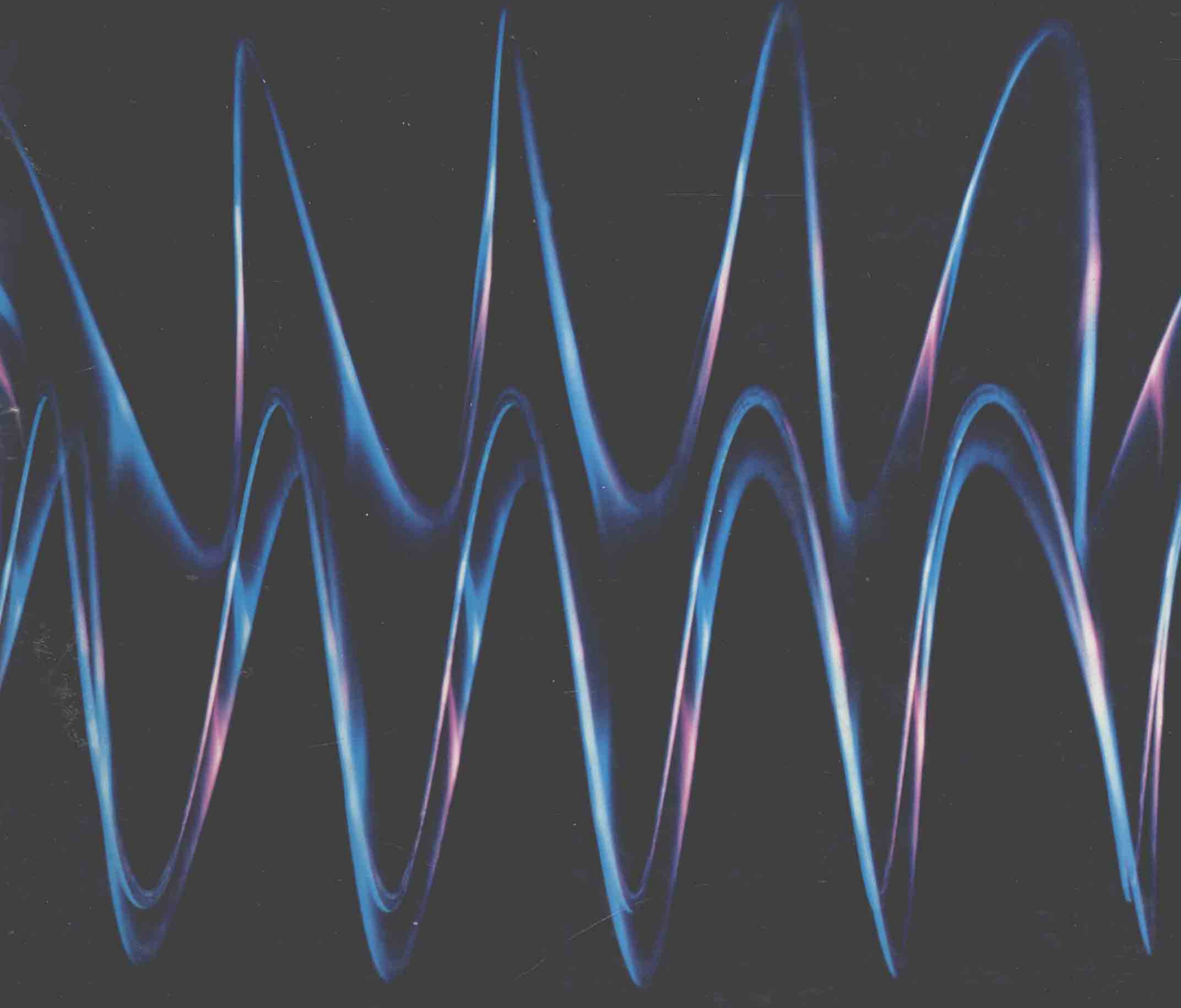
Applied Regression Analysis for Business and Economics

Terry E. Dielman



Applied Regression Analysis

for Business and Economics

Terry E. Dielman

M.J. Neeley School of Business
Texas Christian University

The Duxbury Series in Statistics and Decision Sciences
PWS-KENT Publishing Company
Boston

PWS-KENT
Publishing Company

Sponsoring Editor: Michael Payne
Production Editor: Chris Crochetière
Manufacturing Coordinator: Margaret Sullivan Higgins
Cover Designer: Lindgren Design Associates
Cover Photographer: Jake Armour
Interior Designer: George McLean
Typesetter: The Alden Press
Printer/Binder: Arcata Graphics/Halliday
Cover Printer: John P. Pow Company, Inc.

PWS-KENT Publishing Company is a division of Wadsworth, Inc.

Printed in the United States of America.

1 2 3 4 5 6 7 8 9 — 95 94 93 92 91 90

International Student Edition ISBN 0-534-98327-8

Library of Congress Cataloging-in-Publication Data

Dielman, Terry E., 1952-
Applied regression analysis for business
and economics.
Includes bibliographical references (p. 565)
and index.
1. Economics—Statistical methods. 2. Commercial
statistics. 3. Regression analysis.
I. Title.
HB137.D54 1990 330'.01'519536 90-7764
ISBN 0-534-92238-4

Preface

Applied Regression Analysis for Business and Economics is designed for a one-semester course in regression analysis. This text is primarily intended for business and economics undergraduates or MBAs. The goal of the text is to present regression techniques in a way that avoids unnecessary mathematical rigor. The emphasis is on understanding the assumptions of the regression model, knowing how to validate a selected model for these assumptions, knowing when and how regression might be useful in a business setting, and understanding computer output from commonly encountered statistical packages. The text utilizes the statistical packages of SAS and MINITAB. Other statistical packages containing regression procedures could also be used with little effort. Brief introductions to both MINITAB and SAS are presented in Appendix D.

Actual data drawn from various sources are used throughout the book in the examples and exercises. When data are simulated, an attempt has been made to provide realistic data and situations in which these data might occur. In this way the relevance of the techniques being presented is highlighted for students.

To use the text, little mathematical expertise is necessary beyond basic college algebra. Calculus is only used in Chapter 3 to derive the equations for the least squares estimates of the simple regression coefficients. If desired, this derivation using calculus could be excluded without loss of continuity. No knowledge of linear algebra is assumed in the text. Appendix C does provide a summary of matrices and matrix operations and a brief introduction to the use of matrices in presenting the least squares method for the interested reader. An introductory (or first semester) course in statistics is assumed. Chapter 2 does, however, contain a brief review of most of the concepts covered in an introductory statistics course.

Chapter 3 introduces simple linear regression and the use of MINITAB and SAS regression routines. Chapter 4 provides the extension to multiple linear regression. Chapter 5 discusses the implications of violations of assumptions of the regression model. Also, Chapter 4 presents ways to recognize possible violations and corrections for violations. In Chapter 6, the use of indicator and interaction variables is described. Also, Chapter 6 presents a brief look at analysis of variance models. One-way analysis of variance and its relationship to regression with indicator variables is discussed. Chapter 6 concludes with an examination of randomized block designs and two-way complete factorial designs. Chapter 7 concentrates on the unique aspects of regression using time-series data. Chapter 8 discusses several techniques used to aid in selecting explanatory variables for the regression.

This book, in the form of notes, has been used for several semesters in the regression analysis course I teach at Texas Christian University. The notes in their final version were also used by Professor E.L. Rose in her regression analysis course at the University of Southern California. I would like to thank Professor Rose and the students from both universities for their comments on various versions of the book and for putting up with a set of notes rather than a real text in class. Thanks also to reviewers at various stages of the manuscript, including:

Bruce Cooil
Vanderbilt University

Paul Eaton
Northern Illinois University

Ravindra Khattree
North Dakota State University

Sue Leurgans
Ohio State University

Steven Lippman
University of California-Los Angeles

Robert R. Logan
University of Alaska-Fairbanks

Mark McNulty
Kansas State University

Robert Mogull
California State University-Sacramento

Lewis H. Shoemaker
Millersville University

Leroy Simmons
Loyola College

Peg Young
George Mason University

Special thanks to Professor Khattree who provided very detailed and helpful comments on the manuscript.

Thanks also to Barbara Pfaffenberger for useful and detailed comments on early chapters, Roger Pfaffenberger for answering questions about the text when he had better things to do, and to my typist, Marty Burkhard, for enduring many revisions.

I would also like to express my appreciation to the staff and associates of PWS-KENT Publishing, especially Michael Payne, Susan Hankinson, Marcia Cole, and Chris Crochetière.

Finally, thanks to my wife Karen and my daughter Kelly for putting up with the time I spent on this book and for their encouragement throughout the process.

Contents

CHAPTER

1 An Introduction to Regression Analysis

Computers and telecommunications equipment have buried the present-day manager under a mountain of data. Although the purpose of these data is to assist the manager in the decision-making process, the corporate executive who faces the task of juggling data on many variables may find him- or herself at a loss when attempting to make sense of such information. The decision-making process is further complicated by the dynamic elements in the business environment and the complex interrelationships among these elements.

This text has been prepared to give the manager (or future manager) a tool for examining possible relationships between two or more variables. For example, sales and advertising expenditures are two variables that are commonly thought to be related. When a soft drink company increases its advertising expenditures by paying Michael Jackson $1 billion to do its advertisements, it expects that this outlay will increase sales. In general, when decisions on advertising expenditures of millions of dollars are involved, it would be comforting to have some evidence that, in the past, increased advertising expenditures indeed led to increased sales.

Another example is the relationship between the selling price of a house and its square footage. When a new house is to be listed for sale, how should the price be determined? Is a 4000 ft^2 house worth twice as much as a 2000 ft^2 house? What other factors might be involved in the pricing of houses and how should these factors be included when determining price?

In a study of absenteeism at a large manufacturing plant, management may feel that several variables affect productivity. These variables might include job complexity, base pay, the number of years a worker has been with the plant, and the age of that worker. If high absenteeism can cost the company thousands of dollars, then the importance of identifying its associated factors becomes clear.

Perhaps the most important analytic tool for examining the relationships between two or more variables is regression analysis. *Regression analysis* is a statistical technique for developing an equation that describes the relationship between two or more variables. One variable is specified to be the *dependent* variable or the variable to be explained. The other one or more variables are called the *independent* or *explanatory* variables. Using the previous examples, the soft drink firm would identify sales as the dependent variable and advertising expenditure as the explanatory variable. The real estate firm would choose selling price as the dependent variable, and size as the explanatory variable that explains variations in selling price from house to house.

There are several reasons why business researchers might want to know how certain variables are related. The retail firm may want to develop the variables' interrelationship for control purposes. In other words, how much advertising is necessary to achieve a certain level of sales? An equation that expressed the relationship between sales and advertising would be useful in answering this question. For the real estate firm, the relationship might be used to assign prices to houses entering the market. The management of the manufacturing firm would like to know what variables are most highly related to absenteeism so that the absenteeism rate can be lowered. Thus, the typical reasons for developing an equation that relates two or more variables are (1) for describing the relationship, (2) for control purposes, and (3) for prediction.

Much statistical analysis is a multistage process of trial and error. A good deal of exploratory work must be done to select the appropriate variables for study and to determine the relationships between or among them. This requires that a variety of statistical tests and other procedures be performed and sound judgements made before one arrives at a satisfactory choice of dependent and explanatory variables. The emphasis in this text will be on this multistage process rather than on the computations themselves or an in-depth study of the theory behind the techniques presented. In this sense the text is directed at the applied researcher or the user of statistics.

Except for a few preparatory examples, we will assume that a computer is available to the reader to perform the actual computations. The use of statistical software thus frees the user to concentrate on the multistage "model-building" process.

Most examples will use illustrative computer output to present the results. The two statistical software packages used will be MINITAB and SAS. The main frame version of SAS and PC version 6.1 of MINITAB have been used to develop the materials, and both packages are available in PC and mainframe versions. The output from these two packages is fairly standard and easily understood. Many of the exercises at the end of most chapters are intended to be done with the aid of a computer. Any statistical software package with a regression routine can be used for this purpose. Some of the options available in MINITAB and SAS may not be present in other packages, but this should present no problem in completing the exercises.

A floppy disk has been provided with data sets used in most of the analyses in this text. In each problem where data sets are provided, the MINITAB and SAS commands to read the data will be given. In the commands to read the data files, no disk drive has been specified; only the file name is used. A disk drive may need to be indicated on the commands; for example, `READ 'A:MFINC.DAT' C1` rather than `READ 'MFINC.DAT' C1`. Commands used in MINITAB and SAS to produce appropriate statistical output will be noted throughout the text where appropriate. This book, however, is not intended to provide full information on the use of these statistical packages. For a brief discussion of the use of MINITAB and SAS, see Appendix D. For further information on MINITAB and SAS, the interested reader is referred to one of the following references:

Miller, R. 1988. *MINITAB handbook for business and economics*. Boston: PWS-Kent Publishing Co.

Ryan, B. F., B. L. Joiner, and T. A. Ryan. 1985. *MINITAB handbook*, 2d ed. Boston: PWS-Kent Publishing Co.

SAS Introductory Guide. 3d ed. Cary, N.C.: SAS Institute, Inc.

CHAPTER

2 Review of Basic Statistical Concepts

2.1 Introduction

The purpose of this chapter is to summarize and review many of the basic statistical concepts taught in an introductory statistics course. For the most part, introductory courses in statistics deal with three main areas of interest: descriptive statistics, probability, and statistical inference.

Typically, the problem in question in statistics is one of studying a particular population. A *population*, for purposes of this text, may be defined as the collection of all items of interest to a researcher. The researcher may want to study the sales figures in 1985 for firms in a particular industry, the rates of return of public utility firms, or the lifetimes of a new brand of automobile tires. But because of time limitations, cost, or the destructive nature of testing, not all elements in a population can be examined. Instead, a subset of the population, called a *sample*, is chosen and the characteristic of interest is determined for the items in the sample.

Descriptive statistics is that area of statistics that summarizes the information contained in a sample. This summarization may be achieved by condensing the information and presenting it in tabular form. For example, frequency distributions are one way to summarize data in a table. Graphical methods of summarizing the data also may be used. The types of graphs discussed in introductory statistics courses often include histograms, ogives, and stem-and-leaf plots.

Data also may be summarized by numerical values. For example, to describe the center of a data set, the mean or median is often suggested. To describe variability, the variance, standard deviation, or interquartile range might be used. Each of the numerical values is a single number computed from the data and describes a certain characteristic of a sample.

Describing the information contained in a sample is only a first step for most statistical studies. If the study of a population's characteristics is the researcher's goal, then he or she would like to be able to use the information obtained from the sample to make statements about the population. The process of generalizing from characteristics of a sample to those of a population is called *statistical inference.* The bridge leading from descriptive measures computed for a sample to inferences made about population characteristics is the field of probability.

Statistical sampling is an additional topic discussed in introductory statistics. By choosing the elements of a sample in a particular manner, objective evaluations can be made of the quality of the inferences concerning population characteristics. Without proper choice of a sample, inferences can be made, but there is no way to evaluate these generalizations objectively. Thus, the manner in which the sample is chosen is important.

The most common type of sampling procedure discussed in introductory statistics is simple random sampling. Suppose a sample of n items is desired. To qualify as a *simple random sample* (SRS), the items in the sample are selected so that each possible sample of size n is equally likely to be chosen. In other words, each possible sample has an equal probability of being the one actually chosen. This is one of the pieces of the bridge that probability builds between descriptive statistics and statistical inference. Another piece of the bridge is a description of the behavior of certain of the numerical summaries that are computed as descriptive statistics.

Any numerical summary computed from a sample is called a *statistic.* A researcher will compute a single statistic from one sample chosen from the population of interest and use the numerical value of this statistic to make a statement about the value of some population characteristic. For example, suppose a particular brand of tires is to be studied to determine the average life of these tires. If the average life were known, the tire company might use this information to establish guarantees for its tires. An SRS of tires is chosen and tested to determine the individual lifetimes of the tires. Then the *sample average lifetime* is computed. This sample average can be used as an estimate of the population average lifetime of these tires.

The statistic computed, however, is the sample average lifetime for one particular sample of tires chosen. If a different set of n tires had been chosen, a different sample average might have resulted because of individual variation in the tires' lifetimes. Thus, the sample means themselves vary depending on which set of n tires is chosen as the sample. If this variation in the sample means was without any pattern, then there would be no way to relate the value of the sample mean obtained to the unknown value of the population mean. Fortunately, the behavior of the sample means (and other statistics) from random samples is not without a pattern. This behavior is described by a concept called a *sampling distribution.* Probability again enters the picture because sampling distributions are simply probability distributions. Through knowledge of the sampling distribution of a statistic, procedures can be developed to objectively evaluate the quality of sample statistics used to approximate population characteristics.

In this chapter many of the concepts mentioned previously will be reviewed.

These include descriptive statistics, random variables and probability distributions, sampling distributions, and statistical inference. Because most or all of these topics are covered in an introductory course in statistics, the coverage here will be brief.

For detailed references on introductory statistics, the interested reader is referred to texts such as:

Groeneveld, R. A. 1988. *Introductory statistical methods: An integrated approach using MINITAB.* Boston: PWS-Kent Publishing Co.

Hildebrand, D., and L. Ott. 1987. *Statistical thinking for managers,* 2d ed. Boston: Duxbury Press.

Mendenhall, W. 1988. *A course in business statistics,* 2d ed. Boston: PWS-Kent Publishing Co.

Mendenhall, W., J. E. Reinmuth, R. Beaver, and D. Duhan 1986. *Statistics for management and economics* 5th ed. Boston: Duxbury Press.

Ott, L. 1984. *An introduction to statistical methods and data analysis.* 2d ed. Boston: Duxbury Press.

2.2 Descriptive Statistics

Table 2.1 shows the earnings per share (EPS) for 1985 for a random sample of common stock for 20 firms rated A by the Standard and Poor's Corporation. Examining the 20 members in this list provides little information of use. Just looking at a list of numbers is too confusing even when the sample size is only 20. For larger samples, the confusion would be even greater.

The field of descriptive statistics provides ways to summarize the information in a data set. Summaries can be tabular, graphical, or numerical. One common tabular method of summarizing data is the frequency distribution. Intervals covering the range of the data are constructed and the number of observations in each interval is then tabulated and recorded. Table 2.2 shows one possible frequency distribution for the EPS data.

A graphical representation of a frequency distribution is called a *histogram.* Figure 2.1 shows the histogram representing the frequency distribution in Table 2.2. From the frequency distribution or the histogram, one can obtain a quick picture of certain characteristics of the data. For example, the center of the data and how much variability is present can be observed. The data have been summarized so that these characteristics are more obvious.

Most statistical software packages provide various tabular and graphical methods of summarizing data. Figure 2.2 shows the histogram constructed by MINITAB for the EPS data.

Numerical summaries are single numbers computed from a sample to describe some characteristic of the data set. Some common numerical summaries are sample mean, sample median, sample variance, and sample standard deviation. The

TABLE 2.1 EPS Data for A-Rated Firms, 1985

Firm Name	*EPS*
Allegheny Power Systems	3.41
American Security	1.63
AVEMCO Corporation	1.25
Boston Edison	2.47
Central Vermont Public Service	3.45
Commerce Bancshares	2.83
Dart & Kraft	2.33
Economics Laboratory	0.33
First of America Bank	3.77
Gerber Products	1.16
Hawaiian Electric Industries	2.33
Inter-City Gas	1.56
Key Corporation	1.92
Manufacturers Hanover	6.08
Minnesota Mining and Manufacturing	4.49
Norstar Bancorp	1.87
Pan Canadian Petroleum	1.79
Premier Industrial	0.34
Rochester Gas & Electric	3.81
Signet Banking	2.21

sample mean and sample median are measures of the center or central tendency of the data. The *sample mean* is just the average of all observations in the data set:

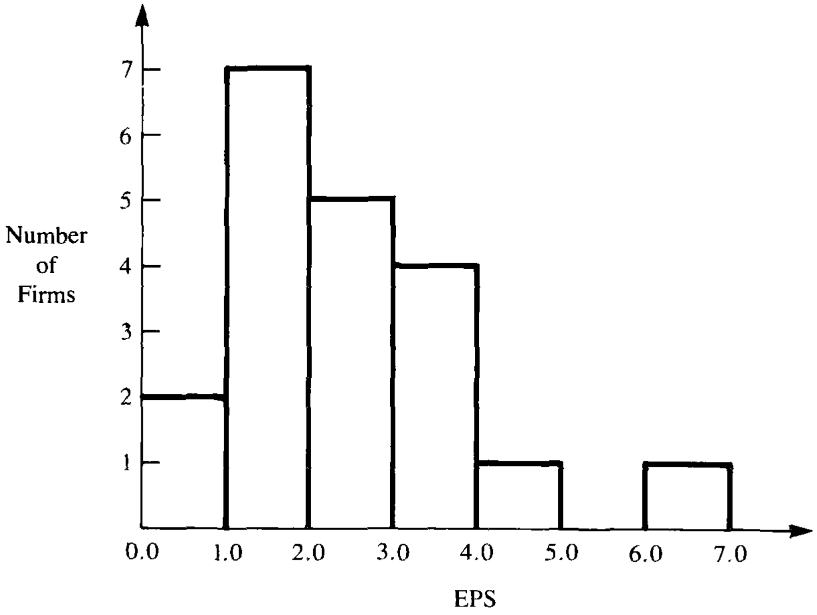
$$\bar{x} = \sum_{i=1}^{n} x_i/n.$$

(See Appendix A for an explanation of summation notation $\Sigma_{i=1}^{n}$). The sample *median* is the midpoint of the data after the data have been ordered. If n, the

TABLE 2.2 EPS Frequency Distribution

EPS	*Number*
At least 0.0 but less than 1.0	2
At least 1.0 but less than 2.0	7
At least 2.0 but less than 3.0	5
At least 3.0 but less than 4.0	4
At least 4.0 but less than 5.0	1
At least 5.0 but less than 6.0	0
At least 6.0 but less than 7.0	1

FIGURE 2.1 Histogram of the EPS Data



number of observations, is even, then the median is chosen to be the average of the two middle observations after the observations have been ordered from smallest to largest. If n is odd, the unique middle value in the ordered data set can be found and used as the median.

The sample variance and sample standard deviation are measures of the data's variability. The *sample variance* is computed as

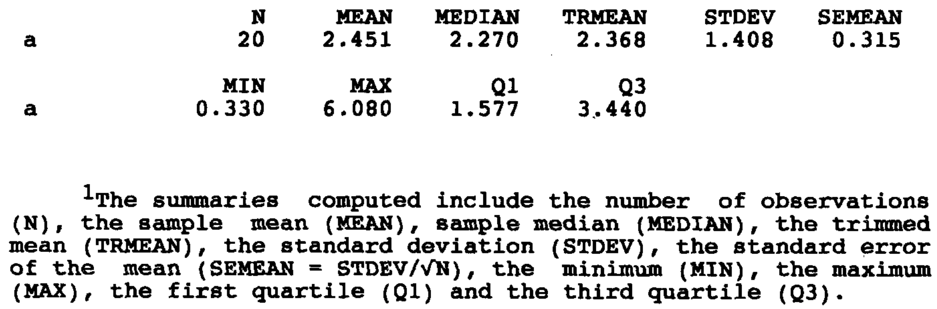
$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n - 1).$$

FIGURE 2.2 MINITAB Histogram for the EPS Data

```
Histogram of a   N = 20

 Midpoint   Count
      0.5       2  **
      1.0       1  *
      1.5       3  ***
      2.0       4  ****
      2.5       3  ***
      3.0       1  *
      3.5       2  **
      4.0       2  **
      4.5       1  *
      5.0       0
      5.5       0
      6.0       1  *
```

FIGURE 2.3 MINITAB Numerical Summaries for EPS Data[1]



```
                 N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
a               20    2.451    2.270    2.368    1.408    0.315

               MIN      MAX       Q1       Q3
a            0.330    6.080    1.577    3.440
```

```
     1The summaries  computed include the number  of observations
(N), the sample  mean (MEAN), sample median (MEDIAN), the trimmed
mean (TRMEAN), the standard deviation (STDEV), the standard error
of the  mean (SEMEAN = STDEV/√N), the  minimum (MIN), the maximum
(MAX), the first quartile (Q1) and the third quartile (Q3).
```

This is the average squared distance of each data point, x_i, from the center of the data, $\bar{x}$. The divisor $n-1$ is used rather than n to provide an unbiased estimator of the population variance. Because s^2 expresses variability in squared units, an intuitively more appealing measure is the *sample standard deviation*, s, which is simply the square root of s^2. Although many other numerical summaries exist, they will not be discussed in this review.

Figure 2.3 shows the results of using MINITAB to compute several descriptive measures for the EPS data, including the sample mean, sample median, sample variance, and sample standard deviation. The MINITAB command used to produce these descriptive measures is

```
DESCRIBE C1
```

where C1 is the column containing the data. In SAS,

```
PROC MEANS;
    VAR A;
```

will produce similar descriptive statistics. Here the EPS data were named A when read in by SAS. Also,

```
PROC CHART;
    VBAR A;
```

will produce a histogram in SAS. For a more extensive list of descriptive statistics in SAS, use

```
PROC UNIVARIATE;
    VAR A;
```