CHAPTER 1 Introduction to Statistics[R]

OBJECTIVES

1. Describe the importance of statistics in health administration.
2. Distinguish between a sample and a population or a universe.
3. Differentiate between a finite and an infinite population.
4. Describe the differences between descriptive and inferential statistics.
5. Differentiate between a statistic and a parameter.
6. Describe the sources of information that are available to the health administrator.

The Role of Statistics in Health Administration

Statistics is a valuable area of applied mathematics for administrators who are developing strategic plans, assessing alternate plans, and evaluating the operation of the health service organization under conditions of uncertainty. Here *statistics* is referred to as the assembly of information, the analysis of data, and the interpretation of results within an administrative or managerial setting. It emphasizes the application of statistical analyses to problems or issues that exist in health service organizations. Specifically, the focus is on the usefulness of statistical analysis in evaluating the internal environment, identifying or solving problems, reducing deficiencies in performance, developing institutional policies, and making decisions that influence not only daily operations but also the strategic position of the health organization.

The importance of statistics in health administration is demonstrated by first considering the problem of monitoring, evaluating, and controlling internal operations. Figure 1.1 is based on the idea that unidentified problems are unlikely to be resolved. It assumes that comparing or contrasting the actual results of activity with a standard of performance identifies the problem. An unfavorable or undesirable difference suggests that a problem exists and that a remedial plan needs to be applied. After initiating the plan, measuring its effects, and assessing the results, the administrator may decide that the problem is resolved and that no further action is required. In contrast, the comparison or contrast may indicate that the remedial plan was only partially successful or that it created additional problems.

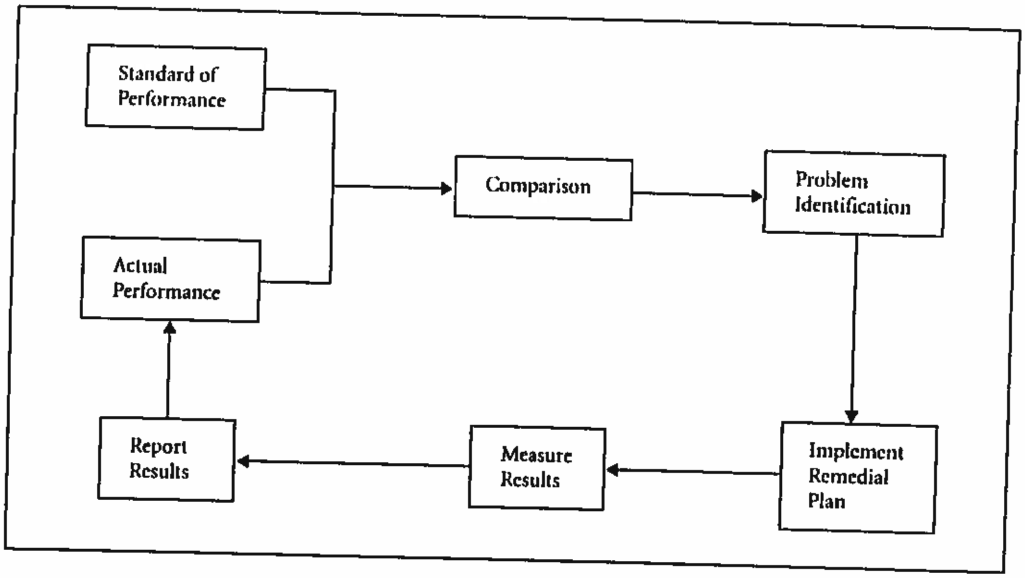

Figure 1.1 **The Role of Statistics in the Internal Control System**

Statistical analyses help to identify problems. Specifically, the standard of performance frequently represents a projection of desired results and might be developed using standard statistical techniques that describe the distributional properties of data that depict the clinical, operational, or financial performance of the organization. For example, the standard of performance might be expressed as the average length of stay or the use of ancillary care by the typical patient assigned to a given diagnostic group or risk category. Obviously, the development of treatment protocols is valuable when evaluating the quality of care, resource use, and related costs. Health organizations must monitor and control costs. A statistical estimate of the relationship between cost and volume enables the institution to set limits on acceptable expenses, to identify situations in which excessive costs were incurred, to isolate the causes, or to implement a remedial plan. Furthermore, simple contrasts between the actual performance and the standard of performance might identify the excessive use of service or resources, outcomes that might compromise the fiscal well-being of the health service organization.

In addition to problem identification and resolution, statistical analyses might be applied to the process of strategic management. Specifically, the situational analysis focuses on the internal and external environment of the health service organization. Evaluating the external environment identifies threats and opportunities. In contrast, when evaluating the internal environment, strengths and weaknesses of the organization should be identified. Focusing on human resources, the clinical function, the administrative function, and information technology, a weakness is an organizational attribute that increases the adverse effects of an external threat or prevents the organization from taking advantage of an opportunity. A strength is a characteristic that insulates the organization from the effects of threats or enhances the organization's ability to take advantage of opportunities that exist or are emerging in the external environment. Accordingly, statistical analysis might be used to

1. Assure the flow of resources, such as labor, supplies, and technology, that are required to provide care
2. Evaluate alternate plans that the administrator might adopt
3. Evaluate new methods of delivering health services or the need to alter the organization's structure
4. Assess the efficiency and effectiveness of the organization in its operating activity
5. Assure the quality of care
6. Ensure the fiscal well-being of the facility in the short term and the adequacy of resources in the long term

This text is intended to explore those statistical techniques that play an increasingly important role in monitoring, evaluating, controlling, and improving the performance of the health service organization.

Basic Definitions

Statistics is used in a variety of contexts, with each having a specific meaning. First, when used in its commonly accepted sense, statistics refers to numeric information and is used when referring to any area that might be expressed numerically. For example, the attributes of an enrolled population, a population at risk, or the daily census might be expressed numerically and might represent the common interpretation of the term statistics. Second, statistics also is used when referring to a body of techniques that are applied in the process of collecting, analyzing, interpreting, and presenting numeric information.

Finally, *statistic* refers to a single value that is derived from a sample selected from a universe or population of interest. A *universe* or *population* is defined as all observations or all theoretically conceivable observations concerning a phenomenon of interest. Populations or universes are further described as finite or infinite. A *finite population* is a universe for which it is possible to list all elements. For example, suppose an inpatient census for a given day represents the population of interest. In this case, it is possible to list all elements or patients that comprise the universe. In contrast, an *infinite population* is unbounded and usually results from the operation of a process under constant conditions. For example, the population of interest might consist of all previous emergency room visits and all of those that might occur in the future. It is not possible to list the elements of an infinite population, and as a result, data derived from a sample of the universe are required to examine the issue of administrative interest.

A *sample* is a subset of the population or universe of interest and conveys information that is of administrative usefulness. Based on the data derived from a sample, we might calculate a statistic that is valuable to the administrator. For example, based on a subset of the appropriate population, the administrator frequently is interested in a statistic such as the average length of stay, the average number of admissions, the average number of visits to the emergency room, or the average number of ancillary services used by patients who are grouped by diagnostic or risk category.

In this text, a sample is differentiated from a population in several ways. First, *parameter* is used when referring to a population, and *statistic* is used when referring to a sample. Furthermore, Greek letters, such as β, μ or σ, are used to identify the parameters of the population. Conversely, statistics derived from samples are represented by letters such as S, r, or b and are used as estimates of their corresponding parameters.

An application of statistical methods requires an understanding of the term *variable*. As indicated previously here, the phenomenon that is of interest to the health administrator exhibits variability or variation. For example, suppose the organization selected a sample of 10 adult discharges and recorded their ages as 23, 45, 34, 35, 67, 54, 64, 43, 32, and 60 years old. In this case, the variable age exhibits variation among the observations. A factor that exhibits variability or volatility or assumes different values is called a variable.

Variables might be classified in terms of their quantitative or qualitative attributes. A *quantitative variable* measures outcomes that are expressed numerically. Examples of quantitative variables include the patient's age, the percentage of outstanding receivables that are recognized as a bad debt, and the days of care the organization provided during a well-specified period. A *qualitative variable* consists of outcomes that without modification are not expressed numerically. Patient satisfaction, represented by very satisfied, satisfied, dissatisfied, and very dissatisfied, and an evaluation of departmental performance, measured by good, fair, or poor, are examples of a qualitative variable. When the administrator applies statistical methods or techniques, the focus of the analysis is on one or more variables.

We frequently apply statistical analyses to problems that involve components of direct patient care, namely, stay-specific and ancillary services. Stay-specific services are components of care that are measured by the number of days of service provided to a patient, to a group of patients, or during a given period. Similarly, we define ancillary services as components of care that are measured by the number of procedures provided to a patient, to a group of patients, or during a specific period. Accordingly, ancillary services are represented by the set of procedures that is provided by the laboratory or radiology and are measured in units of service.

Descriptive and Inferential Statistics

Figure 1.2 presents an overview of statistics, which consists of two distinct branches: descriptive and inferential. The descriptive branch of statistics consists of methods or techniques that are used to rearrange data into a more useful form and methods that are used to measure the properties of a distribution that depicts a set of data. In contrast, the inferential branch of statistics consists of methods that enable the administrator to develop generalizations, estimations, or predictions concerning the phenomenon of managerial interest.

Descriptive Statistics: An Introduction

Descriptive statistics is occasionally limited to methods that might be adopted to describe or present data in a table or as a graph. To illustrate the usefulness of descriptive statistics, consider a set of data that depicts the use of stay-specific services as indicated by the length of stay experienced by 100 patients. In this case, we could present the length of stay for each of the 100 patients in a more useful form.

Assume that the administrator is interested in grouping patients in terms of hospital episodes that extended from 1 to 5 days, 6 to 10 days, 11 to 15 days, and more than 15 days. Adopting these categories, the data depicting the hospital stays of the 100 patients might be presented as the frequency distribution (Table 1.1).

The first column of the table shows the categories that define the frequency distribution. The frequencies in the second column correspond to the number of patients assigned to each group. The distribution indicates that 32 of the 100 patients experienced a 1- to 5-day length of stay.

The relative frequency distribution appears in the final column and indicates the proportion or percentage of patients assigned to each of the categories. These data indicate that 32% of the patients experienced a 1- to 5-day hospital stay. Rather than presenting a list of 100 patients and their corresponding lengths of stay, the frequency distribution is a convenient device that summarizes data and presents numeric information in more useable form.

In addition to presenting information as a graph or table, descriptive statistics, as presented in this text, also includes any treatment of data that does not involve generalizations, predictions, or estimations. Once generalizations, estimations, or predictions are involved, the analysis is inferential rather than descriptive. Hence, inferential statistics allows the administrator to rely on predictions, estimations, and generalizations when developing, implementing, or evaluating decisions.

The distinction between descriptive and inferential statistics might be shown with this simple example. Suppose the administrator is interested in the use of stay-specific services, measured by the length of stay, by patients presenting with diseases of the digestive system, and by patients presenting with diseases of the circulatory system. Five medical records were selected from individuals who were diagnosed with a digestive system disease. The lengths of stay were 4, 6, 7, 10, and

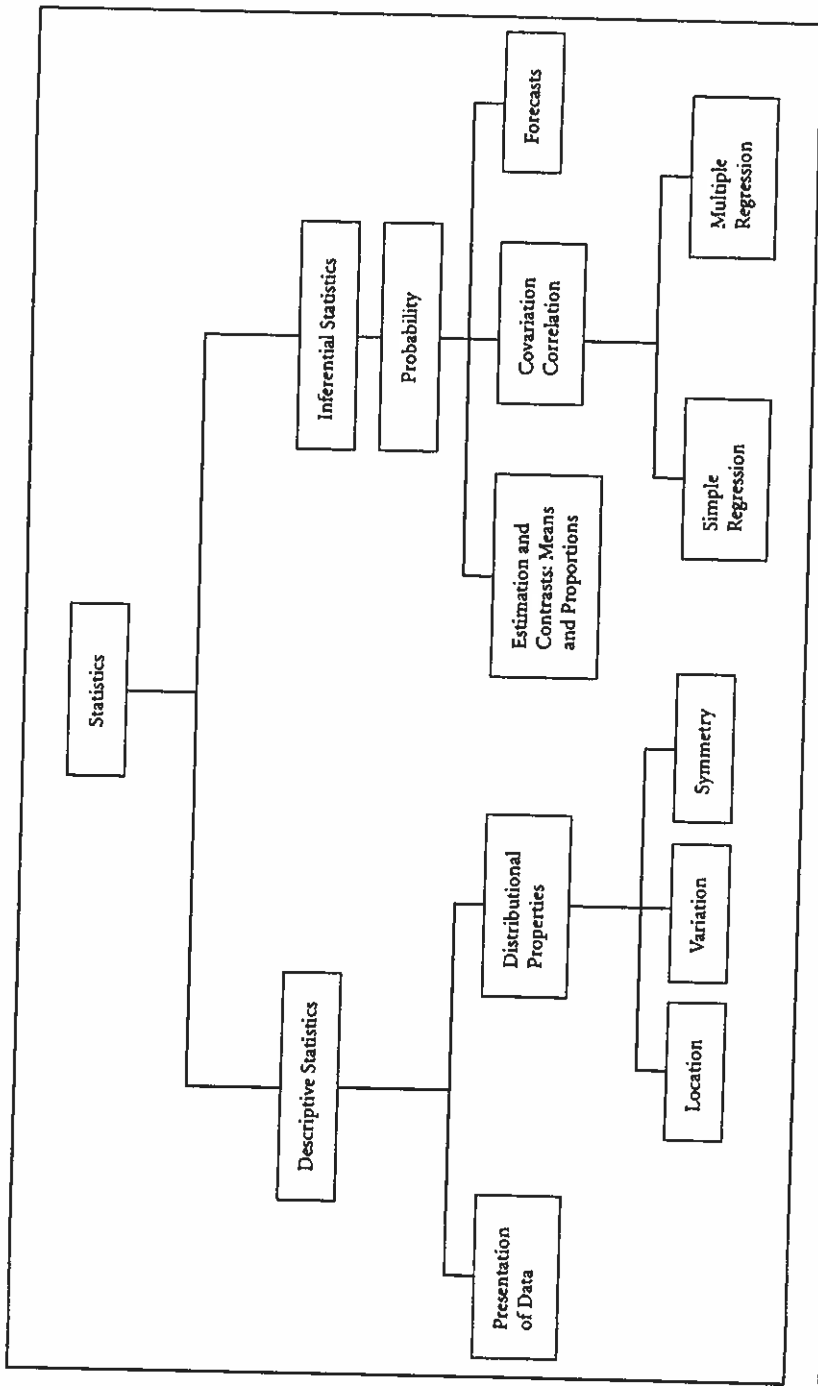

Figure 1.2 The Components of Basic Statistics

Table 1.1 **Frequency and Relative Frequency Distribution of 100 Patients**

Length of Stay	Frequency	Relative Frequency
1 to 5	32	0.32
6 to 10	43	0.43
11 to 15	15	0.15
More than 15	10	0.10
Total	100	1.00

13 days. The medical records for five patients presenting with a circulatory system disease were also chosen. Their corresponding lengths of stay were 5, 7, 9, 10, and 14 days. Based on these data, patients with a digestive system disease experienced an average stay of

$$(4 + 6 + 7 + 10 + 13)/5$$

or 8 days. Similarly, the average stay for patients with a disease of the circulatory system was

$$(5 + 7 + 9 + 10 + 14)/5$$

or 9 days.

Thus far, we used simple arithmetic rules to calculate a single value for each set of data. The resulting averages are descriptive of the data and contribute nothing to the original information. Rather, the calculations simply rearranged the data and enabled us to present the information in a more useable form. Accordingly, the mathematical operations described thus far are descriptive rather than inferential statistics.

Inferential Statistics: Their Role in Health Administration

Based on the previous calculations, suppose that the analyst concluded that patients with a circulatory system disease remained in the hospital longer than individuals diagnosed with a digestive system disease. In this case, the analysis and related conclusion far exceed the information originally provided. When we conclude that one group of patients experienced a longer length of stay than another, we move from descriptive statistics to inferential statistics.

Referring to the example, the difference between the two averages does not necessarily imply that patients with a circulatory system disease remain in the hospital longer than individuals with a digestive system disease. The original data reveal that several patients with a digestive system disease experienced a longer stay than those who were diagnosed with a circulatory system disease. In addition, considerable differences exist among the hospital episodes experienced by patients assigned to the same diagnostic category. It also is possible that a second sample might have results that are reversed. The observed difference between the two averages, given by 9 – 8 (or 1 day), might be attributable to chance and chance alone.

Before concluding that the observed difference implies that one group remained longer than another, several issues must be resolved. First, if the observed difference was very large and without a detailed statistical analysis, we are tempted to conclude that patients with a circulatory system disease experienced a longer length of stay than those with a digestive system disease. In contrast, if the difference is very small, we are tempted to conclude that the two groups experienced the same length of stay. Unfortunately, when expressed in natural or epidemiologic units, the terms *very small* and *very large* lack precision and may vary. Before concluding that one average is larger or smaller than another, it is necessary to develop the criteria that enable us to conclude that a difference is very large or small.

A second and related issue that influences the conclusion concerning the comparison or contrast involves the controls that were applied when the data were selected. It is possible that the process of selecting the data may prevent a meaningful conclusion regarding the observed difference. For example, it is possible that patients with a circulatory system disease presented more secondary conditions than those with a digestive system disease. Suppose further that the sample of patients with a circulatory system disease consisted of older individuals, whereas those in the other group were relatively young. However, advancing age is frequently accompanied by a decline in the individual's health and ability to recuperate. In this situation, the observed difference may be related more to the age distribution of the two samples than to the presenting diagnosis. The discharge status of the patients included in the two samples is of importance to inferences regarding the difference between the two averages. Assume, for example, that several of the patients with a digestive system disease died during the hospital episode, whereas all individuals with a circulatory system disease were discharged alive. Thus, the observed difference between the two averages may be attributable more to the patients' discharge status than to the individuals' diagnosis.

The discussion implies that the statistical treatment of data requires more than collecting data, performing calculations, observing differences, and reaching a conclusion. Rather, the method of assembling information and the application of an appropriate research design are of primary importance when analyzing data and reaching valid conclusions or deriving generalizations. An appropriate application of inferential statistics is only one of several essential ingredients in assessing numerical information, reaching valid conclusions, and implementing policies that are designed to improve the performance of the health service organization.

A Statistical Overview

Figure 1.2 shows that statistics consists of two branches: descriptive and inferential. The descriptive branch consists of two main components. As described by the length of stay example, the first consists of techniques that enable the administrator to rearrange data and present numeric information in a table or graph. The second focuses on the methods that are used to measure or describe a distribution. This text focuses on measures that are used to describe the location, scatter or dispersion, skewness, or lack of symmetry associated with a set of data.

The second branch, inferential statistics, enables the administrator to evaluate numerical data, interpret results under conditions of uncertainty, and develop generalizations, predictions, and estimations. The concept of probability forms the foundation for understanding inferential statistics (Figure 1.2). Before exploring inferential statistics, we examine the basic concepts of probability, several theoretical probability distributions, and the sampling distribution of the mean. A major objective of inferential statistics is to use data assembled from a sample to derive generalizations concerning a parameter of the corresponding population or universe. The focus on probability provides the foundation for not only the development of generalizations under conditions of uncertainty but also the interpretation of theoretical sampling distributions and theoretical probability distributions. Theoretical probability distributions such as the standard normal and the t-distribution enable us to develop inferences concerning the mean, the proportion, and statistics derived by regression analysis (Chapter 5). Similarly, Chapters 7 and 8 indicate that an assessment of differences among multiple proportions and means requires an understanding of the χ^2 and F distributions, respectively.

Figure 1.2 also indicates that covariation and correlation analysis are used to examine the direction of association or relationship in a *bivariate analysis*, defined as one that focuses on only two variables. Covariation indicates the direction of association between two variables and is used to identify a direct or inverse relationship. A direct or positive relationship occurs if the value of one variable increases as the value of the other grows. In contrast, in an inverse or negative relationship, the value of one variable grows while the value of the other declines. Correlation analysis enables us to identify not only the direction of relationship but also the strength of the association that exists in a bivariate analysis.

In turn, covariation and correlation form the basis for understanding and using regression analysis to identify or resolve problems or issues that occur in the health service organization. The purpose of simple regression analysis is to establish the relationship between the dependent variable, representing the administrative interest, and a single independent variable, a factor that is believed to influence the value of the dependent variable. For example, the cost of care is of primary concern to the health administrator and, therefore, might represent the item of interest or the dependent variable. In the short term, it is possible to argue that the volume of service is a primary factor that influences cost, and as measured by days of care or the number of visits, volume might represent the independent variable. In the simplified example, simple regression analysis is an approach that allows the administrator to capture the relationship between cost and the volume of service.

Multiple regression analysis is an extension of its simple counterpart. When multiple regression is performed, the number of independent variables is two or more, enabling us to incorporate several factors in an examination of an administrative phenomenon requiring explanation or prediction. For example, the use of care is obviously important to the typical health service organization. In this case, however, the factors that influence the use of care include the patient's age, gender, insurance status, health status, exposure to lifestyle, occupational or environmental risks, social or racial status, and income. Hence, the inclusion of several independent variables in the analysis allows the administrator to explain the phenomenon of interest or to derive predictions with greater precision.

The text concludes with a discussion of methods that might be used to develop projections or forecasts (Figure 1.2). The development of accurate predictions is fundamentally important to the preparation of the organization's budget and the development of operating plans. The development of projections frequently requires a reliance on *time series data*, defined as information that is recorded and reported at equal intervals. The hospital's daily census, financial statements, and monthly operating reports are examples of time series data. Time series data that are recorded annually may exhibit an increasing or decreasing trend, represented by a long-term rise or fall, respectively. Simple regression analysis might be adopted to estimate trend values or the value of the dependent variable for a given year. Similarly, the time series might exhibit seasonal variation. Seasonal factors result in recurring patterns of variation that occur during the span of 1 year. In such a situation, the method of seasonal decomposition might be applied to the general problem of projecting trend and adjusting estimates for seasonal effects. However, an application of seasonal decomposition to the general problem of developing forecasts requires the use of multiple regression analysis.

In summary, statistics consists of two branches: descriptive and inferential statistics. The primary objective of descriptive statistics is to rearrange one or more sets of numeric information into a more useable form and to describe the properties of the distribution. Inferential statistics is based on theoretical probability and sampling distributions. The primary purpose of inferential statistics is to develop generalizations or estimates concerning the mean or the proportion and to develop indicators of relationship between a dependent variable and one or more independent variables.

Sources of Information

As indicated previously here, numeric information is an essential ingredient in any statistical analysis. Thus, before an application of statistical methods, the analyst must evaluate the relevance, reliability, and validity of data. This section reviews various sources of information that are available to the administrator.

The Internal Information System

The information system that the typical health service organization maintains is an invaluable source of data. In general, the data derived from the information system portray financial, clinical, and operational activity. As such, the information assembled and maintained by the organization is an essential ingredient in the process of planning, monitoring, evaluating, and controlling internal affairs.

For example, the financial reporting system of the organization assembles the fiscal data that are communicated to those who use the information to evaluate the organization and to develop or implement decisions that influence operational activity. The two most important types of financial statements are as follows: The first, represented by the income statement or statement of changes in financial position, summarizes the revenues, expenses, and the net surplus or loss for the period. The income statement is a summary of the resources that the health service organization generates internally. The second report is the balance sheet or the statement of financial position. The balance sheet summarizes the assets or resources of the organization, the obligations of the organization to external entities, and residual ownership or equity at the end of the accounting period. In addition, the typical health service organization also generates reports that show the receipt of supplies, payments to those who provide the resources that are required in operations, and reports that depict outstanding receivables, accounts payable, and capital expenses.

The financial statements are particularly valuable when evaluating periodic changes in four domains of fiscal performance. In particular, the income statement allows the administrator to assess changes in profitability, which refers to the organization's ability to earn revenues in excess of expenses and thereby contribute to internally generated funds that are required to acquire necessary resources. The second domain is debt structure, which measures the organization's relative dependence on external credit sources to acquire assets or resources. The third is liquidity, which is the ability of the organization to honor its currently maturing obligations to outside sources. Finally, the fiscal statements might be used to evaluate the intensity of asset use, which is the amount of revenue earned by providing care relative to the assets of the organization.

In addition to the financial reporting system, the typical health service organization also maintains data that measure clinical and operating activity. The clinical information system assembles data that depict the organization's provision of stay-specific and ancillary care, grouped by patient, diagnostic category, risk group, or departmental unit. For example, the days of care provided to a given patient or patients assigned to a diagnostic group or risk category might enable the analyst to assess the effectiveness or quality of service. Similarly, data that measure the intensity of providing outpatient care or laboratory, radiologic, and rehabilitative services are also of value when assessing the effectiveness of clinical activity.

Information systems that measure operating activity routinely assemble data that depict resource use, measured in real or physical terms. Data depicting the use of resources (such as labor and supplies) that are consumed in the provision of stay-specific or ancillary care are valuable. For example, the use of labor, grouped by occupational category, and consumable supplies, classified by the type of item, to provide each service in the set of procedures offered by the laboratory forms the basis for assessing the efficiency of the unit. When combined with the volume of each procedure, data that measure operating activity allow the administrator to monitor, evaluate, and control the efficiency of operations, as indicated by the consumption of resources per unit of service provided.

Recurrent reports and related statistical analyses may identify a problem or weakness that requires the implementation of a remedial action or policy (Figure 1.1). When a problem or weakness is identified, information may need to be assembled, and after appropriate analysis, a special or nonrecurrent report may need to be prepared so that the organization can address the issue. As a consequence, the data derived from the recurrent reporting system are supplemented by numeric information that enables the administrator to assess the specific problem in detail.

External Sources of Data

In addition to the information that the health service organization gathers, data from external entities may be required to resolve specific problems or to develop policies that are designed to reduce or eliminate weaknesses. For example, data files available to the public assemble information that depicts the prevalence or incidence of mortality and morbidity in a well-defined geographic area such as the county. Public entities, such as state departments of health, publish data that describe the use of inpatient and ambulatory care, grouped by diagnostic classification. Data depicting charges, measures of case complexity or severity, and patient attributes also are frequently available in these sources. Data assembled by external sources might be applied in a comparative analysis that is designed to assess the performance of the organization relative to similar facilities.

Before adopting data derived from external entities, the validity, reliability, applicability, and continuity of the information need to be assessed, a process that requires a distinction between primary and secondary sources. A ***primary source*** is an organization or entity that not only collects information but also publishes related data. A ***secondary source*** is an organization or entity that publishes data that another agency or institution collects.

Data obtained from primary sources may be more reliable than information available from secondary sources. In addition, data derived from primary sources also are more complete than the data from secondary sources. In particular, detailed or disaggregated data appearing in a primary source may be eliminated when presented in a secondary source. Similarly, details regarding variable descriptions, definitions, and methods of collection that appear in a primary source may be omitted or presented in condensed format when the information is presented in a secondary source. Finally, it is possible that data presented in a primary source may be subject to clerical error when the information is transcribed. Hence, information available from primary sources is likely to be more accurate or reliable than similar data appearing in a secondary source.

Occasionally, the health administrator may require data that are not available from internal or external sources. In such a situation, the administrator might rely on a survey or an experiment to obtain needed information.

Surveys

A survey is a process of collecting information without effective control over factors or variables that might influence population characteristics or measures of the item of interest. Most are familiar with surveys and the process of completing a questionnaire or responding to the questions of an interviewer. Suppose the director of human resources is interested in obtaining information concerning the salaries of 2,000 registered nurses in a given market area. In this case, the characteristic of interest, salaries, may be influenced by age, gender, and years of experience. When collecting the information, however, the administrator has no control over the age, gender, or years of experience of respondents because these are individual attributes.

Surveys are frequently used to collect or assemble data. For example, the health service organization might require information that depicts the health status of families residing in the service area or the perception of potential patients regarding the prestige of the facility. A survey may be self-administered, implying that the respondent completes the survey by addressing each item appearing on a prearranged questionnaire. In contrast, a trained interviewer may administer the survey, ask a series of questions that appears on the interview schedule, and record the respondent's answers. Generally, the interview schedule contains instructions that designate the question order, the interpretation of responses, and the method of recording the respondent's answers. The reliance on a trained interviewer usually increases the proportion of surveys that are usable and lowers the tendency of respondents to misinterpret questions. On the other hand, the interviewer may distort responses to the survey. For example, the interviewer may ignore directions for selecting respondents, an outcome that may bias statistical analyses and related results. In addition, the interviewer may influence responses by the method of asking questions or an inadvertent gesture.

Surveys frequently generate required information that is unavailable from internal sources. In these circumstances, the survey may assemble data based on a census or a sample. As previously mentioned, the totality of the phenomenon is referred to as the population or universe. A *sampling frame* lists all of the elements or members of the population. For example, suppose that the director of human relations is interested in the morale of employees who provide outpatient care and develops a survey that is designed to assess their satisfaction with working conditions. In this case, the sampling frame consists of those employed in the emergency room or in ambulatory clinics. In turn, employees who are included in the survey might be selected from the sampling frame.

A survey that includes all members of the population is called a census. In general, the reliance on a census to collect information is expensive and time consuming, features that reduce the desirability of including all members of the population in a survey. For example, the administrator might be interested in the risks present in an insured population. In such a situation, it may not be possible to contact all individuals and assess their health status. Instead, a sample of the insured population might be selected, and based on related data, the administrator is able to develop inferences concerning the health status of those included in the insured population.

A sample is frequently preferred to a census for several reasons. First, a sample can provide valid and reliable data at less cost than a census. In addition, a sample is likely to provide data on a timelier basis than a census because a smaller amount of information must be collected and analyzed. Furthermore, interviewer bias, typographic errors, and clerical errors are typically lower in a sample than in a census. A sample is a smaller undertaking than a census, and the opportunity to ensure the accuracy of data is correspondingly greater.

Although several advantages accompany the decision to administer a survey to a sample of the population, a complete census may be necessary and desirable. Specifically, if the finite population is small and complete information is required, a census is preferred. Thus, the relative advantage of a sample or a census depends on the size of the population and the nature of the administrative issue or problem.

Experiments

Unlike a survey, an experiment collects information when control is exercised over some or all of the variables that might influence the population characteristic or item of interest. For example, the administrator might be interested in examining whether training in dental hygiene reduces the incidence of caries, gingivitis, or periodontal disease. In this case, we might divide patients into two groups that are similar in terms of age, gender, education, and other factors that might influence the effectiveness of the program. We then provide training to one of the groups (called the experimental group) and let the other group (called the control group) continue without training. Because the two groups are similar in terms of age, gender, education, and so forth, differences in the presence of dental disease might be attributed to the training program.