CHAPTER 1

STATISTICS: AN OVERVIEW

1.1 ABOUT THIS CHAPTER

We begin this general overview chapter with a brief discussion on "What is Statistics?" and briefly describe some applications. We then indicate the importance of statistical literacy to our daily lives, discuss why better decisions require better data, and introduce the concept of statistical thinking. We conclude with a brief consideration of the relationship of statistics to mathematics, computer science, and operations research/management science (OR/MS).

1.2 WHAT IS STATISTICS?

> **There are never in nature two beings which are exactly alike.**
> —G.W. Leibniz

The Merriam-Webster Online Collegiate Dictionary defines statistics as "a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data." Our preference is to exclude the "masses of " from the definition. Even today, many situations still require us to deal with small data sets. We might also quibble about the word "numerical." Statisticians sometimes deal with data on preferences, rankings, grades, and colors that are nonnumerical. Finally, we have strong reservations about statistics being a "branch of mathematics" – more on that shortly. Apart from that, we agree with the dictionary definition.

Jon Kettenring, 1997 President of the American Statistical Association (ASA), has more simply referred to statistics as "the science of learning from data" that "is essential for the proper running of government, central to decision making in industry, and a core component of modern educational curricula at all levels." Sallie McNulty, the 2006 ASA President, describes statistics as "the quintessential interdisciplinary science." We agree.

We live in an uncertain world. We can predict the time of the rising and setting of the sun for a specified location with great precision. But the exact outcomes of most other things in life are far less known. Two cars built on the same manufacturing line, even on the same day, will still differ in performance, reliability, appearance, and so forth. One of our major goals, in fact, is to minimize such variability.

The Role of Statistics in Business and Industry, by Gerald J. Hahn & Necip Doganaksoy

Statistics allows us to quantify uncertainty, and, as a consequence, to express the likelihood of future happenings probabilistically – and, when possible, act to impact such outcomes favorably. This typically involves the gathering of appropriate data and their proper analysis.

1.3 AREAS OF APPLICATION

Statisticians are employed in academia, business and industry, government, and by social and research agencies. Statistics is, in fact, used in almost all fields of human endeavor, ranging from actuarial science to zoology. At the time of this writing, the ASA Internet site lists 45 application areas. And we recently learned of one more with the publication of a book on statistics in musicology (see Beran 2003).

So on what kinds of problems do applied statisticians work?

1.3.1 Automobile Gasoline Consumption

An automobile manufacturer wishes to compare the gasoline consumption of one of its models against that of competitors. The results are to be used as the basis of a possible advertising claim. Samples of different model cars are to be test-driven to obtain a statistically valid comparison. How many cars of each model need to be evaluated? How should these be selected? Under what conditions should they be driven? For how long? Who should do the driving? How can one factor in variability in gasoline? How are the resulting data to be analyzed? And how should an advertising claim be worded so that it is both clear to the audience at which it is directed and valid? And more fundamentally, what is the impact of fuel economy on customer purchasing behavior and satisfaction? Statistical concepts and methods, often as part of a multidisciplinary effort, help answer such questions and many more.

This example suggests a common thread in addressing problems statistically: define the question of interest, obtain appropriate data, analyze the data, and then help make decisions from the findings.

1.3.2 Some Other Examples

Peck, Casella, Cobb, Hoerl, Nolan, Starbuck, and Stern (2005) aim to show a general audience "how statistics can provide an organized way of learning from data and how the ensuing knowledge can be used to address important, social, environmental and economic problems." Table 1.1 shows the topics of the 25 essays in this volume.

Also, Peck, Haugh, and Goodman (1998) provide case studies from 20 different application areas, including biology/environment, medical and health care, pharmaceutical, marketing and survey research, and manufacturing.

Chance and *Stats*, two ASA magazines, give further examples. And the 2,000 or so papers presented at the yearly national Joint Statistical Meetings, listed on ASA's Internet site, provide both examples of current hot topics and overwhelming evidence of the rich diversity of subjects addressed by statistics.

TABLE 1.1 Essays in Peck, Casella, Cobb, Hoerl, Nolan, Starbuck, and Stern (2005)

Public Policy and Social Science
• Statistics in the Courtroom
• The Anatomy of a Pre-election Poll
• Counting and Apportionment: Foundations of America's Democracy
• Evaluating School Choice Programs
• Designing National Health Care Surveys to Inform Health Policy
Science and Technology
• Monitoring Tiger Prey Abundance in the Russian Far East
• Predicting the Africanized Bee Invasion
• Statistics and the War on Spam
• Should You Measure the Radon Concentration in Your Home?
• Statistical Weather Forecasting
• Space Debris: Yet Another Environmental Problem
Biology and Medicine
• Modeling an Outbreak of Anthrax
• The Last Frontier: Understanding the Human Mind
• Leveraging Chance in HIV Research
• Statistical Genetics: Associating Genotypic Differences with Measurable Outcomes
• DNA Fingerprinting
• How many Genes? Mapping Mouse Traits
Business and Industry
• To Catch a Thief: Detecting Cell Phone Fraud
• Reducing Junk Mail Using Data Mining Techniques
• Improving the Accuracy of a Newspaper
• Assessing Product Reliability and Safety
• Randomness in the Stock Market
• Advertising as an Engineering Science
Hobbies and Recreation
• Baseball Decision Making
• Predicting the Quality and Price of Bordeaux Wines

And to add a personal element, we describe briefly in Sidebar 1.1 three interesting, and somewhat unusual, problems in which we have been involved.

SIDEBAR 1.1: ALL IN A DAY'S WORK: SOME MEMORABLE APPLICATIONS

Birds, Mating Calls, and Airplane Crashes

Crashes of jet airliners due to the plane's engine ingesting birds upon landing or takeoff are of obvious concern to the manufacturers of airplane engines. An article in the *New York Times* suggested that birds might mistake the sound emanating from certain model jet engines to be mating calls. As a result, instead of flying away, they fly toward the engine and are trapped in it, resulting in malfunction.

This assertion needed to be tested scientifically. So an experiment was to be conducted, at a carefully selected site to evaluate the claim. This involved playing recordings of mating calls and jet engine noise to groups of birds and observing their behavior. Our job was to help plan an experiment that would yield statistically valid results and to analyze the resulting data. It demonstrated that birds are no fools; they are smart enough to differentiate jet engines from mating calls.

Which Brand Food Do Dogs Prefer?

We have been engaged in various consumer preference studies. One dealt with determining preferences among dogs between different brands of food products. The results were to be used to substantiate a proposed advertising claim. Dogs were offered food from the competing brands by selecting from among various bowls placed in front of them over a period of time. The amount of food consumed from each bowl was subsequently measured. We were asked to assess the statistical validity of the experiment used to make this comparison and of the analysis of the resulting data.

Combating Barnacles

Ships and submarines that spend time in harbors are susceptible to marine growth in the form of barnacles, mussels, algae, and other tiny organisms on their hulls. These limit the vessel's maneuverability and speed. Constant plugging up of water inlets by invading organisms is a particular challenge for nuclear-powered vessels. Various coatings to inhibit permanent attachment of marine organisms to submarines had been developed. We were asked to plan a test to compare the effectiveness of alternative coatings and analyze the results. This involved submerging specimens with different types of coating in water and allowing marine growth to occur. The primary measured response was the force required to remove the organisms from the specimens.

1.3.3 Examples from Business and Industry

> **Statistics, as applied to products and services, is not an end in itself, but exists principally to illuminate the way to progress in other fields that require the proper handling of uncertainty.**
> —*Antonio Possolo*

As a prologue to the discussion in subsequent chapters, we present a few examples of how business and industry use statistics:

- Pharmaceutical companies request the appropriate regulating agency to license a new drug – based upon a statistical analysis of the results of carefully planned studies of the drug's performance in treating a specified condition and of its side effects. The request also specifies what type of patients should (and should not) take the drug, how frequently, and at what dosages.
- Providers of consumer credit use statistical models to decide whether or not to approve new loan requests (and establish maximum allowable credit limits) to provide the best possible balance between business gained from authorizing loans and losses suffered due to payment default. Such models are developed

by relating applicant and loan characteristics to payment performance, based upon the analyses of data from past loans.

- A manufacturer of a chemical product uses recent process information to determine how much catalyst to add to the product mix during manufacture so as to come as close as possible to the desired target performance. The algorithm is based upon past data used to characterize the relationship between the process and performance variables.
- Another manufacturer monitors engine measurements (e.g., oil pressure and temperature, coolant temperature, and fuel consumption) of a locomotive in operation on an ongoing basis so as to trigger preventive maintenance, and, if necessary, shut down the engine to avoid an impending field failure. The statistical algorithm used for this purpose is based upon an analysis of past data; it aims to detect the maximum number of impending failures, while minimizing the false alarm rate.
- Broadcasters of television commercials need decide between selling advertising timeslots in a pre-season auction versus on the "spot market" during the year. Spot market prices are determined by recent ratings of the show during which the commercial is to be aired. The preseason auction price is determined based on the expected audience and provides early cash flow and minimal risk. However, it can result in loss of revenue if a show delivers higher ratings than expected. Statistical studies help broadcasters arrive at the best decision by providing algorithms to predict future show performance.

1.4 STATISTICAL LITERACY

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.
—H.G. Wells

The need for an understanding of basic statistical concepts, or statistical literacy, is not limited to the workplace, but pervades our daily lives. It is an important part of being an educated citizen.

1.4.1 Some Examples

We need only pick up a newspaper or watch a television newscast to learn about statistical studies that have important social and national implications. For example:

- The almost daily opinion polls that provide the "pulse of the nation" (with associated statistical error bands) on key issues such as consumer confidence, same-sex marriage, or fast food preferences – based upon a "scientifically

selected" sample of, say, 2,000 respondents in the United States. Such polls become especially important prior to elections. But how good are the results from an appropriately selected sample of 2,000 out of a population of over 225 million citizens of voting age?

- Television ratings estimate the number of people watching different programs nationwide, again based upon a relatively tiny sample of the population of television viewers.
- Environmental studies to assess the impact of global warming rely heavily on statistical analyses of long-term trends in weather data.
- Studies of the harmful side effects of specified drugs or the benefits of a particular type of vitamin.
- The link between war service and various specified illnesses.

1.4.2 Looking at Studies Critically

It is the mark of a truly intelligent person to be moved by statistics.
—George Bernard Shaw

Training in statistics allows us to look critically at studies reported in the press. Such investigations are often committed to selling a point of view. For example, a widely reported study by Waite and Gallagher (2000) alleged "marriage improves longevity." A statistically oriented person would want to know such things as:

- How were the data obtained?
- How was the analysis conducted?
- How many days, weeks, months or years, on the average, does marriage, supposedly, add to our lives?
- Are the results the same for men and women (some claim they apply to men only) and different age groups, etc.?

We would question, moreover, the implication in the article that marriage is what improves longevity. Even if married people live longer, is marriage the reason? Could it simply be that those in good health are more likely to get (and stay) married?

It is well known (although, perhaps, sometimes ignored by politicians) that statistical association found in, often crude, analyses of historical data should not be confused with proof of cause and effect. That is why many scientific investigations, such as the evaluation of the effectiveness of a proposed new drug, employ controlled investigations. In such studies, some subjects are often randomly selected to receive a placebo, instead of the treatments (e.g., different dosages of the new drug) being evaluated, and the results for the various groups are compared. Controlled studies are contrasted with observational studies (often conducted when controlled studies are not feasible) in Sidebar 1.2.

SIDEBAR 1.2: CONTROLLED VERSUS OBSERVATIONAL STUDIES

Controlled studies typically involve a random selection of sample units from a specified "frame"[1] and a random assignment of these units to different "treatments." Controlled studies are highly desirable in many situations, especially when the goal is to establish cause-and-effect relationships. They presumably eliminate bias due to self-selection (by subjects selecting a preferred treatment) and "confounding" variables (discussed shortly). The statistical design of experiments is one form of controlled study that has gained much popularity (Section 2.2.7) in business and industry.

In many applications, it is, however, inappropriate, or even unethical, to conduct a controlled study. Obvious examples are assessing the effect on health of cigarette smoking or of alcohol abuse, or the impact of alternative forms of punishment of past crimes on the incidence of future crimes. In such cases, we are often restricted to conducting an observational study.

An observational study, in contrast to a controlled study, involves the analysis of existing or new data that are typically obtained on some process whose major purpose is *not* that of providing data for analysis or decision making. Instead of randomizing, we have little or no say in how the treatments or exposures are allocated. It is, therefore, often difficult to draw definitive conclusions about cause-and-effect relationships and conclusions based on such studies must be carefully scrutinized. Yet such studies, when carefully planned and conducted, can still provide useful insights and suggest areas for further exploration.

The Framingham Heart Study is one of the best-known observational studies. Starting in 1948, some 5,000 volunteers from Framingham, a Boston, Massachusetts suburb, were monitored for decades through periodic physical examinations, blood tests, and in-depth interviews about their lifestyles. This study provided the first important clues into the effects of a person's diet, exercise, and smoking on cardiovascular health. Some of these associations were subsequently supported by further investigations (see Levy and Brink 2005). The overall study is continuing and now includes some of the original volunteers' children and grandchildren.

For further discussion contrasting these two types of studies, see Flanagan-Hyde (2006), Katz (2006, Chapter 1), and Utts (2005, Chapter 5).

It is not feasible to conduct a controlled study in the "marriage improves longevity" example. So we try, by statistical analyses of data on married and unmarried individuals, to adjust for "confounding" variables that also impact longevity and that may well be correlated with marital status, such as an individual's health or wealth. Adjusting for these variables is usually better than ignoring them, but it is unlikely to provide an unambiguous answer to the question posed. This is because it is usually not possible to identify, measure, and appropriately adjust for *all* confounding variables – and we don't know whether we have.

Campbell (1998), Gelman and Nolan (2002), Gigerenzer (2002), and Spirer, Spirer, and Jaffe (1998) use general interest examples, such as breast cancer

1. A sampling frame is a list or map or other means of identifying and accessing members or sampling units of the target population.

screening, AIDS counseling, domestic violence, experts on trial, and DNA fingerprinting to illustrate questionable analyses of historical data and common misuses of statistics.

1.4.3 Enumerative versus Analytic Studies

Statistical studies typically deal with data obtained in the past, or possibly the present, but the real interest is often about the future. In recognition of this, the renowned statistician, W. Edwards Deming urged us to distinguish between enumerative and analytic studies (see Deming 1975).[2]

Enumerative Studies In an enumerative study, one has a sampling frame that presumably provides a reasonable representation of the target population. A random sample is then taken from that frame. Examples of enumerative studies are:

- Estimating the proportion of defective units in a production lot, based on a random sample drawn from that lot.
- Estimating how many people in the United States are watching a TV show, using a sample of 2,000 homes randomly selected from among all U.S. homes.
- Using an "exit poll" of randomly selected voters on election day to estimate the proportion of voters who assert that they had voted for a particular candidate.

Analytic Studies In many actual situations in business and industry, one is, however, dealing, not with a static population, but with a dynamically changing one or with a future process. In such "analytic studies," one typically wishes to use the results from a current population or process and environment to draw inferences about what will happen in the future for a likely changing population or process and environment. For example:

- In assessing demand for a proposed product, one may use the results of a past survey of consumers to draw conclusions about future purchasing behavior. However, this behavior might change due, for example, to changes in the economy or new competitive offerings.
- In designing a product, one wishes to use the results from a sample of individually built prototype units of a new product built today to extrapolate performance under high-volume production some time in the future.
- In developing a model for credit scoring for loan approval, the payment performance of a past sample of credit applicants is used to decide to whom to give credit in the future. But the payment patterns of future applicants may differ from that of past applicants, especially if there has been a change in economic conditions.

2. Deming used the terminology "analytic problems" rather than "analytic studies."

Thus, the results of an analytic study on an existing process – and the statistical inferences we draw from these – apply to a future process only to the degree to which the two processes are the same.[3]

An understanding of the limitations of analytic studies strongly argues for making investigations as broad as possible. For example, in a study to predict the performance of a future production process, one should try to introduce, as much as possible, the type of variability that is expected to occur in future production – such as day-to-day differences in the environment, differences between operators, and raw material lot differences. This does not eliminate the unknown uncertainty of projecting into the future, but it should help reduce it.

1.4.4 The Role of Statistics in Addressing Life's Problems

People who don't count won't count.
—Anatole France

Statistics helps us address our daily life concerns, such as in the following examples (also see Sidebar 1.3):

- Proper interpretation of medical test results is essential for decision making by patients and caregivers. A series of tests may indicate that an 80-year-old patient has a 1-in-20 and a 1-in-3 chance of dying from a cancer within a year and within 10 years, respectively. Does this warrant a difficult operation, which carries its own estimated risks, given the patient's overall health condition?
- An understanding of how credit scores are constructed, using statistical models and past payment performance, can help us improve our own scores.
- Our utility bills, especially when there is a meter failure, may be estimated statistically from past usage. Such estimates could be slanted in favor of the utility. Our knowledge of statistics can help us achieve an equitable resolution.

SIDEBAR 1.3: STATISTICS IN EVERYDAY LIFE: ANOTHER EXAMPLE

One of us had an elderly mother living in the borough of Queens in New York City. About once a month, he took the train from Schenectady to New York to visit her. When he got there, he needed to make a decision – go to her house from the station by cab or by subway. From past experience, he knew that (adjusting for time of day and day of week) the median time for the cab ride is about 45 minutes. There is, however, a 1-in-10 chance that there are long traffic delays and the ride will take an hour and a half or more. The subway time is an hour, on the average, and 9 times out of 10, the ride does not exceed an hour and 10 minutes. So, forgetting about cost and comfort, he made the following decision. On the way down, he took a cab. This allowed him to increase the time that, in the long run, he

3. Statistical confidence intervals, for example, apply only to enumerative studies. If used in an analytic study, they must be considered as lower bounds reflecting only the *statistical* uncertainty.

could spend with mom. On the way back, he took the subway. This way he reduced the chances of missing the train home. Statistically speaking, he minimized the average time of the trip on the way down and reduced variability (so as to reduce the chances of missing the train) on the way back.

Also, we find it puzzling to have online weather forecasting software tell us that the probability of rain *some time* on a future day is 50%, and then have the same software assert that the probability of rain for *each* of 8 hours during that day is also 50%.

An appreciation of statistics also creates a passion for procuring as much relevant data as possible *before* making a decision. In purchasing a car, we avidly seek data on repair frequency to help us select a model. Or in deciding whether or not to see a movie, we not only seek out the opinions of others who have seen the movie, but also decide how much weight to give to each such assessment – based upon how well the assessor's past judgments correlate with our own.

1.5 BETTER DECISIONS REQUIRE BETTER DATA

> **The value of statistics in industry lies in its ability to influence decision making on a strategic level.**
> —Lynne Hare

Frequently, the ultimate goal in using statistics is to help make informed, and thereby better, decisions in the face of uncertainty. This involves procuring appropriate data and converting such data into useful information with the help of well-selected statistical and/or graphical methods.

Having "appropriate data" is key to the process. Such data are generally not just sitting there for us to analyze, but must be obtained, often by well-targeted and carefully planned investigations, Thus, helping ensure good data is a highly important (and not often fully appreciated) part of our job and is often more important than the statistical analysis itself. A study with good data can, after all, frequently be well analyzed by simple graphical methods. But the world's most sophisticated statistical analysis is unlikely to save a flawed investigation.

Frequently, a series of iterations is required before there is enough information to make a decision. After defining the problem, the available data are evaluated to determine what relevant information may be gleaned therefrom. This often leads to the need for gathering more and better data, which in turn are then analyzed. The process may be repeated as part of what Box, Hunter and Hunter (2005) call an "iterative learning process."

1.6 STATISTICAL THINKING

In contrast to deterministic thinking, statistical thinking recognizes the importance of variability. Its basic concepts are:

- All work occurs in a system of interconnected processes.
- Variation exists in all processes.
- Using data to understand and reduce variation is key to success.

Statistical thinking has broad applicability in situations where there is an opportunity to reduce variation, rework and waste, and for improving quality and productivity. Therefore, a disciplined and databased approach toward identifying the root causes of variability is required. At the higher levels of an organization, this involves clearly communicating the vision and the strategy, adapting a disciplined project management and review system, and taking process variation into consideration in setting goals.

The key ideas of statistical thinking were developed under the sponsorship of the Statistics Division of the American Society for Quality (ASQ) (see Britz, Emerling, Hare, Hoerl, Janis and Shade 2000). Also, Hoerl and Snee (2001) provide further information and examples of business process improvement based on statistical thinking. We illustrate this concept in various places and especially in discussing business processes (Section 12.3.9). And we note that, although statistical thinking was developed principally for business and industry applications, it has much more general applicability.

1.7 RELATED DISCIPLINES

1.7.1 Statistics and Mathematics

The Relationship Statistical methods are derived from mathematical theory; probability theory, in particular, is deeply rooted in mathematics.

At the same time, however, statistics – and especially applied statistics – is much more than a branch of mathematics. Like physics, which has much of its foundation in mathematics, statistics is a discipline of its own. There is much within the realm of statistics that has little to do with mathematics. The planning of studies to get good data and graphical data analysis are two examples.

So How Much Mathematics Do I Need? Practitioners may have little or no training in the theory underlying the tools that they are using. This may not be a serious problem for users of packaged software – as long as they are fully aware of the assumptions underlying these methods (generally a consequence of the theory) and know when to call for expert guidance.

Applied statisticians, in contrast, are likely to be working on more complex problems and frequently need to adapt a particular method to fit the problem and data at hand. This cannot be done without an intimate understanding of the theoretical foundations of statistics. Applied statisticians need to have, as a minimum, the basic knowledge provided by a one-year introductory course in mathematical statistics.[4]

4. Statisticians, mostly in academia, who are developing new methods, need to know additional theory. This, in turn, may require added course work in mathematics.

The application domain, in addition, often involves some inherent mathematical complexity. For example, optimization of a chemical process requires understanding of the differential equations that characterize the system.

1.7.2 Statistics and Computers

Statisticians have been closely associated with computers from the beginning of the "computer age." The renowned statistician John Tukey is credited with coining the computer terms "bit" (for binary digit) in 1946 and "software" (in a computing context) in 1958. We shall be discussing the role of computers and computations in the application of statistics in business and industry throughout this book, starting with Section 2.4.1.

Those who apply statistics need be proficient computer analysts. In that sense, they are hardly different from the rest of today's scientific community. In addition, they also need to be familiar with statistical software and have a strong interest in data storage, representation, and manipulation.

1.7.3 Statistics and OR/MS (Operations Research/Management Science)

OR/MS, also sometimes referred to as decision science or decision technology, is a field that is closely related to statistics. The Institute for Operations Research and the Management Sciences (INFORMS) states that "Members of the OR/MS profession apply scientific tools and methods to improve systems and operations and to assist in managerial decision making." OR/MS has been described as a scientific approach to analyzing problems and making decisions. This field emerged in an effort to leverage for nonmilitary applications – such as factory layout planning, manufacturing planning, transportation, and capacity planning – some of the highly successful logistics work done during World War II.

The INFORMS Internet site states that OR/MS professionals typically concern themselves with such problems as:

- What new sanitary facilities will be needed to serve the population of Sun Valley, Idaho in some designated future year?
- How can a dress manufacturer lay out patterns to minimize wasted material?
- How often should the sales force of a frozen yogurt company call on its customers?
- How many elevators should be installed in a new office building?

These questions sound very similar to ones that statisticians are likely to encounter. In fact, statistics is an important part of OR/MS and OR/MS professionals generally receive training in statistics. In turn, statisticians employ OR/MS methods extensively and need to be knowledgeable in them. It is often difficult to differentiate whether one is engaged in statistics or in OR/MS work – and quite unnecessary to do so.

We provide an introductory discussion of OR/MS methods in Section 12.7.

1.7.4 Other Domains

Problems that are addressed by statistics can also often be usefully analyzed by recently developed tools from artificial intelligence (AI), knowledge discovery in databases (KDD) and data mining, and soft computing. These methods include case-based reasoning, fuzzy logic, genetic algorithms, machine learning, and neural networks. Those who apply statistics, therefore, need, as a minimum, to be familiar with these tools (and, hopefully, those engaged in such work are similarly knowledgeable in statistics).

1.7.5 Some Closely Related Specialized Areas

In some areas of application, statistical and related quantitative approaches have led to knowledge areas of their own. Notable among these is actuarial science, focusing on the uncertainties associated with granting insurance and setting rates and terms for insurance policies (Section 11.7.6). In addition, there are such hybrid areas as biometrics, chemometrics (Section 13.6.7), econometrics, and environmetrics. These focus on the application of statistical, mathematical, OR/MS, and other quantitative approaches to problems in biology, chemistry, economics, and the environment, respectively. They require a high degree of knowledge both in the subject area and in statistics, and often also rely heavily on computer technology.

1.8 MAJOR TAKEAWAYS

- Statistics has been broadly defined as "the science of learning from data." It has important application in most fields of human endeavor from actuarial science to zoology – and notably in business and industry.
- Statistical literacy is an important part of being an educated citizen; it helps us look at studies reported in the press critically and make better decisions in addressing life's problems.
- Many studies are analytic (one is interested in what will happen in a future process) rather than enumerative (one wants to draw inferences about an existing population). A future process is likely to differ from the current process under investigation. We, therefore, need to make analytic studies as broad as possible and recognize their limitations.
- Better business decisions require better data. Planning to get good data is a key part of most statistical studies and may often be more important than the subsequent statistical analysis.
- Statistical thinking emphasizes that all work occurs in a system of interconnected processes, that variability is present in every process, and that such variability needs to be addressed by databased studies.
- Statistical methods are derived from mathematical theory. However, statistics is a discipline in its own right.

- Statisticians have been closely associated with computers from the beginning of the computer age.
- Statistics is closely related to OR/MS.
- Problems that are addressed by statistics can also often be usefully analyzed by recently developed tools from AI, KDD and data mining, and soft computing.
- Knowledge areas closely akin to statistics, such as actuarial science (and biometrics, chemometrics, econometrics, and environmetrics), have also evolved.

DISCUSSION QUESTIONS

General

1. Identify an application of statistics currently in the news. How did statistics contribute?

2. Identify a decision-making situation from your own experience for which a statistically based approach led, or might have led, to a better decision.

3. What, in your favorite sport, are some tactical or strategic decisions that can benefit from a statistical approach? How would you go about using such an approach?

4. We state in Section 1.4.1 that "environmental studies to assess the impact of global warming rely heavily on statistical analyses of long-term trends in weather data." Research this comment and elaborate.

5. Give some further examples of analytic studies.

6. We argue in Section 1.4.3 for making analytic "investigations as broad as possible." How would you go about doing this in the three cited examples of such studies?

7. In the description of Statistical Thinking in Section 1.6, it is asserted that "all work occurs in a system of interconnected processes." Elaborate on this statement and give some examples.

Technical

1. Consider again an application of statistics in the news. What do you think was the technical approach used? Critique, from a statistical perspective, the discussion of this application by the media.

2. What are some of the issues that needed to be considered in planning the study (Section 1.3.2) to obtain a valid assessment of whether birds can differentiate between mating calls and the sound of jet engines? How would you address these?

3. In discussing public opinion polls in Section 1.4.1, we raise the question "how good are the results from an appropriately selected sample of 2,000 out of a population of over 225 million citizens of voting age?" Respond to this question and indicate the degree of precision about a population that a sample of 2,000 can provide. What assumptions are being made and how might these be violated in practice?

4. Elaborate on how you would go about identifying potential confounding variables in a particular observational study and adjusting for them.

5. In each of the three examples of enumerative studies in Section 1.4.3, what is the target population and the sampling frame? How well does the frame represent the population? How might you select a random sample from the frame?

6. We assert in Section 1.4.4 that it is "puzzling to have online weather forecasting software tell us that the probability of rain *some time* on a future day is 50%, and then have the same software assert that the probability of rain for *each* of 8 hours during that day is also 50%." Explain.