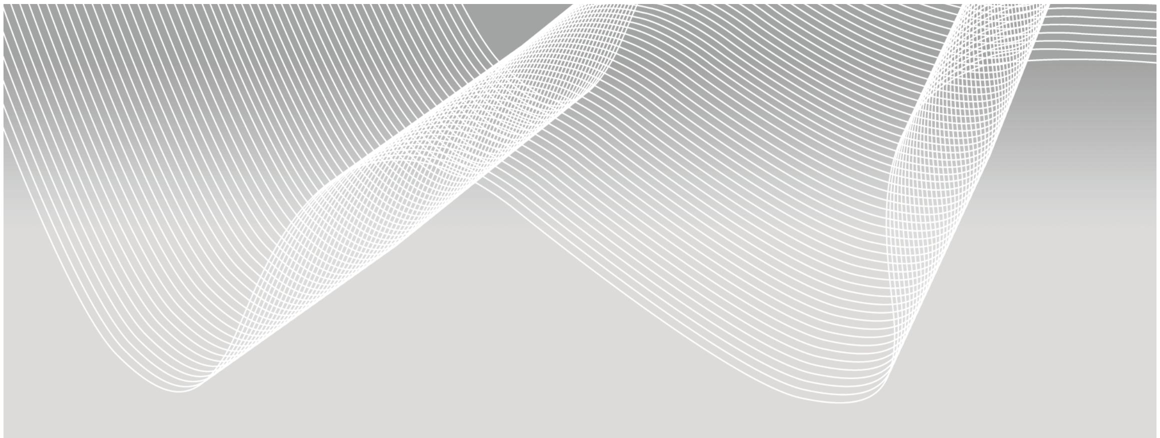
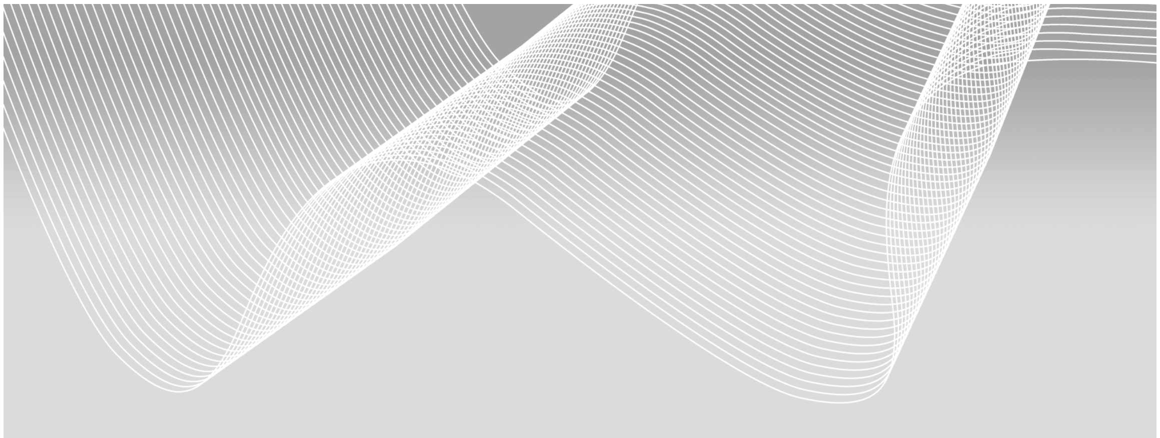


Part 1

Basic Foundation



Chapter 1

Introduction to Statistics

Why Study Statistics?

With the advent of the Internet and new techniques for gathering data, the amount of information in the modern business enterprise is literally exploding. Despite the potential value of this information, most organizations are still struggling to use this information in any meaningful way. Some companies, such as AT&T, are turning to more sophisticated statistical analysis of this information to try and uncover patterns in the data that will give them a competitive edge. In many ways, this new use of statistical techniques is fundamentally different from the historical use of statistical analysis in business organizations. The emphasis is more on exploration and discovery than on the more traditional hypothesizing and confirming in the past. Newer software, such as JMP software used in this text, has been specifically designed for exploration and discovery. Nonetheless, the traditional topics of statistical inference still play an important role in the discovery process as we will discuss in a later section and at various points in the text.

Statistics is also part of the growing field of *business analytics*. Analytics derives from the Greek word "analutiká" pertaining to mathematical or logical analysis. The term was widely adopted by Internet companies such as Google.[1] In the IT (information technology) realm, many software companies such as SAS, SPSS, and Oracle promote their analytics products under the broad umbrella of *business intelligence* (BI). However, our use of the term here is much broader than tracking Web statistics or even statistical analysis. A recent book on the topic defined analytics as "the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions." (Davenport and Harris, 2007.) Companies cited in the book are as diverse as Harrah's Entertainment Inc., Netflix, and the Boston Red Sox baseball team. There is even a free digital magazine on analytics sponsored by Informs (Institute for Operations Research and the Management Sciences).[2]

From the above discussion, it is apparent that statistical analysis has become an increasingly important tool in the business arsenal. As Davenport and Harris put it, "Organizations (and people) that make good use of statistical methods are likely to be more successful than those which do not."[3] That is why it is so important for business leaders of the future to have a basic fundamental knowledge of statistical reasoning. Although you may not be "doing statistics" in your future jobs, at the very least, you need to be able to understand enough to converse with those who are. However, statistics can be a challenging subject for students. We will try to help you feel comfortable with the subject by avoiding unnecessary mathematical complexity, while at the same time presenting the theory and application in a way that will hopefully convince you of its value to managers.

In this first chapter, we introduce some essential terminology of statistics and introduce an important distinction between the more traditional view of statistics and what we call the "modern" view. We also discuss important aspects of collecting data, which is the raw material for statistical analysis, some ethical considerations in statistics, and the role of the computer and information technology in statistical analysis.

The Traditional View of Statistics

Our view of the role of statistics in organizations has changed over the past two decades. Here we will try to describe this evolution by contrasting two views of statistics that we will call the "traditional" and the "modern" view. As with any attempt to simplify and "pigeon hole" diverse points of view, this distinction is somewhat artificial and not totally accurate, but hopefully it will convey the general shift in the way statistical analysis is approached.

The traditional view of statistics is as a set of techniques that are utilized to process numerical data. A typical definition of statistics might be something like the following:

> *Statistics* is the body of methods for the collection, analysis, presentation, interpretation, and use of data.

As a body of methods, statistics is applied in many different fields such as psychology, engineering, medicine, and the physical sciences. In the traditional view, it has also been recognized that businesses and other organizations have a need to effectively process and use numerical data. In any organization, managerial success often depends heavily on the availability of valid, timely, and relevant information. An auditor depends on a sample of accounts receivable to determine whether the stated amounts in the company's accounting statements are materially correct. A human resource manager needs an estimate of the average salary for Java programmers in the local area. A marketing manager wants to chart sales by product line for his or her next presentation. A purchasing manager needs an estimate of the average time it takes to obtain parts from a particular supplier to plan lead times for materials. As applied in the business world, the primary objective of statistical analysis in the traditional view has been to enable managers to make better decisions.

The traditional view of statistics in business was heavily influenced by the historical precedents of the way statistics was used in the scientific process, both in the physical and social sciences. The process starts with problem identification, then collecting data, followed by analyzing data and finally reaching a conclusion. The emphasis in most textbooks about statistics has been on the last two phases of analyzing the data, including the type of statistical tests to perform and the computational procedures to follow, and reaching a conclusion.

The Modern View of Statistics

There are several business and social trends over the past two decades that have changed the way in which many organizations view statistics. An understanding of these trends is important in comprehending how our view of statistical analysis has changed in recent years. Our purpose here is not to give a detailed historical analysis of these trends but only to give a broad overview of their influence. Each of them has had a subtle, but profound, influence on the way that organizations view and use statistics.

A Process View of Organizations

One trend in business that has had a major influence on our view of statistical analysis is the "process view." The view of an organization as a set of processes initially emerged from W. Edwards Deming's work in Japan after World War II through various management movements of the past 50 years such as total quality management and Six Sigma. At its simplest level, a process is something that takes inputs and transforms them into outputs.

Figure 1.1 Process As a Set of Activities

Inputs
Process
Outputs

An obvious example of a process is a manufacturing operation that takes raw materials and parts and produces physical goods as outputs, such as automobiles, refrigerators, or computers. However, processes are not restricted to manufacturing. For example we can talk about the hiring process that takes inputs of applicants and produces newly trained employees for the organization. Or we can view an advertising campaign as a process that transforms digital, paper, and other information forms into a communication message that increases brand awareness for the organization. We can talk about the accounts receivable process that takes invoices for products and services and turns them into cash flow for the company. In fact, the process view of an organization asserts that all work of any type is done in a process. Several major movements that arose from this view—in particular Six Sigma—have led not only to an increasingly significant role for statistics in the organization but also to a more dynamic view of statistics as an important tool of organizational change and improvement.

Changes in Statistical Pedagogy

Also during this period, important changes were occurring in academics that would fundamentally change pedagogical thinking regarding statistics. This rethinking of how best to teach the subject of statistics was heavily influenced by the process view of business organizations and the Six Sigma movement. However, it was also motivated by a realization that the traditional methods of teaching statistics in academic institutions, and even in secondary education, were not producing the intended results. In short, students who had taken a statistics class did not know much more about statistics when they had finished the course than when they had started, and even worse, the course typically had minimal impact on their way of thinking about the world. This new approach to the teaching of statistics can be described as the statistical thinking approach (Hoerl and Snee, 2002).

> *Statistical thinking* is a philosophy of learning and action based on three fundamental principles:
>
> - All work consists of a system of interconnected processes.
> - Variability exists in all processes.
> - Understanding and reducing variability are the keys to improving a process.

Understanding variability is at the heart of statistical thinking. Variability can be conceptualized as a deviation from an average or expected value. One of the keys to understanding process variability is to realize that there are two kinds of variability, common cause and special cause. *Common cause variability* is inherent in the process itself and is due to many small causes that are unpredictable given the current state of knowledge. Common cause variability needs to be measured, but it requires no special action. It is simply the expected random variability of the process. *Special cause variability*, on the other hand, indicates that something has changed about the process or that an outside force has exerted an effect on the process. Special cause variability requires managerial action to correct the process and restore it to its proper functioning.

As an example, consider a machine at a cereal manufacturer that fills each box with 32 ounces of corn flakes. Now there are rarely exactly 32 ounces in each box because of minor variations in the process. For example, corn flakes are never exactly the same size and this can produce minor variations in weight. Also, such things as the amount of moisture in the air, the position of the box under the machine, and other factors can cause slight variations. If the process is working properly, however, the average amount of cereal over all of the boxes should be 32 ounces. The variability inherent in the process is likely known to the company and this variability is what we are calling common cause variability. On the other hand, suppose that excess moisture in the air on a particular day has caused some soggy flakes to become lodged in the mechanism of the filling machine. This particular blockage causes some boxes to be underfilled while others are overfilled when flakes become dislodged. This is not part of the normal variability inherent in the process and requires action by the company to correct the situation. This is an example of what we are calling special cause variability.

The key to successful process control is to be able to distinguish between common cause and special cause variability. A measure of common cause variability is the first step to doing this because this measure tells the manager how much variability is to be expected. Variability outside this range must then be due to special cause variation.

For a more everyday life example, suppose that you hear on the news one night that juvenile crime is up 7 percent this year in the town where you live. What does this increase mean? Is there a crime wave going on? Do the police and local city government need to take action? We cannot adequately address these questions without knowing something about the common cause variability. Juvenile crime is also a result of a process, albeit a very complex socioeconomic process. For this reason, we would expect some natural variability in crime rates from year to year even if there were no change in the underlying factors that lead to reported crimes.

If we knew that juvenile crime rates typically fluctuated up or down by 10 percent a year then we would not be too alarmed by the 7 percent increase this year. On the other hand, if we know that juvenile crime rates typically vary by only 1 or 2 percent a year, then the 7 percent increase this year would be quite alarming. Without an awareness of the normal or common cause variability, we would be constantly reacting to random variations as if they were important. This can lead to a great deal of stress, expense, and many unintended consequences associated with the process. Sadly, we do this all too often in public policy and even in the business environment.

As we will see later, comparing a deviation or sample result with expected or normal variability is essentially the process of statistical inference. Making inferences in statistics can be reduced to the question of "Are the data that we observe within the range of common cause variability and therefore due to chance, or are they outside that range and therefore indicative of something important?" It will help you to better appreciate the role of statistics in business if you understand that all data arises from a process, that the outputs of a process are inherently variable, and that managers must make decisions about the process and what steps to take based on that data. Statistics helps managers make more informed and better decisions.

Changes in Information Technology

The last two decades have seen rapid and enormous changes in computer and information technology. The impact of the Internet and e-commerce on businesses is well known. Somewhat less well known is the impact of large-scale data warehouses. A *data warehouse* is a repository of the transaction data of an organization specifically designed to support querying and reporting. Often a data warehouse can contain terabytes of valuable information about an organization's suppliers, customers, and other business partners. The development of these large repositories of information has also led to the development of *data mining* and other business intelligence tools to extract, organize, and make sense of this massive amount of information. Data mining is used for two distinct purposes: knowledge discovery and prediction. For purposes of knowledge discovery, sophisticated statistical techniques for sorting and classifying information are utilized in an attempt to find new and meaningful patterns in data. For purposes of prediction, statistical techniques are used for forecasting and deriving predictive models. In either case, data mining involves the use of statistical techniques to extract meaningful information from the massive collections of data to provide competitive advantage for the organization.

How Is This Modern View of Statistics Different?

The result of the three influences discussed above is a subtle but real shift in the thinking about the role of statistics in organizations. We might describe this shift as moving from "affirmation" to "discovery." In the traditional role of statistics, the purpose was more to affirm our hypotheses about the way things work (the scientific process), and the resulting emphasis was on conducting scientific studies and hypothesis testing. The modern view is more anchored in the discovery mode of trying to uncover new patterns in data or to find new ways to improve processes. The primary problems are to gain an understanding of the variability inherent in all business processes and to devise strategies for responding to that variability.

The importance of this shift from affirmation to discovery is that most statistical software has been designed in the traditional mode and is organized around clusters of statistical techniques. Statistical training, therefore, necessitated learning the assumptions and uses of these different techniques. The JMP software utilized in this text uses an organization quite different from the traditional software tools. From the very beginning, JMP has been oriented to the task of discovery and has incorporated the tools of process management, including Six Sigma. The implication of this is that this book is organized quite differently from the traditional business statistics text. Its organization is more in tune with the philosophy of data exploration and in learning from the data rather than in starting with preconceived ideas to be tested. That does not mean that the traditional techniques of statistics are ignored or that we will not discuss the assumptions and uses of these techniques. It does mean, however, that the orientation and organization of the material will differ from the traditional text just as JMP is different from traditional software. We will explore more of the implications of the JMP software in a later section of this chapter and throughout the book.

Important Concepts in Statistics

No matter which view of statistics we take as our starting point, there are some basic concepts that need to be learned. In this section we will cover some of those concepts and some basic terminology of statistics.

Populations and Samples

Almost everything in the field of statistics starts with the distinction between a population and a sample. In statistical analysis, a *population* is the entire collection of elements that we are interested in for a particular study. The elements could be people (such as employees, customers, voters, etc.), groups (such as companies, sales regions, etc.), or things (products, services, political positions, etc.). A *sample* is any part of the population. (Note, however, that a good sample should be representative of the population.)

A *census* is the collection of data on every element in the entire population, whereas a sample is only part of the population. In many, if not most, cases, doing a census is difficult to justify because of the length of time it would take and the added expense. This means that in the world of business, we are often dealing with samples.

Parameters and Statistics

A *population parameter* is a descriptive measure of the population, whereas a *sample statistic* is a descriptive measure of a sample. In this text, population parameters will always be designated by Greek symbols (e.g., $\alpha, \beta, \mu, \pi,$and σ), while we will use English symbols for sample statistics (e.g., $a, b, \overline{X}, p,$and s). The word "statistic" also has another meaning in popular usage. People often refer to any single item of quantitative data as a statistic and to collections of such items as statistics. Thus we sometimes hear on the news about the recent release of government unemployment statistics. We will use the term data to refer to raw numbers and reserve the term statistics to refer to values arrived at by doing further analysis of the raw data.

Descriptive and Inferential Statistics

Descriptive Statistics

Descriptive statistics is concerned with describing data by classifying, summarizing, and graphing either population or sample data. In other words, descriptive statistics can apply to either a population or a sample. For example, suppose that a personnel manager of an airline company is faced with negotiating a new wage contract. Charts and summary measures of the current wages paid to pilots, mechanics, flight attendants, and other employee groups would give the negotiators a composite overview, or summary, of wage levels and differences among workers. It's true—a picture is often worth a thousand words!

In addition to graphical displays of data, in many cases we require numerical values that summarize particular characteristics of the data such as central tendency or the variability or dispersion in the data. For example, we may need to use descriptive measures for financial projections or building financial models. Although pictures are nice, they are difficult to incorporate into a financial model.

The important aspect of descriptive statistics is that it simply describes the data and do not go beyond this description to make inferences on the basis of that information.

Inferential Statistics

Inferential statistics involves making inferences that go beyond the known data. Most statistical inference is inductive in nature. Induction is the logical process of using specific knowledge about a sample from a population to infer something about the population as a whole. For example, a bank economist may use a sample of the unemployment levels in 20 Ohio cities to estimate the statewide unemployment rate, or an auditor may use a sample of accounts receivable records to infer that the account statements are accurate.

Sampling and Nonsampling Error

Firms apply statistical sampling to everything from inspecting raw materials to auditing accounts or assessing customer satisfaction with their products. In each case, only part of the population is inspected or measured. But that limited information is used to make a decision about the entire population. There is some risk of making a mistake because the decision must be made on the basis of partial information. For example, a bank economist estimating statewide unemployment in Ohio might happen to get several cities with unusually high unemployment in a sample of 20 cities. She then might estimate state unemployment at a higher rate than it really is.

Sampling error is the deviation that exists whenever a sample statistic differs from the population parameter because of the chance failure of a sample to perfectly represent the population from which it is taken.

Sampling error is due to chance (the "luck of the draw") and is a characteristic of statistical analysis. So we can always expect it and must account for it in our analysis. In fact, the analysis of sampling error is a trademark of inferential statistics.

Statistical investigations often contain errors that are not due to sampling error. These *nonsampling errors* or bias can arise from the nature of the estimator used, or from unrecognized systematic causes, such as unskilled interviewers or inaccurate measuring devices. Data recording and transcription errors are additional sources of nonsampling error. Biased sample statistics do not accurately represent their population values because they are consistently off the mark, much like a speedometer that always registers too low or a bathroom scale that always reads too high. If you don't know about the bias, you may be in for a surprise, but if you do know the amount and direction of bias, you can sometimes make a correction for it. However, we will usually want to eliminate bias if at all possible.

Bias or systematic errors can occur for many reasons. One common source is the administrative procedures and the tasks associated with taking, recording, or reporting data. **Moreover, these nonsampling errors can occur in both census and sample data**. In fact, if large masses of data have to be handled, nonsampling error may be so large that a smaller sample that eliminates nonsampling error may be more accurate in estimating the population value than doing an entire census. This is one argument that some have advanced for doing a sample rather than a complete census of the U.S. population every 10 years.

Accuracy, however, is not the only criterion of a good sample. A sample of two items can be very "accurate," in the sense of no systematic bias, but may be too small to provide a reliable estimate of the population parameter. The sample must be large enough to ensure that it is representative. Although a sample of two items may not have any individual inaccuracies, a decision maker should feel much more confident using a sample of 1,000 rather than a sample of two to provide an indication of the population characteristic.

Statistical Inference and the Discovery Process

As we stated earlier, modern statistical software such as JMP is oriented more toward exploration and discovery rather than hypothesizing and confirming. However, that does not mean that the traditional methods of statistical inference are not relevant to a study of statistics. The reason for the continued importance of statistical inference relates to the distinction between population and samples, and the existence of sampling error. Even though we are interested in the entire population, usually we have data only on a sample taken from that population. We know that sampling error is always present when we take a sample. Therefore, if we discover, for example, that there is a difference between males and females with respect to ATM usage at the bank, does that difference generalize to the population at large? How large is that difference in the population? In other words, is the difference real or just the result of sampling error, and what is the magnitude of that difference in general? We cannot convincingly answer these questions without the methods of statistical inference. For this reason, even though our purpose may be exploration and discovery, we still need the traditional tools of inferential statistics to know whether or not our discoveries are "real" or just a chance result of this particular sample data.

Collecting Data

Most of the topics in this text will presume that the data is already available. However, the data collection process is very important because it will determine the quality of the data, and, therefore, the quality of the conclusions and inferences we draw from that data. In this section, we will try to highlight some key issues in the collection of the data that have important ramifications for the quality of the investigation. In a business context, there are three basic types of issues that are usually involved in a statistical analysis:

- Estimating a value
    - What is the average income in a certain section of town?
    - What percentage of consumers prefers our soft drink to our competitor's?
- Making a comparison between groups
    - Are people who work in self-managed teams more productive than people who work alone?
    - Are the returns on investments in IT-related projects the same as those on more traditional capital investment projects?
- Studying relationships between variables
    - Is there a relationship between amount of training and worker productivity?
    - What is the relationship between the amount spent on Web advertising and sales for the following month?

No matter which business issue is being addressed, the first major question is whether the data that already exist are adequate for our needs or if we need to collect additional data ourselves. Data that already exist but were gathered for another purpose are called *secondary data* because they are used for a purpose that was not the primary reason for gathering the data in the first place. For example, a Montana lawn care firm could find that late summer billings for its services were strongly associated with customers that had sprinkler systems. Although the billing and sprinkler data were not originally gathered to address this issue, a decision to expand the services of the company to include sprinkler installation with a service contract could be justified on this evidence and may be expected to increase future sales. Data gathered for the purpose of the current study are called *primary data*.

There are two principal sources of secondary data, the firm's internal records and resources, and external data from outside the organization. The firm's own records and databases are often its most useful source of information. This data is typically gathered as a consequence of monitoring and documenting everyday business activities and not for the purposes of a statistical study. Other common forms of data typically maintained by organizations include personnel records and accounting information.

Data of this type, however, may also be very useful for other purposes. When data are used for any purpose other than the one for which they were originally gathered, the data are referred to as secondary data. Organizations large and small use well-designed information systems to supply a variety of statistical data about past and present operations. Modern organizations often purchase business analytics software to help mine the vast quantities of information that they collect to help them manage their operations more effectively and to gain a competitive advantage over other firms.

Advantages and Disadvantages of Secondary Data

The obvious advantage of secondary data is usually the lower cost of gathering the information. Often, as with most government data, we can even get the information without direct cost, although we may have to pay an employee to gather the information. Another advantage of secondary data is that we can usually get the information very quickly.

However, there are a number of issues with secondary information.

- Are there errors in the data?
- Do the definitions and measures used in the study correspond to our needs?
- Are the data current?
- Were the data gathered in an appropriate manner?

Most of the problems with secondary data are based on the fact that we have no control over how the data were gathered and reported. Issues with measurement and differing ways of defining variables are a frequent difficulty. For example, there are a variety of secondary sources of productivity data in different industries. However, there are a variety of ways of defining productivity, and different sources may use slightly different definitions. Using these data for other purposes can then be problematic. This issue of measurement and definition can also arise with internal data. For example, with internal data, we can often find data that closely, but not exactly, match our needs. A common instance of this is in trying to forecast demand for the firm's products and services. The data that we are likely to find in the firm's records are sales data. However, sales and demand are not the same thing. One reason that they can differ is that there are times when a customer desires to purchase our products or services (there is a demand), but for one reason or another we are not able to provide the product or service. We do not have the product in inventory, or we are at our maximum capacity for providing the service. This demand may never be recorded as a sale if the customer goes elsewhere to buy the product or service, or the sale may occur during a different time period than demand even if the customer is willing to wait. In either event, the sales data are not the same as demand.

Collecting Primary Data

Given these issues with secondary data, it may well be worth the time and expense of gathering our own data for a statistical investigation. There is an old saying "When you want something done right, do it yourself." This is often good advice when it comes to primary versus secondary data.

Key Issues in Collecting Data

Given the fact that we are frequently going to gather our own data, there are a number of questions that arise. Most of these questions revolve around four key issues.

- Who are we going to gather data from and how do we select them?
- How much data are we going to collect?
- How are we going to get the data?
- How much control are we going to exercise over the situation when the data are collected?

The first two issues will be dealt with in more detail in later chapters when we talk about sampling and estimation. It is the last two issues that we want to explore here.

The third issue relates to how we get the data. To obtain the data, we can either ask the respondents (a survey), or we can observe the data we want to record. For example, we can ask individuals which of a set of multimedia ads are the most attention-getting (survey) or we can observe them as they watch the ads and do other tasks (observation).

Surveys and Observation

Surveys and observation are processes that yield data about the status or elements of a population. *Surveys* typically ask for responses from the elements of the population. With *observation*, we simply observe the behavior in which we are interested. Examples of surveys might be an employee survey of attitudes toward current benefits packages or a Web survey regarding the usability of our Web site for current and potential customers. An example of an observational study would be the use of "mystery shoppers" that some retail firms employ where they send other personnel into their stores to pose as shoppers and gather observational data on store employee behavior. Another example of an observational study might be the use of a contracted firm's personnel to go to competitors' stores and gather pricing information.

One of the major difficulties with using observation is that if the respondent is aware that they are being observed, this may change their behavior. This means we usually need to observe someone without their knowledge. This can, of course, bring up ethical issues and concerns about privacy. Care must also be taken that there is no bias introduced by the observer. Since the observers are often human, their observations are potentially subject to their experiences and biases. Care must be taken to ensure that the observers are objective and that the subjects' responses typify their normal behavior. In addition, there are three conditions that must be satisfied before we can use observation.

- The phenomenon that we are interested in must be something that is observable. For example, we cannot use the observation method to study attitudes toward a product since such attitudes are not directly observable.
- The phenomenon to be measured must be one that occurs fairly frequently. It wouldn't make much sense, for example, to use the observational method to study industrial accidents since such accidents would be fairly rare and it would take a long time to get just a few observations.
- The phenomenon to be measured must occur over a fairly short time span. For example, studying the potential effects of a change in benefits on long-term retention of key employees would not work very well using observation since the time span would be so long that the results would not be of much value.

If all of these criteria are met, observation can provide high quality data that haven't been biased by socially desirable behavior or faulty measurement devices. If not, you would probably be better off just asking someone a direct question in a survey.

Surveys, directly asking the respondents for information, can be done through interviews or questionnaires. Interviews, whether in person or over the phone, require a trained interviewer asking prescribed questions of a respondent and recording the response on a standard form. The response rate is usually higher than with mail or Internet questionnaires because another person is either directly in front of the respondent or is talking to them in real time. More detailed information can also be obtained because questions can be explained and clarified as needed.

In-person interviews are expensive, sometimes costing as much as $200 or more per respondent. According to CensusScope, the 2000 U.S. Census cost approximately $6.5 billion, or $56 per housing unit. Phone interviews are a less expensive and widely used method of obtaining data through personal interviews. However, recent technologies such as Caller ID and the National Do Not Call Registry have made phone interviews increasingly difficult.

Questionnaires involve giving each respondent the same preset set of questions to answer. Questionnaires have been traditionally administered through the mail but increasingly the Internet is being used as well to administer Web-based surveys. Regular mail questionnaires are a relatively impersonal means of collecting external statistical data. Web-based surveys are a rapidly growing type of survey administration that exhibits many of the same characteristics we associate with regular mail surveys. Great care must be exercised in designing both types of questionnaires to avoid misunderstandings, to eliminate questions that can yield biased results, and to ensure that the needed data can be obtained and tabulated.

Experiments and Post-Hoc Studies

Another fundamental issue in gathering primary data relates to how much control the researcher exercises over the situation. In *experiments*, the researcher actively manipulates at least some of the factors involved. For example, in a marketing research study, a researcher may systematically vary product features to investigate which features are important to purchase intentions. Or a grocery chain might physically change the point-of-sale displays in a sample of stores to observe the impact on sales. In non-experimental, situations, the researcher simply takes the factors as they exist in the environment and conducts a so-called *post-hoc* or after-the-fact study. The primary advantage of experiments over post-hoc studies is in demonstrating causality. The active manipulation of experimental factors is crucial to the demonstration of causality. It is virtually impossible to argue causality from post-hoc tests because we cannot rule out other plausible explanations as having caused the relationship between variables or differences between groups. Experiments where some factors are directly manipulated and others controlled through randomization are crucial to arguing for a causal influence.

A Classification of Data-Gathering Situations

We can combine the two distinctions that we have just discussed to classify data-gathering situations into four different types as depicted in Table 1.2. It should be noted that any one study can use a combination of these methods. For example, in a given study we might manipulate some variables (experiment) but simply take others as they exist (post-hoc). We may ask respondents for some information and observe other data.

Table 1.2 Examples of Data-Gathering Situations

	Survey (Self-Report)	Observation
Experiment	Volunteers are shown films, in one group with violent content and another group mild content. They are then asked their opinions on capital punishment.	One group of workers is organized in teams, and in another group individuals work alone. We observe their productivity (units of output).
Post-hoc	We survey a group of consumers about their shopping habits and look at the differences between those who are married and those who are single.	We observe a number of shoppers and whether or not they stop and observe an in-store display. We also note their gender and whether or not they have children with them.

In a business context, much of the data we use are historical such as sales, expenses, and other performance measures, which are observed values and of necessity post-hoc. However, some areas of the firm often use survey type data. For example, the marketing function often deals with survey data obtained from customers that incorporate post-hoc variables but may also include experimental manipulations as well.

Types of Data

In addition to knowing how we want to collect the data, it is also important to understand what kinds of data we have collected. The types of calculations we can perform on the data and the types of analyses we can do depend on the nature of the data we have. Most statistical packages, including the JMP software bundled with this text, make use of these distinctions. The first distinction relates to whether or not the data are inherently numerical in nature or are not numerical.

Quantitative versus Qualitative Variables[4]

We are all familiar with *quantitative data*, such as the total unit sales of a product line, or data on income and employment levels. Quantitative data are inherently numerical in nature, and it makes perfect sense to add the values or calculate averages. For example, we might add unit sales across product lines to get total sales or average a group of incomes. *Qualitative data*, on the other hand, are normally represented by text or characters. For example a variable representing the different sales regions of a company (Eastern, Midwest, Southern, or Western) is not a numerical variable, although we may sometimes assign numbers to it. Similarly, common variables such as gender, organizational units, and product lines are qualitative variables. Qualitative variables are often used in grouping data from quantitative variables. For example, we might examine average sales by sales region for a particular company.

Discrete and Continuous Variables

For quantitative variables, a further distinction is necessary for understanding different statistical procedures. This distinction is between discrete and continuous variables.

Discrete variables: Variables that can assume only separate and distinct (countable) values are referred to as discrete. Although there are exceptions, in general, discrete variables are integers and do not assume fractional values.

Continuous variables: Variables that can take on any value in a range of (measurable) values are referred to as continuous. In general, continuous values can take on any integer or fractional value.

You can often make this differentiation by asking yourself whether the numbers (the data) result from something being counted (i.e., are discrete) or something being measured. Discrete variables can assume only a finite number of values within any finite interval on the real number line. Continuous variables, on the other hand can take on an infinite number of values within an interval on the real number line. For example, number of employees in an organization would be a discrete variable while age of employees would be a continuous variable. So a firm might have 205 employees (discrete), and their average age (continuous) might be 42 years of age. If we wanted to be more precise, we could refine the average age to 42.2 years, or 42.23 years, or 42.234 years, and so on. In theory, there is no limit to the precision with which we could measure age. We can measure it in years, tenths of years, hundredths of years, etc. In other words, it is continuous and can take on any value greater than zero. However, we cannot further refine the variable for number of employees to a higher level of precision, and it cannot be fractional (the number of employees, cannot be 205.5). Therefore, this variable is discrete because it can only take on certain values. There are discrete variables that are not integer in value, but they are relatively rare. Most examples come from the fashion industry. For example, in the United States, but not in Europe, shoe sizes come in fractional units, but the variable is still really a discrete variable. For example, you can buy a size 9 ½ shoe but you cannot buy a size 9 ¼ shoe. Another example is hat sizes. However, in most cases discrete variables only take on integer values.

Although we have discussed discrete and continuous variables as a dichotomy of distinct types, in practice, we often treat this as a continuum. At one end of the continuum are variables that are obviously discrete and can take on only a few possible values. For example, the number of defects in a sample of 10 parts obviously can only take on the values of 0 through 10. Something like weight of a shipment is obviously a continuous variable because it can take on any value, including fractional values, depending on the precision of our scale. In a statistical analysis, we would treat these variables quite differently. However, in some cases, for analysis purposes we might treat a discrete variable as if it were continuous. For example, if our variable represented the number of hammers sold by Home Depot worldwide in a given month, that variable is, strictly speaking, discrete. We cannot sell one-half of a hammer. However, the number of possible values that this variable can assume is certainly in the hundreds of thousands if not millions. Because the number of possible values is so large, we may well treat this as a continuous rather than a discrete variable. This type of situation often arises in a discussion of probability distributions and will arise again in Chapter 6.

Scales of Measurement

A third distinction related to the data is in terms of the measurement scale that produces the numbers. Statistics deals with numbers or data. Data are measured on a scale, and it is important to know the type of scale because the kinds of operations you can perform on the data will depend on the properties of the underlying scale. For example, addition and subtraction aren't appropriate for certain types of scales. For other scales, addition and subtraction may be fine, but ratios are not appropriate. S.S. Stevens (1946) defined a classification of scales that has been commonly used since. According to Stevens, we can distinguish four scale categories: (1) nominal, (2) ordinal, (3) interval, and (4) ratio.

Nominal Scales

The *nominal scale* is used to describe data that are categorical or qualitative in nature. For example, residents of an area may be classified according to the type of automobile they drive or their religious affiliation. Nominal scale data are the types of data that we ordinarily treat as text or characters. For example the type of car might be classified as Ford, Chevrolet, Chrysler, Toyota, Nissan, etc. If numbers are assigned to represent the different values, and they often are, the numbers have limited meaning. The only meaning present in the numbers is to indicate that the cars come from the same or different manufacturers. Even the order of the numbers is meaningless for a nominal scale. Just because the number to represent a particular car is higher than the number for another car does not imply that one car is somehow "better" than another.

Ordinal Scales

The *ordinal scale* is used for ranked observations according to some order of preference, where the rankings convey a relative position (i.e., on a scale from high to low or low to high). For example, a supermarket survey may show that consumers rank dessert preferences as cheesecake (1st), ice cream (2nd), pie (3rd), cake (4th), cookies (5th), and pudding (6th). Note that an ordinal ranking conveys more information than a nominal classification, but it does not imply anything about the equality of the differences between ranks. Thus the preferential difference between the first two items (cheesecake and ice cream) is not necessarily the same as that between the last two (cookies and pudding). This means that addition and subtraction are unacceptable operations for ordinal scales. Therefore, asking for the average ranking is still not a meaningful question. However, asking what is the middle value such that half of the values are above and half below does make sense because order has meaning.

Interval Scales

The *interval scale* is used to describe data that have a constant unit of measure, but not necessarily a natural zero point. For example, assume the temperature in Phoenix is 110 °F and in Anchorage is 55 °F. We could say the temperature in Phoenix if 55 degrees higher than in Anchorage because each degree is a known (constant) increment. However, we could not say that Phoenix is twice as warm as Anchorage! This is because zero on the Fahrenheit scale is not an absolute zero (i.e., absolute zero—where there is no heat—is at -460 °F). The zero points on the Fahrenheit and Celsius scales are in a sense arbitrary. Therefore, a ratio of 2:1 on the Fahrenheit scale is not a ratio of 2:1 on the Celsius scale.[5]

While temperature is an obvious example of an interval scale, other scales are not so obvious and often there is debate about whether a particular measure is ordinal or interval. Consider, for example, the common "Likert Scale" items commonly found on questionnaires and student teacher evaluations. There is considerable doubt as to whether or not these are interval scales or only ordinal in nature. Your school most likely has some form of evaluation instrument where students evaluate their classes and instructors at the end of the term. These questions are often on a point scale, say from 1 to 7 where 1 is the most negative evaluation and 7 is the most positive. That means that 4 serves as a type of neutral or zero point. Treating this as an interval scale implies that the differences between the numbers are constant so that the difference between the ratings of 4 and 5 are the same as the differences between the ratings of 5 and 6. However, this is often a doubtful assumption. The difference between a rating of 4 (neutral) and 5 (positive) may in fact be much larger psychologically than the difference between 5 and 6, which are two positive ratings. Although researchers often treat such data as if they were interval in nature, this practice can lead to misleading results if there are severe differences.

Ratio Scales

The *ratio scale* is used to describe data that not only have a constant unit of measure so that one can talk about differences, but also have a natural zero point. For example, $100,000 of profit is indeed twice as much as a $50,000 profit. This ratio will be the same no matter which currency the profits are measured in. If we convert the currencies to euros or the Chinese yuan, the ratio will still be 2:1. This is because there is a natural zero point. Zero money is zero money, as you may be all too aware, no matter whether measured in dollars, pounds, euros, yuan, or any other currency. Data that have ratio scale properties contain the most information and are frequently the most useful for statistical purposes. We can perform any kind of arithmetic operations on ratio scales.

Computers and Statistical Analysis: Introducing JMP

As discussed earlier, the modern business enterprise records massive amounts of data that are instantly available to almost anyone in the organization. Much of this information is loosely organized and largely unexplored. The business questions related to the data are also largely unstructured and ill formed. Often the issue is about "learning" about customers, suppliers, or the general business environment. Traditional statistical software with structured menus organized around statistical techniques is ill-equipped for this situation. What is needed is software designed for data exploration with a visual orientation. This is precisely why JMP (pronounced Jump) software was originally created. Visually oriented and structured with drill-down capabilities, JMP is ideally suited for data exploration and learning.

JMP was developed by SAS, a leading provider of business analytics software and services. JMP has proven to be especially useful in areas such as process improvement and Six Sigma efforts. JMP is very visually oriented, and you should find it very easy to use. It is logically structured around the task of understanding a set of data and what the numbers tell us. The next chapter will provide a more detailed introduction to JMP. You will have many opportunities to utilize the software throughout the text, and by the end of the text you should be very comfortable with JMP.

The Inland Northwest Credit Union

To illustrate statistical principles and the use of JMP throughout this text, we will use a common example of a fictitious credit union located in the state of Washington. We will simply introduce the company here and will then use the example throughout the text. Although the company is fictitious, the situations described in the text are situations that can arise in any real-world business environment.

The Inland Northwest Credit Union (INCU), a member of the National Credit Union Administration (NCUA), was founded at a local high school during the 1920s and has since grown to be one of the largest credit unions in its region. The credit union now has over 1,000 employees and over 120,000 members in the state of Washington and three counties in Idaho. With over $3 billion in assets, INCU is one of the largest credit unions in the state of Washington.

The senior levels of management at INCU consist of Susan Strong (CEO), Hank Wilson (CFO), Darrell Young (CTO), Mary Warner (COO), Alice Hansen (Vice President of Human Resources), and Ann Rigney (Vice President of Marketing). Other employees of INCU will be introduced as we go through the text and introduce new situations and problems to be solved at the credit union.

Summary

Statistics is part of the growing field of business analytics. The traditional view is of statistics as the body of methods used for the collection, analysis, presentation, and interpretation of data. In this view, statistical studies entail an organized sequential process of (1) identifying the problem, (2) establishing the criteria for solution, (3) collecting data, (4) analyzing and testing the data, and (5) drawing conclusions.

A more modern view of statistics derives from three general social phenomena that have developed over the last thirty plus years. The first is viewing a business as a set of processes with the key business issues being understanding and managing these processes. The second, partly an outgrowth of the first, is the change in statistical pedagogy around the viewpoint of statistical thinking. Statistical thinking states that (1) all work is a process, (2) all processes exhibit variability, and (3) understanding and reducing variability are the keys to process improvement. The third and final force for change is the tremendous changes in information technology and the changes in the amount and type of data maintained by the typical business. Together these forces point to the need for a different view of statistical pedagogy and the need for different types of software tools.

No matter what the viewpoint, learning statistics involves learning new concepts and terminology. Distinctions between populations and samples, and the associated distinction between parameters and statistics, are crucial to understanding inferential statistics. The concepts of sampling and nonsampling error are introduced here and will be discussed in more detail later in the text.

Data, the primary input to statistical analysis, come from both internal sources (company accounting, customer, and operational data) and external sources, such as government data, national and international trade as well as other organizations, and increasingly, from global Web sources. The data collection phase requires important decisions regarding the use of primary or secondary data, internal versus external sources, surveys versus observations, and experimental versus post-hoc studies.

As societies become more affluent and knowledge-oriented, information systems and large-scale data warehouses take on increasing importance. Specialized software such as JMP is increasingly being used to make sense of all this information. JMP will be utilized throughout the text to illustrate the practice of statistical analysis in business. The Inland Northwest Credit Union was introduced as an example setting to anchor our discussions and illustrations in a common business setting.

Chapter Glossary

business analytics The use of data, statistical and quantitative analysis, and explanatory and predictive models to drive business decisions.

business intelligence Systems and technologies for gathering, storing, accessing, and analyzing business data to improve strategic decision making and organizations' competitive positioning.

census The collection of data from an entire population.

common cause variability Random variability in the process that arises from many small causes that are not predictable. It does not require action unless we are going to change the entire process to reduce this variability.

data mining The process of analyzing data from different perspectives to extract information and relationships useful for future predictions.

data warehouse An integrated central time-variant repository of organizational data used to support management decision making.

descriptive statistics Concerned with summarizing, displaying, and describing a set of numbers. It does not involve any inference that goes beyond the given data.

experiments Involve the deliberate control and manipulation of some variables when studying a sample or population.

inferential statistics Making inferences about unknown values based on known information.

nonsampling error Errors due to ways that a statistical study is conducted that can occur in a census of a population as well as in a sample.

observation Gathering data without the direct questioning of the participants in the study. Usually involves visual or auditory information.

population The set of all elements that we are interested in for a particular study.

population parameter A measured characteristic of a population.

post-hoc A study that takes the measures of variables and other factors as they exist without any active manipulation.

primary data Data gathered for the particular purpose of a statistical study.

sample A subset of the population.

sample statistic A measured characteristic of a sample.

sampling error The difference between a sample statistic and a population parameter that is due to the chance failure of the sample to be perfectly representative of the population.

secondary data Data used in a statistical study that were originally gathered for some other purpose.

special cause variability Nonrandom variability introduced into the process by a specific identifiable cause either from outside the process or because of a change in the process itself.

statistical thinking The philosophy of learning and action based on the principles that (1) all work consists of a system of interconnected processes, (2) variability exists in all processes, and (3) understanding and reducing variability is the key to improving a process.

statistics The body of methods for the collection, analysis, presentation, interpretation, and use of data.

survey A process of gathering information about a population or sample by asking questions.

Questions and Problems

1. A friend of yours has questioned why you are studying statistics as part of your business studies. She claims that business is all about finances and that accounting, finance, and maybe a little marketing are all you need to know. How do you respond to your friend?

2. Explain the difference between parameters and statistics.

3. Elaborate on the difference between sampling error and nonsampling error.

4. The Clean Air Coalition recently conducted a survey of 1000 car owners in the U.S. and found that the average car owner drives 24.7 miles per day during the work week and between 9,300 and 10,100 miles per year total.

a) Indicate whether this is

   i. sample or population data

   ii. primary or secondary data

b) What type of inference would the Coalition be likely to make from this data?

c) Are their conclusions likely to be accurate? Explain.

5. A competitor's Web site claims that their company's software is faster, more reliable, and more accurate than the software produced by your organization. Explain how you might conduct a study to gather data to determine whether your competitor's claim is justified by the facts.

References

Davenport, Thomas H., and Harris, Jeanne G. 2007. *Competing on Analytics: The New Science of Winning*. Boston, MA: Harvard Business School Press.

Hoerl, Roger and Snee, Ronald D. 2002. *Statistical Thinking: Improving Business Performance*. Pacific Grove, CA: Duxbury.

Stevens, S. S. 1946. "On the Theory of Scales of Measurement." *Science*, 103(2684):677–680.

Notes

[1] www.google/analytics

[2] http://www.analyticsmagazine.com/

[3] Davenport and Harris. 2007.

[4] JMP makes this distinction between data types as Numeric columns versus Character columns. The JMP classification will be discussed in more detail in Chapter 2.

[5] The Kelvin scale, on the other hand, does have an absolute zero point and is, therefore, a ratio rather than an interval scale.