McGraw-Hill Series in
Probability and
Statistics

STATISTICAL METHODS IN THE BIOLOGICAL AND HEALTH SCIENCES

J. Susan Milton

THIRD EDITION

STATISTICAL METHODS IN THE BIOLOGICAL AND HEALTH SCIENCES

Third Edition

J. Susan Milton
Radford University

WCB McGraw-Hill

Boston Burr Ridge, IL Dubuque, IA Madison, WI New York San Francisco
St. Louis Bangkok Bogotá Caracas Lisbon London Madrid
Mexico City Milan New Delhi Seoul Singapore Sydney Taipei Toronto

WCB/McGraw-Hill

A Division of the McGraw-Hill Companies

STATISTICAL METHODS IN THE BIOLOGICAL AND HEALTH SCIENCE,
THIRD EDITION

This book is printed on acid-free paper.

6 7 8 9 0 QPF/QPF 0 7 6 5

ISBN 0-07-290148-9

Vice president and editorial director: *Kevin T. Kane*
Publisher: *John-Paul Lenney*
Sponsoring editor: *William K. Barter*
Marketing Manager: *Mary K. Kittell*
Project Manager: *Cathy Ford Smith*
Production Supervisor: *Laura Fuller*
Freelance design coordinator: *Mary I. Christianson*
Supplement coordinator: *Rita Hingtgen*
Compositor: *Interactive Composition Corporation*
Typeface: *10.5/12 Times Roman*
Printer: *Quebecor Printing Book Group/Fairfield, P.4*

p. 131: excerpt from Mark Stevens, *Three Mile Island*, © 1980. Reprinted with permission from Random House, Inc.

Library of Congress Cataloging-in-Publication Data

Milton J. Susan (Janet Susan)
Statistical methods in the biological and health sciences / J.
Susan Mılton — 3rd ed
p cm.
Includes bıblıographıcal references and ındex
ISBN 0-07-290148-9
J Lıfe scıences—Statıstıcal methods J Title
QH323 5 M54 1999
519 5-dc21 **98 18214**
CIP

www mhhe com

ABOUT THE AUTHOR

J. Susan Milton is professor of statistics at Radford University. Dr. Milton received a B.S. from Western Carolina University, an M.A. from the University of North Carolina at Chapel Hill, and a Ph.D. in statistics from Virginia Polytechnic Institute and State University. She is a Danforth Associate and is a recipient of the Radford University Foundation Award for Excellence in Teaching. Dr. Milton is the author of *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences,* as well as *Introduction to Statistics, Probability with the Essential Analysis, Applied Statistics with Probability,* and *A First Course in the Theory of Linear Statistical Models.*

PREFACE

It has become increasingly evident that the interpretation of much of the research in the biological and health sciences depends to a large extent on statistical methods. For this reason, it is essential that students in these fields be exposed to statistical reasoning early in their careers. This text is intended for a first course in statistical methods for undergraduate students in the biological and health sciences. However, it can also be used to advantage by graduate students with little or no prior experience with statistical methods.

This text is not a statistical cookbook, nor is it a manual for researchers. We attempt to find a middle road—to give the student an understanding of the logic behind statistical techniques as well as practice in using them. Knowledge of calculus is not assumed,. and readers with an adequate background in high school algebra should be able to follow the arguments presented.

We chose the examples and exercises specifically for the student of biological and health sciences. These are drawn from genetics, general biology, ecology, and medicine, and, except where indicated otherwise, data are simulated. However, the simulation is done with care so that the results of the analysis are consistent with recently reported research. In this way, the student will gain some insight into the types of problems that interest current workers in the biological sciences. Many exercises are open-ended to stimulate some classroom discussion.

It is assumed that the student has access to some type of electronic calculator. Many such calculators are on the market, and most have some built-in statistical capability. Use of these calculators is encouraged, for it allows the student to concentrate on the interpretation of the analysis rather than on the arithmetic computations. Instruction in the use of the TI83 graphing calculator is given in the text. This calculator, which is relatively new on the market, will perform most of the statistical tests presented in the text. It will also derive many of the confidence intervals described and has available most of the statistical tables discussed in the manuscript.

We should point out that most of the data sets presented are rather small so that the student will not be overwhelmed by the computational aspects of statistical analysis. This does not imply that very small samples are acceptable in biological research. In fact, most major research projects involve a tremendous investment in time and money and result in a large body of data. Such data lend themselves to analysis via the electronic computer. For this reason, we include some instruction in the interpretation of computer output. The package chosen for illustrative purposes is SAS (Statistical Analysis System: SAS Institute, Inc., Raleigh, North Carolina). This was done because of its widespread availability and ease of use. We do not intend to imply that it is superior to other well-known packages such as SPSSX (Statistical Package for the Social Sciences), BMD (Biomedical Computer Programs, University of California Press), or

MINITAB (Duxbury Press). An introduction to SAS together with the computer code required to generate the output is given in an optional Technology Tools section at chapter end.

This is a substantial revision of the second edition of the text. Reviewers' comments have been incorporated into the text to strengthen the discussions in many places. New exercises have been added throughout the text. A Technology Tools section introducing SAS programming and the TI83 graphing calculator has been added at the end of many chapters. New discussions include those of back-to-back stem-and-leaf plots, a simplified discussion of variance comparisons, and an expanded *T* table. The text continues to place a heavy emphasis on the finding and interpretation of *P* values.

A number of different courses can be taught from this book. They can vary in length from one semester to one year. It is difficult to determine exactly what material can be covered in a given time, since this is a function of class size, academic maturity of the students, and the inclination of the instructor. However, we do offer some guidelines for the use of this text in the chapter summaries below.

Chapter 1 This is an introduction to descriptive statistics. The notion of population versus sample is introduced early and stressed. The exploratory data analysis (EDA) topics of stem-and-leaf diagrams and box plots have been expanded. The importance of assessing shape, location, and variability is emphasized.

Chapter 2 This chapter introduces probability from an intuitive point of view. Tree diagrams are introduced and their use in solving genetics problems is emphasized. Counting techniques are given and tied to the problem of calculating probabilities via the classical method. If time does not permit coverage of the entire chapter, we suggest that Sections 2.1 and 2.2 be covered.

Chapter 3 This chapter covers the axioms of probability and the theorems that follow from the axioms. The topics of independence, conditional probability, and Bayes' Theorem are found here. A section entitled "Diagnostic Tests and Relative Risk" is included, presenting applications of conditional probability that are of special interest to students in the medical fields. This chapter can be skipped if time does not permit its coverage.

Chapter 4 This chapter covers discrete random variables only, introducing the notions of density, cumulative distribution, and expectation.

Chapter 5 This chapter parallels the ideas presented in Chapter 4 but applies them to continuous random variables. A subsection on the normal probability rule and its application to the construction of medical charts is given.

Chapter 6 In Chapter 6 we discuss point and interval estimation of the mean as well as hypothesis testing on the value of this parameter. A section on random sampling and randomization is included. The use of the *P* value is explained and emphasized throughout this chapter and the remainder of the text. A section on the effect of sample size on the length of confidence intervals and on the power of a test is given.

Chapter 7 This chapter is a short chapter on inferences concerning the variance and the standard deviation of a random variable. The discussion of variance comparisons has been simplified to include a rule-of-thumb check for equality. The formal *F* test is still included in the text.

Chapter 8 In Chapter 8 we discuss inferences on a proportion and the comparison of two proportions, with the Central Limit Theorem used to justify the techniques given.

Chapter 9 In Chapter 9 we compare two means via point and interval estimation and hypothesis tests. Preliminary F tests for comparing variances are discussed. Both the pooled and the Smith-Satterthwaite procedures for comparing means based on independent samples are explained. Discussion of how to use SAS for these tests is included. The chapter ends with a section about paired data.

Chapter 10 Chapter 10 introduces techniques used to compare the means of more than two populations, including discussions of the one-way classification model, randomized blocks, and the two-way classification model. The material includes a discussion of the effectiveness of blocking and Bonferroni T tests for conducting paired comparisons. Notes on computing are given throughout the chapter.

Chapter 11 This chapter presents a thorough discussion of simple linear regression and correlation. A section on multiple regression has been added.

Chapter 12 Categorical data problems are considered here, with an emphasis on tests of independence and tests of homogeneity in 2×2 and $r \times c$ tables.

Chapter 13 In this chapter, distribution-free alternatives to the classical procedures given in earlier chapters are presented. The material includes sections on the Lilliefors test for normality, Bartlett's tests for equal variances, and a small sample binomial test on proportions.

Many courses on this level are one semester in length, and it would be difficult to cover the entire text in that time. Sections that can be omitted with little loss of continuity are labeled as optional.

Thanks are due to Maggie Rogers, Bill Barter, and Cathy Smith for their encouragement and direction during the revision of this text and to Joann Fisher for the typing of the manuscript. My appreciation goes to Tonya Porter for her help in the preparation of the solutions manual. I also wish to recognize Joan Savage and Charlene Lutes for their help as biological consultants. Special thanks are offered to the following reviewers for their helpful suggestions: Charles M. Biles, Ph.D., Humboldt State University; John E. Boyer, Jr., Kansas State University; Annette Bucher, Colorado State University; Christiana Drake, University of California; Dr. R. K. Elswick, Jr., Medical College of Virginia, Virginia Commonwealth University; Thomas J. Glover, Hobart and William Smith Colleges; Golde I. Holtzman, Virginia Tech (VPI); Mark Krailo, University of Southern California; Benny Lo, NW Polytechnic University; Christopher Morrell, Loyola College; Lisa Sullivan, Boston University; Andrew Jay Tierman, Saginaw Valley State University; Mark S. West, Auburn University; and Robert F. Woolson, Ph.D., The University of Iowa.

J. Susan Milton

CONTENTS

1

Descriptive Methods

Statistics has become an indispensable tool for most scientists. What is statistics and how can statistical techniques be used to answer the practical questions posed by scientists?

Statistics has been defined as the art of decision making in the face of uncertainty. We begin by describing a typical problem that calls for a statistical solution. We use this example to introduce some of the language underlying the field of statistics. The terms are used here on an intuitive level. They are defined in a more technical sense later, as the need arises.

A researcher studying heart disease in persons 18 years old or older has identified four factors as being potentially associated with the development of the disease: age, weight, number of cigarettes smoked per day, and family history of heart disease. The researcher wants to gather evidence that either confirms these factors as contributing to the development of the disease or shows them to be unimportant. How should she or he proceed?

This is inherently a statistical problem. What characteristics identify it as such? Simply these:

1. Associated with the problem is a large group of objects (in this case, people) about which inferences are to be made. This group of objects is called the *population.*
2. Certain characteristics of the members of the population are of particular interest. The value of each of these characteristics may change from object to object within the population. They are called *random variables:* variables because they change in value; random because their behavior depends on chance and is somewhat unpredictable.
3. The population is too large to study in its entirety. So we must make inferences about the population based on what is observed by studying only a portion, or *sample,* of objects from the population.

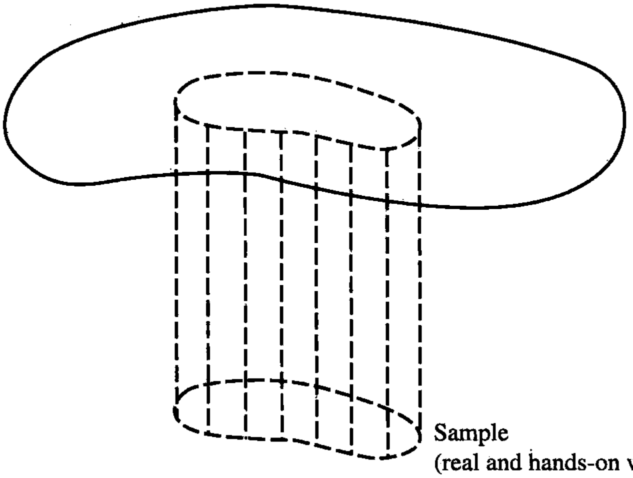
In the study of factors affecting heart disease, the population is the set of all persons suffering from the disease. The random variables of interest are the patient's age and weight, number of cigarettes smoked per day, and family history. It is impossible to identify and study every person with heart disease. Thus any conclusions that are reached must be based on studying only a portion, or a sample, of these people.

Random variables fall into two broad categories: continuous and discrete. A *continuous random variable* is a variable that prior to the experiment being conducted can assume any value in some interval or continuous span of real numbers. Measurements of things such as time, length, height, age, weight, speed, temperature, and pressure are usually assumed to be continuous. The variable age in the study of heart disease is continuous, as is the variable weight. The age of a person in the study conceivably can lie anywhere between 18 and, say, 110 years, a continuous time span. The person's weight may lie anywhere from perhaps 90 to 600 pounds! A *discrete random variable* is a variable that assumes its values at isolated points. Thus the set of possible values is either finite or countably infinite. Discrete random variables often arise in practice in connection with count variables. The number of cigarettes smoked per day is discrete. If we count a portion of a cigarette smoked as being a cigarette smoked, then the set of possible values is {0, 1, 2, 3, 4, 5, . . .}, a countably infinite collection. If family history is studied by recording the number of natural parents and grandparents who experienced heart disease, then this variable also is discrete. Its set of possible values is {0, 1, 2, 3, 4, 5, 6}, a finite collection. Random variables are usually denoted by capital letters.

A descriptive measure associated with a random variable when the variable is considered over the entire population is called a *parameter.* Parameters are usually denoted by Greek letters. To remember that parameters describe populations, just remember that both of these words begin with the letter *p*. One commonly encountered parameter is the population average value, or the population mean. This parameter is denoted by the Greek letter μ. For example, in the study of heart disease, the researcher would be interested in determining the average number of cigarettes smoked per day by members of the population. The exact value of this parameter cannot be obtained unless every member of the population is surveyed. Since this cannot be done, the exact value of μ will remain unknown even after our study is complete. However, we will be able to use statistical methods to approximate its value based on data obtained from a sample of patients drawn from the population.

A descriptive measure associated with a random variable when the variable is considered only over a sample is called a *statistic.* This is easy to remember because the words *statistic* and *sample* both begin with the letter *s*. Statistics serve two purposes. They describe the sample at hand, and they serve as approximations for corresponding population parameters. For example, the average number of cigarettes smoked per day by members of a sample of heart disease patients is a statistic. It is called a sample average or sample mean. Its value for a given sample probably will not equal the population mean μ exactly. However, it is hoped that it is at least close in value to μ.

A statistician or user of statistics is always working in two worlds. The ideal world is at the population level and is theoretical in nature. It is the world that we would like to see. The world of reality is the sample world. This is the level at which we really operate. We hope that the characteristics of our sample reflect well the characteristics of the population. That is, we treat our sample as a microcosm that mirrors the population. The idea is illustrated in Figure 1.1.



Population
(ideal but theoretical
world whose
characteristics are
described by parameters)

Sample
(real and hands-on world
whose characteristics are
described by statistics)

FIGURE 1.1
The sample is viewed as a miniature population. We hope that the behavior of the random variable under study over the sample gives an accurate picture of its behavior in the population.

We shall be interested primarily in three questions concerning the behavior of a random variable. These are:

1. What is the location of the variable? That is, about what value does the variable fluctuate?
2. How much variation is involved? That is, do the observed values of the variable tend to cluster closely together or are they widely spread?
3. What is the shape of the distribution? That is, do the values tend to fall into a bell-shaped, flat, u-shaped, or some other distinctive pattern?

In this chapter we introduce some graphical and some analytical techniques that can be used to answer these questions.

1.1
DISTRIBUTION TABLES: DISCRETE DATA

Recall that a discrete random variable is a random variable that can assume at most either a finite or a countably infinite number of possible values. Discrete random variables arise frequently in survey data. For example, we might want to compare the opinion of females concerning the issue of abortion to that of males. Hence one variable in the study is "sex." This variable is discrete since it assumes only the two naturally occurring values "male" or "female." We might ask the question, Do you favor legalized abortion on demand during the first trimester of pregnancy? Since the answer to this question varies from respondent to respondent, we are dealing with a random variable. The researcher could decide to record each response as "yes," "no," "undecided," or "refuses to answer." In this way a discrete random variable with four possible values is created. To understand and summarize such data, it is helpful to display the data in table or graphical form. These tables or graphs usually display the possible values of the random variable along with information on the number of times each value occurs. These counts are called *frequency counts* or simply *frequencies*. Example 1.1.1 illustrates the idea.

EXAMPLE 1.1.1. A comparative study of two adult homes in western Virginia is conducted. The purpose of the study is to determine the type of patients being served and to ascertain where patients go upon discharge from the home. Four discrete random variables are involved. They are sex (coded by the researcher as F = female or M = male), diagnosis (coded as MR = mentally retarded, MI = mentally ill, PI = physically ill), age, and destination after leaving the home (coded as 1 = died, 2 = home of relative, 3 = hospital, 4 = street, 5 = another home for adults, 6 = nursing home, 7 = not discharged at present). (Data presented are for one home and are taken from a larger study conducted by the statistical laboratory and Debbie Thompson, Department of Social Work, Radford University, 1990.)

Sex	Diagnosis	Age	Destination	Sex	Diagnosis	Age	Destination
M	MI	29	2	F	MI	72	6
M	MR	35	7	M	MI	52	7
F	PI	34	7	F	PI	31	7
M	MI	36	7	M	PI	35	7
F	MR	25	7	M	PI	42	7
F	MI	20	7	F	MI	29	2
F	PI	31	7	F	MR	61	7
F	PI	89	1	F	MI	18	3
M	MR	42	7	F	MR	64	7
M	MI	41	7	M	PI	51	7
F	PI	47	7	F	PI	30	7
M	PI	41	2	F	MR	35	7
M	MI	87	7	M	PI	40	6
F	MR	56	1	M	MR	76	3
F	MR	50	7	M	PI	59	7
F	PI	28	7	F	MI	71	6
M	MR	35	7	F	MI	62	7
F	PI	23	7	F	MI	65	3
F	MR	39	3	M	MR	51	7
M	PI	42	7	F	MR	18	7

The frequencies for the variable *diagnosis* are shown in Table 1.1. Notice that this table lists the category into which the response falls along with the number of observations per category.

In most studies frequency counts are obtained, and they do give valuable insight into the behavior of the random variable under study. However, frequency counts alone can be misleading. For example, suppose that we hear that 10 new cases of acquired immunodeficiency syndrome (AIDS) were diagnosed at a particular hospital during

TABLE 1.1
Frequency distribution for the variable *diagnosis* of Example 1.1.1

Category	Frequency
MI (mentally ill)	12
MR (mentally retarded)	13
PI (physically ill)	15

the month of June. Is this cause for alarm? Maybe—maybe not. It depends, of course, on the number of persons screened for the disease. Ten cases discovered among 20 persons tested certainly paints an entirely different picture than does 10 cases found among 1000 persons tested. To put a frequency count into perspective we report the count relative to the total, thus forming a *relative frequency*. Table 1.2 gives the frequencies and relative frequencies for the variable *diagnosis* of Example 1.1.1. Relative frequencies can be multiplied by 100 to yield the percentage of observations falling into each category. This information is useful since percentages are readily understood by everyone. Table 1.3 gives the complete summary of the variable *diagnosis*.

Table 1.4 is a summary table of the data as produced by SAS. SAS, which stands for Statistical Analysis System, is a statistical computing package that is in widespread use among data analysts, statisticians, and researchers. Some key SAS basics will be explained in the Technology Tools sections of this text. Notice that SAS has automatically listed the diagnosis values in alphabetical order. It has also included a column called "cumulative frequency" and one called "cumulative percent." The word *cumulative* means to accumulate or sum the values. Thus the cumulative frequency 25 is obtained by adding the number of mentally retarded patients (13) found in the second row to the number of mentally ill patients (12) found in the first row; the

TABLE 1.2
Frequency and relative frequency distributions for the variable *diagnosis* of Example 1.1.1

Category	Frequency	Relative frequency
MI (mentally ill)	12	12/40 = .300

TABLE 1.3
Complete summary table for the variable *diagnosis* of Example 1.1.1

Category	Frequency	Relative frequency	Percentage
MI (mentally ill)	12	12/40 = .300	30.0
MR (mentally retarded)	13	13/40 = .325	32.5
PI (physically ill)	15	15/40 = .375	37.5