another way, what is the "big picture"? The big picture is basically this: In any situation in which statistics may be applicable and useful, the beginning is the question for which an answer is sought. Are our Medicare claims properly completed? Does a color-coding scheme for emergencies reduce emergency room time? Do the hospitals in a region differ in terms of costs? Will an education strategy work?

In attempting to answer any of these questions, it is generally true that not all the data that might bear on the answer will ever be available. In some cases, though it might be possible to access all the relevant data, it might just be too costly to do so. This would be true, for example, with regard to Medicare claims in a home health agency. Because it would be very costly to examine every claim, the answer must rely on a subset of the claims. In other cases, it might never be possible to access all records or all people who might be necessary to provide a definitive answer. With regard to the question of whether an education intervention will increase the knowledge women have of breast cancer, it would be physically impossible to assess all women who might ever be potential subjects of such an education effort.

The consequence of this inability to access all the data that may be relevant to a decision means that it will be necessary, generally, to rely on only a subset of the data—a *sample*—to make whatever decision is called for. Statistics is about the rules and procedures for using a subset of the data to make the decisions desired. In learning statistics, one learns these rules and procedures, when and to what types of data to apply, and the confidence that one can have in using the results of the sample data to make inferences about the total population. This is the basic function of statistics.

In considering the function of statistics as the process of using a sample to make inferences about a larger population, it is important to point out that in many cases this is the only way, and often the best way, to reach decisions. In the case of the acceptability of Medicare claims, for example, it is highly likely that if the staff of a home health agency were required to review every one of the files, they would become tired, bored, and generally unhappy with the process. They might make mistakes or errors in judgment that they would perhaps not make if they were working with only a sample of records. When they had finished their audit of the entire population of

claims, they could very well have less useful information than they would have had under the limitations of a sample. And, in any case, the cost would be prohibitive.

1.4 Some Initial Definitions

Before proceeding much further in this discussion, it is essential to make certain that everyone is clear about a number of terms that will crop up again and again in this text.

Populations and Samples

Populations are those groups of entities about which there is an interest. Populations may be made up of people—for example, all citizens of the United States or all patients who have shown up or ever will show up at a specific emergency room clinic. Populations may be made up of organizations—for example, all hospitals in the United States, or all long-term care facilities in New York state. Populations may be made up of political entities—for example, all the countries in the world or all the counties in California. All the persons who might ever receive a particular type of assessment are a population, and all people who ever will have an magnetic resonance imaging (MRI) could be considered another population, or these two groups together could be considered a population.

Populations tend to be large groups of individual persons, objects, or items from which samples can be taken.

In general, we are interested in characteristics of populations as opposed to characteristics of samples. We might wish to know the average cholesterol level of all persons age 55 or older (a population). We might wish to know the daily bed occupancy rate for hospitals in the United States (a population). Or we might wish to know the effect of a specific drug on cholesterol levels of some group of people (a population). If we knew these specific pieces of information, we would know a *parameter*. Parameters are information about populations. In general, except for some data collected by a complete census of the population (even most complete censuses are not complete), we do not know parameters. The best we can usually do is estimate parameters based on a subset of observations taken from populations.

Samples are subsets of populations.

Samples are subsets of populations. If a population of interest consists of all patients who have shown up or ever will show up at a specific emergency room clinic, a sample from that population could be all the patients who are there on a specific afternoon. If a population of interest consists of all long-term care facilities in New York state, a sample from that population might be all these facilities in Buffalo, Syracuse, and Albany. If a population of interest is all persons who have used or ever will use a cholesterol-reducing drug, a sample from that population might be all persons who received prescriptions for such a drug from a specific physician. An individual member of a sample might be referred to as an element of the sample or, more commonly, as an observation.

Information from samples can be used to make estimates of information about populations (parameters). When a specific value from a sample is used to make an estimate of the same value for a population, the sample value is known as a statistic. Statistics are to samples what parameters are to populations. If the parameter of interest is, for example, waiting time in emergency rooms, an estimate of that parameter could be the average waiting time for a small, carefully selected group of emergency rooms. The estimate would be a statistic. In general, we can know values of statistics but not parameters, even though we would wish to know the values of parameters.

Random and Nonrandom Samples

The samples, or subsets of a population, may be selected in a random manner or in a nonrandom manner. All patients in an emergency room on a specific afternoon would probably not constitute a *random sample* of all people who have used or will use an emergency room. All the hospitals in Buffalo, Syracuse, and Albany might be a random sample of all hospitals in New York state, but they probably would not be. All persons who received prescriptions for a cholesterol-reducing drug from a specific physician would, in general, not be a random sample of all persons who take such drugs. All of these examples would probably be considered nonrandom samples. Nonrandom samples may be drawn in many ways. In general, however, we are not interested in nonrandom samples. The study of statistics is based on and assumes the presence of random samples. This requires some discussion of what constitutes a random sample.

Random samples are drawn in a manner whereby every member of the population has a known probability of being selected.

A random sample is a sample drawn in a manner whereby every member of the population has a known probability of being selected. At a minimum, this means that all members of the population must be identifiable. Frequently, there is a gap between the population of interest and the population from which the sample is actually drawn. For example, suppose a health department wished to draw a random sample of all families in its area of responsibility to determine what proportion believed that the health department was a possible source of any type of health services—prevention, treatment, advice—for members of the family.

The *target population* is all families in the area of responsibility. If we assume that this is a county health department, a random sample would assign a known probability of selection to each family in the county. In general, this would mean that each family in the county would have an equal probability of selection. If there were, for example, 30,000 families in the county, each one would have a probability of 1/30,000 of being selected as a member of the sample.

But, in general, it would be very difficult to be certain that every family in the county had exactly a 1/30,000 probability of being selected for the sample. The difficulty arises from the problem of devising an economically feasible mechanism of identifying and contacting every possible family in the county. For example, one relatively inexpensive way of collecting the information desired would be to contact a sample of families by telephone and ask them questions. But some families do not have phones, making their probability of selection into the sample not 1/30,000 but simply zero. Other families may have more than one phone, and if care is not taken to ensure that the family is not contacted twice, some families might have twice the chance (or more) of being selected into the sample. Still other families—especially in the present age of telemarketing—would decline to participate, which would make their probability of being included zero as well.

Other mechanisms of identifying all families in a county or other area have equal difficulties. Voter rolls list only registered voters. Tax rolls list only persons who pay taxes. Both of these rolls also contain single persons. A decision would have to be made about whether a single person constitutes a family. In summary, then, it is very often nearly impossible, or at least very expensive, to draw a truly random sample from a target population. What often happens instead is that the sample drawn is actually from a population very close to the target population but not the target population exactly. Instead of all the families in the county being the population from which the sample is drawn, the population may be all families with telephones.

The sampled population is the population from which the sample is actually drawn.

The population from which the sample is actually drawn is known as the *sampled population*. Inferences from the sample to the population are always to the sampled population; one hopes that these inferences hold for the target population as well. Given that the population sampled may not be exactly the target population desired, there still needs to be a mechanism for assuring that each member of the population to be sampled has a known probability generally equal of being selected. There are lots of ways of assuring randomness in specific settings. Shuffling cards is a way of assuring that each person has an equal chance of getting the good cards and the bad cards during the deal—essentially a random distribution of the cards. Rolling dice is a way of ensuring a random distribution of the faces of a die. Flipping a coin is a way of ensuring the random appearance of a head or a tail.

But sampling from a population of all families served by a health department is more complicated. One workable mechanism might be to put every family's name on equal-sized slips of paper, put all the slips of paper in a box, shake the box, and without looking at the slips of paper, draw out the number of slips desired. But this approach, although it would produce a random sample of families, would be both cumbersome and time consuming. Excel provides several mechanisms that can be used to draw random samples.

There are basically four different types of random samples: systematic samples, simple random samples, stratified samples, and cluster samples.

Systematic Samples

Systematic samples are samples drawn by first dividing the population into subsets equal to the number of observations ultimately desired in the sample and then drawing a specific observation from each subset. If the total population of interest consisted of 30,000 families and the sample to be drawn was to consist of 100 families, the first step in drawing a systematic sample would be to divide the total population into 100 subsets. Suppose we have the 30,000 families on a list in alphabetical order. The common way to divide the families into 100 subsets would be to take the first 300 families as the first subset, the second 300 as the second subset, and so on to the end of the list.

A systematic sample is drawn from equally divided subsets of a population.

The next step in drawing a systematic sample would be to select randomly one family from the first subset of 300. Then the corresponding family from each of the other 99 subsets would be selected to fill out the sample. For example, if the family in position 137 in the alphabetical list were selected at random, then family number 437 (the 137th family in the second subset of 300) would be taken as the next member of the sample, and family number 737 would be selected as the third member, all the way to family number 39,837. This would produce a sample of 100 families, all of which are spaced 300 families apart in the alphabetically sorted list.

A systematic sample actually represents a single observation in statistical terms. This is because once the selection is made from the first subset, all other observations are fixed. If a sample of 100 is to be selected systematically from a population of 30,000, 300 different samples can be selected. These 300 samples correspond to each of the 300 families that can be selected as the first element of the sample. Because systematic samples are samples made up from a single random selection, the results expected from statistics do not actually apply to systematic samples. Nevertheless, systematic samples are often treated as if statistics do apply appropriately to them. This is generally considered acceptable for drawing inferences about populations.

Simple Random Samples

Simple random samples are samples drawn in such a way that every possible sample of a given size has an equal likelihood of being selected for the sample. If the total population of interest consisted of 30,000 families and the sample to be drawn was to consist of 100 families, every possible sample of 100 families would have an equal likelihood of being drawn in a simple random sample. Before the widespread availability of computers, simple random samples were typically drawn by associating each element of the population with a number from a random number table. If the number in the random number table was in a certain range, the element was included in the sample; if not, the element was not included in the sample. The advent of personal computers, and especially such programs as Excel, has eliminated the need for random number tables. Excel can generate lists of random numbers that can be used to select simple random samples.

Simple random samples assume that each sample has an equal likelihood of being selected.

Whereas only 300 different systematic samples of size 100 could be drawn from a population of 30,000 families, there are far more simple random samples of size 100. The number of different simple random samples of 100 families that can be taken from a population of 30,000 families is so large that it would take several lines of text to write it out completely. It is approximately the number 46,815 followed by 285 zeros. And each one of this very large number of samples has an equal likelihood of being selected as the one simple random sample taken.

Stratified Samples

Stratified samples are samples drawn by dividing the total population into two or more groups, or strata, and then drawing a specified proportion of each stratum for the sample. The specified portion might be proportional to the stratum size, or it might be equal to the number drawn from other strata, regardless of whether the stratum sizes are equal. Within each stratum, the sample may be drawn by simple random sampling or by systematic sampling. The sample is typically drawn by simple random sampling.

A stratified sample is drawn by dividing the population into strata and then drawing a specified proportion from each stratum.

Consider how a stratified sample might apply to our sample of 100 families from a population of 30,000 families. Suppose we know that within our population of 30,000 families, 3,000 have Hispanic surnames. If we want to draw a stratified sample that would guarantee that it had proportional representation of families with Hispanic surnames, we could first divide the total population into two strata: those with Hispanic surnames and those with other surnames. Then we could take a sample of 10 families from among those with Hispanic surnames and a sample of 90 families from among those who do not have Hispanic surnames.

In general, stratified samples are drawn for two reasons:

1. The researcher wishes to ensure that the groups represented by the strata are appropriately represented in the final sample.
2. There is reason to believe that the subject of interest to the study is closely related to the characteristics upon which the strata are based.

In the latter case, for example, if a health care professional wished to estimate average height among teenagers 17 to 19 years of age, it would probably be useful to stratify on sex. As teenagers, males are likely to be taller than females. At preadolescence, it might be useful to stratify on sex because females are likely to be taller than males.

Cluster Samples

Cluster samples are samples drawn by first dividing the sample into several groups, or clusters. The sampling then proceeds in two or more stages. This discussion is of a two-stage cluster sample only. In the first stage, a set of the clusters is drawn using either systematic or simple random sampling, although simple random is most commonly employed. In the second stage, either all members of the cluster or a sample of members of the cluster are selected to be included in the final sample.

Cluster samples are drawn via two stages: Groups are drawn first using random or systematic sampling techniques, and then samples are drawn from those groups.

In the case of our sample of 100 families from among 30,000, the families to be selected could be divided into ZIP code areas first. The sample of ZIP code areas randomly selected is the first stage. In the second stage, families could be selected randomly from those ZIP code groups selected to fill out the sample of 100. Typically, cluster samples are used when the collection of data from a simple random sample would involve a great deal of travel time. The use of clusters limits the travel required for data collection to only those clusters selected. A major drawback of cluster sampling is that it is likely to increase the variability of those statistics about which estimates are to be made.

Cluster Samples versus Stratified Samples

Cluster samples and stratified samples differ from one another. In using cluster samples, only a few of the groups or clusters actually have members represented in the final sample. When using stratified samples, all groups, or strata, have members represented in the final sample. For the material presented in this book, it is assumed that the data were drawn in what would be considered either a simple random method or a stratified method with the number of observations drawn from each stratum proportional to stratum size.

Variables, Independent and Dependent

Throughout this book there are frequent references to the term "variable." A *variable* is a characteristic of an observation or element of the sample that is assessed or measured. A value for a variable across all members of a sample (e.g., the average height of preadolescent teens) is typically referred to as a *statistic*. The comparable value for the population is a parameter. Most statistical activities are an attempt either to determine a value for a variable from a sample (and thus to be able to estimate the population value) or to determine whether there is a relationship between two or more variables.

A variable is a characteristic of an observation or an element of the sample that is assessed or measured.

In order to show a relationship between two or more variables, the variables must vary. That is to say that they must take on more than one value. Any characteristic of a population that does not vary is a constant. There can be no relationship between a constant and a variable. This is equivalent to saying that there can be no way of accounting for the value of any variable by referring to a constant. For example, if we wished to describe variations in adult-onset diabetes rates among persons with Hispanic surnames, it would be useless to employ Hispanic surname as an explanation, because it is constant for all these people. It cannot explain differences. But if we wished to describe differences in adult-onset diabetes among all the people living in, say, New Mexico, Hispanic surname or non-Hispanic surname might be a useful variable to employ.

Variables, Categorical and Numerical

Variables are typically classified as either of two types: categorical or numerical. Numerical variables are further classified as either discrete or continuous. These distinctions are important for the type of statistic that may effectively be employed with them.

Categorical variables are distinguished simply by name.

Categorical variables are variables whose levels are distinguished simply by names. Hispanic and non-Hispanic surname is a two-level categorical

variable that roughly distinguishes whether a person is of Hispanic ancestry. Sex is a two-level categorical variable that in general divides all persons into male or female. Categorical variables can take on more levels as well. Type of insurance coverage, for example, is a multilevel categorical variable that may take on the values Medicare, Medicaid, voluntary not-for-profit, for-profit, self-pay, and other. Other categorical variables may take on many levels.

Although a variable may be represented by a set of numbers, such a representation does not automatically mean that it is not a categorical variable. The ICD-9-CM code is a categorical variable, even though the codes are represented as numbers. The numbers simply classify diagnoses into numerical codes that have no actual numerical meaning. Another type of categorical variable that is assigned a number is the dummy variable. The dummy variable is a two-level categorical variable (e.g., sex) that is assigned a numerical value, usually the values 1 and 0. The value 1 might be assigned to female and 0 assigned to male, or vice versa. Although this type of variable remains a categorical variable, it can be treated as a numerical variable in some statistical applications that require numerical variables. However, a categorical variable with more than two levels, such as the ICD-9-CM code, can be treated as a numerical variable in analysis only by dividing the multilevel categorical variable into a number of two-level categorical variables that can be treated as dummy variables.

Numerical variables are distinguished by number.

Numerical variables are, as the name implies, variables whose values are designated by numbers. But unlike ICD-9-CM codes, the numbers have some meaning relative to one another. At the very minimum, a numerical variable whose value is, for example, 23 is presumed to be larger than a numerical variable whose value is 17. Numerical variables may be measured on three scales: the ordinal scale, the interval scale, and the ratio scale.

Ordinal Scale

The ordinal scale is a scale in which the values assigned to the levels of a variable simply indicate that the levels are in order of magnitude. A common ordinal scale is the *Likert scale*, which requests a response to one of usually five alternatives: strongly agree, agree, undecided, disagree, or strongly disagree. These responses are then assigned values of 1 to 5, or 5 to 1, and treated as values of a numerical variable. Treating Likert scale

responses like numerical variables assumes that the conceptual difference between strongly agree and agree is exactly the same, for example, as the conceptual difference between undecided and disagree. If that cannot be assumed, then ordered variables, such as Likert scale variables, even if assigned numerical values, should not be treated as numerical variables in analysis but must be treated as categorical variables.

Interval Scale

The *interval scale* is a scale in which the values assigned to the levels of a variable indicate the order of magnitude in equal intervals. The commonly employed measures of temperature, Fahrenheit and Celsius, are interval scales. For Celsius, for example, the value of 0 refers not to the complete absence of heat but simply to the temperature at which water freezes. One hundred on the Celsius scale refers to the temperature at which water boils at sea level. The distance between these two has been divided into 100 equal intervals. Because this is an interval scale measurement, it is accurate to say that the difference between 10 degrees Celsius and 15 degrees Celsius is the same as the distance between 20 degrees Celsius and 25 degrees Celsius. But it is not accurate to say that 20 degrees Celsius is twice as warm as 10 degrees Celsius.

Ratio Scale

The ratio scale is a scale in which the values assigned to the levels of a variable indicate both the order of magnitude and equal intervals but, in addition, assumes a real zero. The real zero represents the complete absence of the trait that is being measured. Temperature measured on the Kelvin scale has a real zero, which represents the complete absence of heat. At a more prosaic level, the number of patients in an emergency room is measured on a ratio scale. There can be zero patients in the emergency room, representing the complete absence of patients, or there can be any number of patients. Each new patient adds an equal increment to the number of patients in the emergency room. In general, any variable that is treated as numeric for statistical analysis is assumed to be measured on at least an interval scale and most commonly on a ratio scale.

Discrete and Continuous Numerical Variables

Discrete numerical variables are variables that can take on only whole number values. Discrete numerical variables are typically the result of the counting of things, persons, events, activities, and organizations. The number of persons in an emergency room is a discrete numerical variable.

There must always be a whole number of persons in the room—for example, 23. There can never be 23.7 persons in an emergency room. The number of children born to an unmarried woman, the number of organizations that are accredited by a national accrediting body, the number of physicians on a hospital staff, the number of health departments in a state—these are all discrete numerical variables.

Continuous numerical variables are variables that can take on any value whatsoever. They can be whole numbers, such as 47, or they can be numbers to any number of decimal places, such as one-third (which is 0.33333 . . . and so on forever). The amount of time that a person spends in an emergency room is a continuous variable that can be stated to any level of precision (in terms of minutes, seconds, parts of seconds) that we have the ability and interest to measure. Body temperature, pulse rate, height, weight, and age are all continuous numerical variables. Measures that are created as the ratio of one variable to another, such as cost per hospital admission or cost per day or proportion of children fully immunized, are also continuous numerical variables.

Probabilities of occurrence of discrete or continuous numerical variables cannot be found in the same ways. In general, it is possible to find the exact probability of the occurrence of a discrete outcome based either on an a priori distribution or on empirical information. Probabilities of outcomes for continuous numerical variables, however, can only be approximated. Despite this, the distribution of continuous numerical variables has been extensively researched and in the form of the normal distribution: Particularly, it forms the basis of most statistical analyses for numerical variables, whether the variables are measured as discrete or continuous.

EXERCISES FOR SECTION 1.4

1. For each of the following sets of entities, decide whether you think the set is more likely to be a population or a sample, and explain why.

- **a.** All hospitals in the United States
- **b.** All patients in the emergency room of a given hospital on a given day
- **c.** One hospital from each of the 50 largest cities in the United States
- **d.** All health departments in a given state
- **e.** The patients who visit a single physician
- **f.** Operating room procedures for February 11 in a single hospital

2. What mechanisms might you use to obtain a list of all members of the following target populations, and how successful might you be?

- **a.** Emergency room visitors for the past six months at a single hospital emergency room
- **b.** Hospitals in the United States
- **c.** Patients visiting a single health department
- **d.** All food service facilities in a health department catchment area
- **e.** All people in a single hospital catchment area
- **f.** People who will dial 911 in the next six months in a given municipality

3. Determine whether each of the following is a systematic sample, a simple random sample, a stratified sample, a cluster sample, or a nonrandom sample, and indicate why.

- **a.** A sample drawn by randomly selecting 50 pages from a telephone book and taking the fifth name on each page
- **b.** A sample drawn by dividing all persons visiting an emergency room in the past six months into male and female and randomly selecting 100 from each group
- **c.** A sample drawn by selecting the person who arrives at a doctor's office at a time closest to a randomly selected time of day (say, 9:10 A.M.) and each person coming closest to that time on 40 subsequent days
- **d.** Any five cards drawn from a well-shuffled deck
- **e.** A sample drawn by randomly selecting six health departments from among those in a state and then randomly selecting six staff members from each of the six health departments
- **f.** A sample taken by selecting 20 hospitals in such a way as to ensure that they are representative of the types of hospitals in the United States

4. For each of the following random variables, determine whether the variable is categorical or numerical. If the variable is numerical, determine whether the phenomenon of interest is discrete or continuous.

- **a.** The number of clients at a health department Maternal and Child Health (MCH) clinic
- **b.** The primary reason for an MCH clinic visit
- **c.** The length of time in minutes spent by a client waiting to be seen at the clinic
- **d.** Whether children at the clinic have the recommended immunizations
- **e.** The weight of children seen at the clinic
- **f.** The income of families of clients seen at the clinic

5. Determine whether each of the following scales is nominal, ordinal, interval, or ratio and indicate why.

- **a.** The classification of patients into male and female
- **b.** The designation of male patients as 0 and female patients as 1
- **c.** The number of live births to a woman coded 0, 1, 2, 3, or more
- **d.** The measured pulse rate
- **e.** The number of staff members in a health department

1.5 Five Statistical Tests

This book consists of two sections. The first section, which comprises the first six chapters, is essentially preparatory material designed to equip the user of the book for the second section. The first section includes this introduction, a chapter on the use of Excel for statistical analysis, a chapter on data acquisition and preparation for statistical analysis, a chapter on descriptive presentation of data and Excel's graphing capability, a chapter on probability, and a chapter on the examination of data distributions.

The second section of the book, which comprises Chapters 7 through 14, is concerned with hypothesis testing. Hypothesis testing is essentially the act of determining whether data from a sample can be seen to support or not support a belief in independence between two or more variables in a population; one is commonly considered a *dependent variable* and the other or others are thought of as *independent variables*.

Five separate statistical tests that address this question of independence are discussed in this book. They are the chi-square test, the *t* test, *analysis of variance (ANOVA)*, regression analysis, and *Logit*. In practical terms, each of these tests can be thought of as testing whether sample values for two or more variables could have been drawn from a population in which the variables are independent of one another. Without considering specifically what independence means in regard to any one of these tests, it is possible to distinguish between these five tests on the basis of the type of data for which they are able to assess independence.

Chi-Square Test

The chi-square test can be used to assess the independence of two variables, both of which are categorical. Either of the two variables may take on two

or more levels or categories, but the data itself are measured simply as named categories. For example, the chi-square can be used to determine whether coming to an emergency clinic for a true emergency or for a visit that is not an emergency (a two-level categorical variable) is independent of whether one comes during the day or during the night (another two-level categorical variable). Or a chi-square test could be used to determine whether the desire of women for an additional child (a two-level, yes–no variable) is independent of the number of children she already has in the three categories—for example, one, two or three, and four or more. The chi-square can be used on variables that take on larger numbers of values as well, but in practical terms, it is unusual to see a chi-square that involves variables having more than three or four levels.

t Test

The *t* test can be used to assess the independence of two variables. The *t* test assumes that one variable is a numerical variable measured in either discrete or continuous units and the other is a categorical variable taking on only two values. For example, the *t* test can be used to determine whether the score people receive on a test of knowledge about breast cancer measured on a 20-point scale (a numerical variable) is independent of whether those people were specifically and consciously exposed to knowledge about breast cancer or were not (a two-value categorical variable). Or a *t* test could be used to determine whether the cost of a hospital stay (a numerical variable) was independent of whether the patient was a member of an HMO or was not (a two-level categorical variable).

Analysis of Variance

Analysis of variance, or ANOVA, an extension of the *t* test, can be used to assess the independence of two or more variables. ANOVA assumes that one variable is a numerical variable measured in either discrete or continuous units and the others are categorical variables that may take on any number of values rather than only two. ANOVA, for example, could be used to assess not only whether a knowledge score about breast cancer was independent of exposure to knowledge about breast cancer but also whether the score might be independent of several different types of exposure. Exposure could be the reading of a brochure, a one-on-one discussion with a physician, both, or neither. Analysis of variance could also be used to determine whether the length of a hospital stay (a numerical variable) was independent of the hospital in which the stay took place over five separate hospitals (a categorical variable taking on five values).

Regression

Regression, a logical last stage in this progression, is a technique that can test the independence of two or more numerical variables measured in either discrete or continuous units. Regression may also include one or more categorical variables, any one of which can take on only two values (in which case, it is often referred to as analysis of covariance). Regression, then, could test the independence, for example, of the cost of a hospital stay (a numerical variable) and the length of a hospital stay (a second numerical variable) across an essentially unlimited number of hospitals. Or it could assess the independence of the dollar value of all hospital billings (a numerical variable) and the number of patients admitted (a second numerical variable) for a sample of for-profit and not-for-profit hospitals (a categorical variable taking on two values).

Logit

Logit, an extension of regression, can examine the independence of two or more variables where the dependent variable is a dichotomous categorical variable and the independent variable or variable set may be categorical (taking on only two values) or numerical—either discrete or continuous. Logit could be used, for example, to assess the independence of the outcome of an emergency surgical procedure (measured as successful or unsuccessful) and such variables as the degree of presurgery trauma, the length of time between the emergency and the surgical procedure, the age of the patient, and so on.

EXERCISES FOR SECTION 1.5

1. Consider which type of analysis could be used to assess independence for each of the following sets of data and state why this would be so (the dependent variable is given first).

- **a.** Hospital length of stay per admission and insurance type, including Medicare, Medicaid, private not-for-profit, private for-profit, and self-pay
- **b.** Cost per hospital stay and sex, age, and whether medical or surgical
- **c.** Whether a woman desires an additional child and the number of children now living categorized as none, one or two, and three or more
- **d.** Blood pressure readings for a group of people before and after the initiation of a six-week exercise and diet regimen

e. Hospital length of stay per admission for the first digit of the ICD-9/ICD-10 code

f. Birth weight for newborns measured as low or normal and gestational age, mother's age, and whether she is a smoker or a nonsmoker

2. Suggest a dependent variable and at least one independent variable for a question that could be analyzed using each of the following:

a. Chi-square analysis

b. A *t* test

c. Analysis of variance

d. Regression

e. Logit

KEY TERMS

analysis of variance (ANOVA)
categorical variable
chi-square statistic
cluster sample
continuous numerical variable
dependent variable
discrete numerical variable
dummy variable
independent variable
interval scale
Likert scale
Logit
Medicare
numerical variable
parameter
population
random sample
sample
sampled population
simple random sample
spreadsheet
statistic
stratified sample
systematic sample
t test
target population
variable