Modern Engineering Statistics

THOMAS P. RYAN
Acworth, Georgia

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

Modern Engineering Statistics

BICENTENNIAL
1807 WILEY 2007
BICENTENNIAL

The Wiley Bicentennial–Knowledge for Generations

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce

William J. Pesce
President and Chief Executive Officer

Peter Booth Wiley

Peter Booth Wiley
Chairman of the Board

Modern Engineering Statistics

THOMAS P. RYAN
Acworth, Georgia

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Wiley Bicentennial Logo: Richard J. Pacifico

Library of Congress Cataloging-in-Publication Data:

Ryan, Thomas P., 1945–
Modern engineering statistics / Thomas P. Ryan.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-08187-7
1. Engineering–Statistical methods. I. Title.
TA340.R93 2007
620.0072–dc22

20060521558

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Preface

Statistical methods are an important part of the education of any engineering student. This was formally recognized by the Accreditation Board for Engineering and Technology (ABET) when, several years ago, education in probability and statistics became an ABET requirement for all undergraduate engineering majors. Specific topics within the broad field of probability and statistics were not specified, however, so colleges and universities have considerable latitude regarding the manner in which they meet the requirement. Similarly, ABET's *Criteria for Accrediting Engineering Programs*, which were to apply to evaluations during 2001–2002, were not specific regarding the probability and statistics skills that engineering graduates should possess.

Engineering statistics courses are offered by math and statistics departments, as well as being taught within engineering departments and schools. An example of the latter is The School of Industrial and Systems Engineering at Georgia Tech, whose list of course offerings in applied statistics rivals that of many statistics departments.

Unfortunately, many engineering statistics courses have not differed greatly from mathematical statistics courses, and this is due in large measure to the manner in which many engineering statistics textbooks have been written. This textbook makes no pretense of being a "math stat book." Instead, my objective has been to motivate an appreciation of statistical techniques, and to do this as much as possible within the context of engineering, as many of the datasets that are used in the chapters and chapter exercises are from engineering sources. I have taught countless engineering statistics courses over a period of two decades and I have formulated some specific ideas of what I believe should be the content of an engineering statistics course. The contents of this textbook and the style of writing follow accordingly.

Statistics books have been moving in a new direction for the past fifteen years, although books that have beaten a new path have often been overshadowed by the sheer number of books that are traditional rather than groundbreaking.

The optimum balance between statistical thinking and statistical methodology can certainly be debated. Hoerl and Snee's book, *Statistical Thinking*, which is basically a book on business statistics, stands at one extreme as a statistics book that emphasizes the "big picture" and the use of statistical tools in a broad way rather than encumbering the student with an endless stream of seemingly unrelated methods and formulas.

This book might be viewed as somewhat of an engineering statistics counterpart to the Hoerl and Snee book, as statistical thinking is emphasized throughout, but there is also a solid dose of contemporary statistical methodology.

This book has many novel features, including the connection that is frequently made (but hardly ever illustrated) between hypothesis tests and confidence intervals. This connection is illustrated in many places, as I believe that the point cannot be overemphasized.

I have also written the book under the assumption that statistical software will be used (extensively). A somewhat unusual feature of the book is that computing equations are kept to a minimum, although some have been put in chapter appendixes for readers interested in seeing them. MINITAB is the most frequently used statistical software for college and university courses. Minitab, Inc. has been a major software component of the Six Sigma movement and has made additions to the MINITAB software to provide the necessary capabilities for Six Sigma work. Such work has much in common with the field of engineering statistics and with the way that many engineers use statistics. Therefore, MINITAB is heavily relied on in this book for illustrating various statistical analyses, although JMP from SAS Institute, Inc. is also used.

This is not intended, however, to be a book on how to use MINITAB or JMP, since books have been written for that purpose. Nevertheless, some MINITAB code is given in certain chapters and especially at the textbook Website to benefit users who prefer to use MINITAB in command mode. Various books, including the *MINITAB User's Guide*, have explained how to use MINITAB in menu mode, but not in command mode. The use of menu mode is of course appropriate for beginners and infrequent users of MINITAB, but command mode is much faster for people who are familiar with MINITAB and there are many users who still use command mode. Another advantage of command mode is that when the online help facility is used to display a command, all of the subcommands are also listed, so the reader sees all of the options, whereas this view is not available when menu mode is used. Rather, the user has to navigate through the various screens and mentally paste everything together in order to see the total capability relative to a particular command.

There are, however, some MINITAB routines for which menu mode is preferable, due in part to the many subcommands that will generally be needed just to do a standard analysis. Thus, menu mode does have its uses.

Depending on how fast the material is covered, the book could be used for a two-semester course as well as for a one-semester course. If used for the latter, the core material would likely be all or parts of Chapters 1–6, 8, 11, 12, and 17. Some material from Chapters 7 and 14 might also be incorporated, depending on time constraints and instructor tastes.

For the second semester of a two-semester course, Chapters 7, 9, 10, 13, 14, and 15 and/or 16 might be covered, perhaps with additional material from Chapters 11 and 12 that could not be covered in the first semester. The material in Chapter 12 on Analysis of Means deserves its place in the sun, especially since it was developed for the express purpose of fostering communication with engineers on the subject of designed experiments. Although Chapter 10 on mechanistic models and Chapter 7 on tolerance intervals and prediction intervals might be viewed as special topics material, it would be more appropriate to elevate these chapters to "core material chapters," as this is material that is very important for engineering students. At least some of the material in Chapters 15 and 16 might be covered, as time permits. Chapter 16 is especially important as it can help engineering students and others realize that nonparametric (distribution-free) methods will often be viable alternatives to the better-known parametric methods.

There are reasons for the selected ordering of the chapters. Standard material is covered in the first six chapters and the sequence of those chapters is the logical one. Decisions had to be made starting with Chapter 7, however. Although instructors might view this as a special topics chapter as stated, there are many subject matter experts who believe that tolerance intervals and prediction intervals should be taught in engineering statistics courses. Having a chapter on tolerance intervals and prediction intervals follow a chapter on confidence intervals is reasonable because of the relationships between the intervals and the need for this to be understood. Chapter 9 is an extension of Chapter 8 into multiple linear regression and it is reasonable to have these chapters followed by Chapter 10 since nonlinear regression is used in this chapter. In some ways it would be better if the chapter followed Chapter 14 since reliability models are used, but the need to have it follow Chapters 8 and 9 seems more important. The regression chapters should logically precede the chapter on design of experiments, Chapter 12, since regression methods should be used in analyzing data from designed experiments. Processes should ideally be in a state of statistical control when designed experiments are performed, so the chapter on control chart methods, Chapter 11, should precede Chapter 12. Chapters 13 and 14 contain subject matter that is important for engineering and Chapters 15 and 16 consider topics that are generally covered in a wide variety of introductory type statistics texts. It is useful for students to be able to demonstrate that they have mastered the tools they have learned in any statistics course by knowing which tool(s) to use in a particular application after all of the material has been presented. The exercises in Chapter 17 provide students with the opportunity to demonstrate that they have acquired such skill.

The book might also be used for self-study, aided by the Answers to Selected Exercises, which is sizable and detailed. A separate *Solutions Manual* with solutions to all of the chapter exercises is also available. The data in the exercises, including data in MINITAB files (i.e., the files with the `.MTW` extension), can be found at the website for the text: ftp://ftp.wiley.com/public/ sci_med/engineering_statistics.

I wish to gratefully acknowledge the support and assistance of my editor, Steve Quigley, associate editor Susanne Steitz, and production editor Rosalyn Farkas, plus various others, including the very large number of anonymous reviewers who reviewed all or parts of the manuscript at various stages and made helpful comments.

THOMAS P. RYAN

Acworth, Georgia
May 2007

CHAPTER 1

Methods of Collecting and Presenting Data

People make decisions every day, with decision-making logically based on some form of data. A person who accepts a job and moves to a new city needs to know how long it will take him/her to drive to work. The person could guess the time by knowing the distance and considering the traffic likely to be encountered along the route that will be traveled, or the new employee could drive the route at the anticipated regular time of departure for a few days before the first day of work.

With the second option, an experiment is performed, which if the test run were performed under normal road and weather conditions, would lead to a better estimate of the typical driving time than by merely knowing the distance and the route to be traveled.

Similarly, engineers conduct statistically designed experiments to obtain valuable information that will enable processes and products to be improved, and much space is devoted to statistically designed experiments in Chapter 12.

Of course, engineering data are also available without having performed a designed experiment, but this generally requires a more careful analysis than the analysis of data from designed experiments. In his provocative paper, "Launching the Space-Shuttle *Challenger*—Disciplinary Deficiencies in the Analysis of Engineering Data," F. F. Lighthall (1991) contended that "analysis of field data and reasoning were flawed" and that "staff engineers and engineering managers . . . were unable to frame basis questions of covariation among field variables, and thus unable to see the relevance of routinely gathered field data to the issues they debated before the *Challenger* launch." Lighthall then states "Simple analyses of field data available to both Morton Thiokol and NASA at launch time and months before the *Challenger* launch are presented to show that the arguments against launching at cold temperatures could have been quantified. . . ." The author's contention is that there was a "gap in the education of engineers." (Whether or not the *Columbia* disaster will be similarly viewed by at least some authors as being a deficiency in data analysis remains to be seen.)

Perhaps many would disagree with Lighthall, but the bottom line is that failure to properly analyze available engineering data or failure to collect necessary data can endanger lives—on a space shuttle, on a bridge that spans a river, on an elevator in a skyscraper, and in many other scenarios.

Intelligent analysis of data requires much thought, however, and there are no shortcuts. This is because analyzing data and solving associated problems in engineering and other areas is more of an art than a science. Consequently, it would be impractical to attempt to give a specific step-by-step guide to the use of the statistical methods presented in succeeding chapters, although general guidelines can still be provided and are provided in subsequent chapters. It is desirable to try to acquire a broad knowledge of the subject matter and position oneself to be able to solve problems with powers of reasoning coupled with subject matter knowledge.

The importance of avoiding the memorization of rules or steps for solving problems is perhaps best stated by Professor Emeritus Herman Chernoff of the Harvard Statistics Department in his online algebra text, *Algebra 1 for Students Comfortable with Arithmetic* (`http://www.stat.harvard.edu/People/Faculty/Herman_Chernoff/Herman_Chernoff_Algebra_1.pdf`).

> Memorizing rules for solving problems is usually a way of avoiding understanding. Without understanding, great feats of memory are required to handle a limited class of problems, and there is no ability to handle new types of problems.

My approach to this issue has always been to draw a rectangle on a blackboard and then make about 15–20 dots within the rectangle. The dots represent specific types of problems; the rectangle represents the body of knowledge that is needed to solve not only the types of problems represented by the dots, but also any type of problem that would fall within the rectangle. This is essentially the same as what Professor Chernoff is saying.

This is an important distinction that undoubtedly applies to any quantitative subject and should be understood by students and instructors, in general.

Semiconductor manufacturing is one area in which statistics is used extensively. International SEMATECH (SEmiconductor MAnufacturing TECHnology), located in Austin, Texas, is a nonprofit research and development consortium of the following 13 semiconductor manufacturers: Advanced Micro Devices, Conexant, Hewlett-Packard, Hyundai, Infineon Technologies, IBM, Intel, Lucent Technologies, Motorola, Philips, STMicroelectronics, TSMC, and Texas Instruments. Intel, in particular, uses statistics extensively.

The importance of statistics in these and other companies is exemplified by the *NIST/SEMATECH e-Handbook of Statistical Methods* (Croarkin and Tobias, 2002), a joint effort of International SEMATECH and NIST (National Institute of Standards and Technology), with the assistance of various other professionals. The stated goal of the handbook, which is the equivalent of approximately 3,000 printed pages, is to provide a Web-based guide for engineers, scientists, businesses, researchers, and teachers who use statistical techniques in their work. Because of its sheer size, the handbook is naturally much more inclusive than this textbook, although there is some overlap of material. Of course, the former is not intended for use as a textbook and, for example, does not contain any exercises or problems, although it does contain case studies. It is a very useful resource, however, especially since it is almost an encyclopedia of statistical methods. It can be accessed at `www.itl.nist.gov/div898/handbook` and will henceforth often be referred to as the *e-Handbook of Statistical Methods* or simply as the *e-Handbook*.

There are also numerous other statistics references and data sets that are available on the Web, including some general purpose Internet statistics textbooks. Much information, including many links, can be found at the following websites: `http://www.utexas.edu/cc/stat/world/softwaresites.html` and `http://my.execpc.com/~helberg/statistics.html`. The *Journal of Statistics Education* is a free, online statistics publication devoted to statistics education. It can be found at `http://www.amstat.org/publications/jse`.

Statistical education is a two-way street, however, and much has been written about how engineers view statistics relative to their work. At one extreme, Brady and Allen (2002) stated: “There is also abundant evidence—for example, Czitrom (1999)—that most practicing engineers fail to consistently apply the formal data collection and analysis techniques that they have learned and in general see their statistical education as largely irrelevant to their professional life.” (It is worth noting that the first author is an engineering manager in industry.) The Accreditation Board for Engineering and Technology (ABET) disagrees with this sentiment and several years ago decreed that *all* engineering majors must have training in probability and statistics. Undoubtedly, many engineers would disagree with Brady and Allen (2002), although historically this has been a common view.

One relevant question concerns the form in which engineers and engineering students believe that statistical exposition should be presented to them. Lenth (2002), in reviewing a book on experimental design that was written for engineers and engineering managers and emphasizes hand computation, touches on two extremes by first stating that “. . . engineers just will not believe something if they do not know how to calculate it . . .,” and then stating “After more thought, I realized that engineers are quite comfortable these days—in fact, far too comfortable—with results from the blackest of black boxes: neural nets, genetic algorithms, data mining, and the like.”

So have engineers progressed past the point of needing to see how to perform all calculations that produce statistical results? (Of course, a world of black boxes is undesirable.) This book was written with the knowledge that users of statistical methods simply do not perform hand computation anymore to any extent, but many computing formulas are nevertheless given for interested readers, with some formulas given in chapter appendices.

1.1 OBSERVATIONAL DATA AND DATA FROM DESIGNED EXPERIMENTS

Sports statistics are readily available from many sources and are frequently used in teaching statistical concepts. Assume that a particular college basketball player has a very poor free throw shooting percentage, and his performance is charted over a period of several games to see if there is any trend. This would constitute observational data—we have simply observed the numbers. Now assume that since the player’s performance is so poor, some action is taken to improve his performance. This action may consist of extra practice, visualization, and/or instruction from a professional specialist. If different combinations of these tasks were employed, this could be in the form of a designed experiment. In general, if improvement is to occur, there should be experimentation. Otherwise, any improvement that seems to occur might be only accidental and not be representative of any real change.

Similarly, W. Edwards Deming (1900–1993) coined the terms *analytic studies* and *enumerative studies* and often stated that “statistics is prediction.” He meant that statistical methods should be used to improve future products, processes, and so on, rather than simply “enumerating” the current state of affairs as is exemplified, for example, by the typical use of sports statistics. If a baseball player’s batting average is .274, does that number tell us anything about what the player should do to improve his performance? Of course not, but when players go into a slump they try different things; that is, they *experiment*. Thus, experimentation is essential for improvement.

This is not to imply, however, that observational data (i.e., enumerative studies) have no value. Obviously, if one is to travel/progress to “point B,” it is necessary to know the starting point, and in the case of the baseball player who is batting .274, to determine if the starting point is one that has some obvious flaws.

When we use designed experiments, we must have a way of determining if there has been a “significant” change. For example, let’s say that an industrial engineer wants to determine if a new manufacturing process is having a significant effect on throughput. He/she obtains data from the new process and compares this against data that are available for the old process. So now there are two sets of data and information must be extracted from those two sets and a decision reached. That is, the engineer must compute *statistics* (such as averages) from each set of data that would be used in reaching a decision. This is an example of *inferential statistics*, a subject that is covered extensively in Chapters 4–15.

DEFINITION

A *statistic* is a summary measure computed from a set of data.

One point that cannot be overemphasized (so the reader will see it discussed in later chapters) is that experimentation should generally not be a one-time effort, but rather should be repetitive and sequential. Specifically, as is illustrated in Figure 1.1, exprimentation should in many applications be a never-ending learning process. Mark Price has the highest free throw percentage in the history of the National Basketball Association (NBA) at .904, whereas in his four-year career at Georgia Tech his best year was .877 and he does not even hold the career Georgia Tech field goal percentage record (which is held by Roger Kaiser at .858). How could his professional percentage be considerably higher than his college percentage, despite the rigors of NBA seasons that are much longer than college seasons? Obviously, he had to experiment to determine what worked best for him.

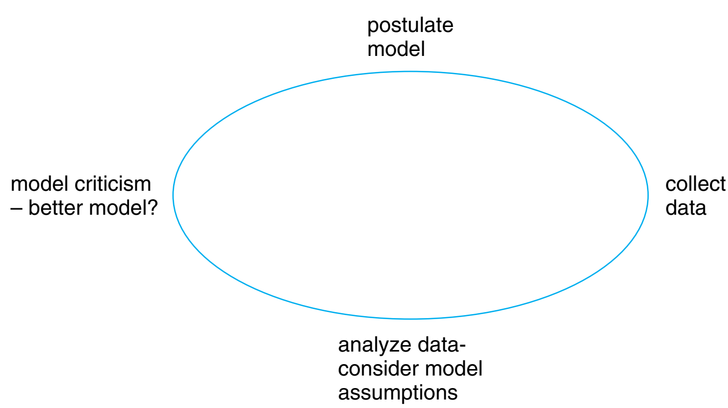

Figure 1.1 Repetitive nature of experimentation.

1.2 POPULATIONS AND SAMPLES

Whether data have been obtained as observational data or from a designed experiment, we have obtained a *sample* from a *population*.

DEFINITION

A *sample* is a subset of observations obtained from a larger set, termed a *population*.

To the layperson, a population consists of people, but a statistical population can consist of virtually anything. For example, the collection of desks on a particular college campus could be defined as a population. Here we have a *finite* population and one could, for example, compute the average age of desks on campus. What is the population if we toss a coin ten times and record the number of heads? Here the population is conceptually infinite as it would consist of all of the tosses of the coin that could be made. Similarly, for a manufacturing scenario the population could be all of the items of a particular type produced by the current manufacturing process—past, present, and future.

If our sample is comprised of observational data, the question arises as to how the sample should be obtained. In particular, should we require that our sample be *random*, or will we be satisfied if our sample is simply *representative*?

DEFINITION

A *random sample* of a specified size is one for which every possible sample of that size has the same chance of being selected from the population.

A simple example will be given to illustrate this concept. Suppose a population is defined to consist of the numbers 1, 2, 3, 4, 5, and 6, and you wish to obtain a random sample of size two from this population. How might this be accomplished? What about listing all of the possible samples of size two and then randomly selecting one? There are 15 such samples and they are given below.

12	15	24	34	45
13	16	25	35	46
14	23	26	36	56

Following the definition just given, a random sample of size two from this population is such that each of the possible samples has the same probability of being selected.

There are various ways to obtain a random sample, once a *frame*, a list of all of the elements of a population, is available. Obviously, one approach would be to use a software program that generates random numbers. Another approach would be to use a random number table such as Table A at the end of the book. That table could be used as follows. In general, the elements in the population would have to be numbered in some way. In this example the elements are numbers, and since the numbers are single-digit numbers, only one column of Table A need be used. If we arbitrarily select the first column in the first set of four columns, we could proceed down that column; the first number observed is 1 and the second is 5. Thus, our sample of size two would consist of those two numbers.

Now how would we proceed if our population is defined to consist of all transistors of a certain type manufactured in a given day at a particular facility? Could a random sample be obtained?

In general, to obtain a random sample we do need a frame, which as has been stated is a list of all of the elements in the population. It would certainly be impractical to "number" all of the transistors so that a random sample could be taken. Consequently, a *convenience sample* is frequently used instead of a random sample. The important point is that the sample should be representative, and more or less emulate a random sample since common statistical theory is based on the assumption of random sampling.

For example, we might obtain samples of five units from an assembly line every 30 minutes. With such a sampling scheme, as is typical when control charts (see Chapter 11) are constructed, every item produced will not have the same probability of being included in any one of the samples with this *systematic sampling* approach, as it is called.

Such a sampling approach could produce disastrous results if, unbeknown to the person performing the sampling, there was some cyclicality in the data. This was clearly shown in McCoun (1949, 1974) in regard to a tooling problem. If you imagine data that would graph approximately as a sine curve, and if the sampling coincided with the periodicity of the curve, the variability of the data could be greatly underestimated and the trend that would be clearly visible for the entire set of data would be hidden.

As a second example, assume that every twenty-first unit of a particular product is nonconforming. If samples of size three happen to be selected in such a way (perhaps every 15 minutes) that one nonconforming unit is included in each sample, the logical conclusion would be that one out of every three units produced is nonconforming, instead of one out of twenty-one.

Because of these possible deleterious effects, how can we tell whether or not convenience samples are likely to give us a true picture of a particular population? We cannot, unless we have some idea as to whether there are any patterns or trends in regard to the units that are produced, and we may not know this unless we have a time sequence plot of historical data.

Another point to keep in mind is that populations generally change over time, and the change might be considerable relative to what we are trying to estimate. Hence, a sample that is representative today may not be representative six months later. For example, the racial composition of a particular high school could change considerably in a relatively short period of time, as could the sentiments of voters in a particular district regarding who they favor for political office.

Consequently, it is highly desirable to acquire a good understanding of the processes with which you will be working before using any "routine" sampling procedure.

1.3 VARIABLES

When we obtain our sample, we obtain data values on one or more *variables*. For example, many (if not most) universities use regression modeling (regression is covered in Chapters 8 and 9) as an aid in predicting what a student's GPA would be after four years if the student were admitted, and use that predicted value as an aid in deciding whether or not to admit the student. The sample of historical data that is used in developing the model would logically have the student's high school grade point average, aptitude test scores, and perhaps several other variables.

If we obtained a random sample of students, we would expect a list of the aptitude test scores, for example, to vary at random; that is, the variable would be a *random variable*. A mathematically formal definition of a random variable that is usually found in introductory statistics books will be avoided here in favor of a simpler definition. There are two general types of random variables and it is important to be able to distinguish between them.

DEFINITIONS

A *random variable* is a statistic or an individual observation that varies in a random manner. A *discrete* random variable is a random variable that can assume only a finite or countably infinite number of possible values (usually integers), whereas a *continuous* random variable is one that can theoretically assume any value in a specified interval (i.e., continuum), although measuring instruments limit the number of decimal places in measurements.

The following simple example should make clear the idea of a random variable. Assume that an experiment is defined to consist of tossing a single coin twice and recording the number of heads observed. The experiment is to be performed 16 times. The random variable is thus the number of heads, and it is customary to have a random variable represented by an alphabetical (capital) letter. Thus, we could define

$$X = \text{the number of heads observed in each experiment}$$

Assume that the 16 experiments produce the following values of X:

$$0 \quad 2 \quad 1 \quad 1 \quad 2 \quad 0 \quad 0 \quad 1 \quad 2 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 2 \quad 0$$

There is no apparent pattern in the sequence of numbers so based on this sequence we would be inclined to state that X (apparently) *varies* in a *random* manner and is thus a random variable.

Since this is an introductory statistics text, the emphasis is on *univariate* data; that is, data on a single random variable. It should be kept in mind, however, that the world is essentially *multivariate*, so any student who wants to become knowledgeable in statistics must master both univariate and multivariate methods. In statistics, "multivariate" refers to more than one response or dependent variable, not just more than one variable of interest; researchers in other fields often use the term differently, in reference to the independent variables. The graphs in Section 1.4.4 are for two variables.

1.4 METHODS OF DISPLAYING SMALL DATA SETS

We can enumerate and summarize data in various ways. One very important way is to graph data, and to do this in as many ways as is practical. Much care must be exercised in the use of graphical procedures, however; otherwise, the impressions that are conveyed could be very misleading. It is also important to address at the outset what one wants a graph to show as well as the intended audience for the graph.

We consider some important graphical techniques in the next several sections. There are methods that are appropriate for displaying essential information in large data sets and there are methods for displaying small data sets. We will consider the latter first.

■ EXAMPLE 1.1

The data in Table 1.1, compiled by GaSearch Pricing Data for November 2001, is a sample of natural gas purchasers in the state of Texas with over 1,500,000 Mcf throughput.

Data

TABLE 1.1 Gas Pricing Data for November 2001

Purchaser Name	Average Cost per Mcf
Amoco Production Corp.	\$2.78
Conoco Inc.	2.76
Duke Energy Trading and Marketing	2.73
Exxon Corporation	2.71
Houston Pipe Line Co.	3.07
Mitchell Gas Services LP	2.95
Phillips Petroleum Co.	2.65
Average Top State of Texas Production	2.79

Discussion

With data sets as small as this one, we really don't need to rely on summary measures such as the average because we can clearly see how the numbers vary; we can quickly see the largest and smallest values, and we could virtually guess the average just by looking at the numbers. Thus, there is no need to "summarize" the data, although a graphical display or two could help identify any outliers (i.e., extreme observations) and one such graphical display of these data is given in Section 1.4.6.

There is, however, a need to recognize that sampling variability exists whenever one takes a sample from a population, as was done in this case. That is, if a different sample of gas purchasers had been selected, the largest and smallest values would have been different, and so would any other computed statistic. Sampling variability is introduced in Chapter 4 and plays an important role in the material presented in subsequent chapters. ■

1.4.1 Stem-and-Leaf Display

Few data sets are as small as this one, however, and for data sets that are roughly two to three times this size, we need ways of summariing the data, as well as displaying the data. Many college students have had the experience of seeing the final exam scores for their