INTRODUCTION TO STATISTICS

J.S.Milton
P.M.McTeer
J.J.Corbet

INTRODUCTION TO STATISTICS

J. Susan Milton
Radford University

Paul M. McTeer
Radford University

Boston, Massachusetts Burr Ridge, Illinois Dubuque, Iowa
Madison, Wisconsin New York, New York San Francisco, California St. Louis, Missouri

WCB/McGraw-Hill

A Division of The McGraw-Hill Companies

Introduction to Statistics

This book is printed on acid-free paper.

2 3 4 5 6 7 8 9 0 FGR FGR 9 0 3 2 1 0 9 8 7

ISBN 0-07-042528-0

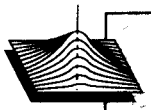
This book was set in Times Roman by York Graphic Services, Inc.
The editors were Jack Shira, Maggie Lanzillo Rogers, and Caroline Jumper.
Editorial assistance was provided by Linda McPhee Smith.
The designer was Robin Hessel Hoffmann.
The production supervisors were Natalie Durbin and Tanya Nigh.
The cover was designed by Armen Kojoyian.
Illustrations were by Accurate Art, Inc., and York Graphic Services, Inc.
Quebecor Printing Fairfield, Inc., was printer and binder.

Library of Congress Cataloging-in-Publication data

Milton, J. Susan (Janet Susan)
Introduction to statistics / J. Susan Milton, Paul M. McTeer, James J. Corbet.
p. cm.
Includes bibliographical references (p. A95-A96) and index.
ISBN 0-07-042528-0 (alk. paper)
1. Statistics. I. McTeer, P. M. II. Corbet, J. J. (James J.) III. Title.
QA276. 12.M55 1996 96-22270
519.5 — dc20 CIP

INTERNATIONAL EDITION

When ordering this title, use ISBN 0-07-114523-0



PREFACE

Over the past several years there has been a renewed interest in the understanding and application of statistics. Students studying the material for the first time, however, find it difficult to grasp the concepts involved because most problems are word problems that require a certain amount of reasoning ability, not simply rote memorization. Each problem looks totally different; it is only with practice and experience that students begin to recognize patterns and problem types. To develop this recognition, students need to have available a large number of interesting examples and exercises that present the same concept in different settings. This need has been met in this book. Our goal in writing this book is to provide students with a knowledge of the mechanics of problem solving and also with an idea of when and how to apply statistical methods. The student who completes a course using this book will have an adequate background for doing basic research or reading research papers in his or her field.

This *Introduction to Statistics* is intended for students studying the subject at the precalculus level. Chapter 1, Descriptive Methods, introduces students to concrete examples of data handling. In that chapter, we present many standard techniques for displaying and summarizing data sets. We also introduce some of the graphical techniques of EDA (exploratory data analysis). The chapter on probability, Chapter 2, is quite complete. It can be covered in its entirety in about 2 weeks, or in less time if Section 2.1 is used as a reading assignment and Sections 2.7 through 2.9 are omitted. Chapter 3 introduces discrete random variables and places particular emphasis on the Binomial and Poisson distributions. The concept of expectation is explained and is tied to the notion of fair wagers and fair odds. Chapter 4 parallels Chapter 3 and introduces the general concepts underlying continuous random variables. The normal curve is studied in detail. Chapter 5 begins the study of statistical inference. In this chapter the student begins to apply probability to make inferences about a population based on a sample drawn from the population. The T and X^2 distributions are introduced, the notion of a confidence interval is explained, and the language of hypothesis and significance testing is developed. Chapter 6 develops inferential techniques for proportions. It also introduces the first two-sample procedure, that for comparing two proportions. Chapter 7 applies the concepts of Chapters 5 and 6 to the study of control charts. Some of the more commonly employed control charts in statistical quality control are developed. Chapter 8 presents techniques used to compare two means and two variances. This chapter is a natural extension of Chapter 5. Chapter 9 gives a thorough development of simple linear regression and provides an introduction to multiple regression. Regression and correlation techniques are compared and contrasted. Chapter 10 introduces techniques used to analyze categorical data via X^2 tests of

association. The ideas of Chapter 8 are extended to k samples in Chapter 11. The one-way classification model as well as randomized complete blocks are discussed. Scheffé, Duncan, and Bonferroni multiple-comparison tests are presented. The book ends with a chapter on distribution-free tests that can be used as alternatives to the normal theory procedures that were developed in the rest of the book. Throughout the book, the data sets presented are rather small so that students will not be overwhelmed by the computational aspects of statistical analyses. However, we do not intend to imply that statisticians routinely rely on very small samples.

SPECIAL FEATURES

Technology Tools Students are encouraged to use their hand-held calculators or computers since our purpose is to emphasize the interpretation of statistical results rather than the mere computation of these results. Instruction in the use of three of the leading graphics calculators—the TI82, TI83, and TI85—and for use of the statistical software package Minitab is included in sections clearly labeled *Technology Tools*. (For more information about Minitab, please contact Minitab, Inc., 3081 Enterprise Drive, State College, Pennsylvania, 16801-3008.)

Experimental Exercises The book also includes a series of Experimental Exercises. These are "hands on" exercises that can be completed in class. They are classroom-tested and their purpose is to allow the students to do some data collection on their own. We have found that the use of some or all of these exercises will help bring statistics to life and will enhance the enjoyment of the course for students.

Applications Problems and examples in the text are drawn from a variety of disciplines, including business, sports, psychology, economics, and medicine. They range from the simple to the challenging. Some problems are icon-coded so that, if desired, assignments can be targeted toward particular interest groups. The following subjects are represented by these icons:

 Business/Economics

 Social Sciences

 Biology/Physical Science

 Health Science

Study Tools Each chapter ends with a comprehensive vocabulary list and a set of review exercises to help students prepare for exams. Answers to most odd-numbered and review exercises are contained in an appendix.

SUPPLEMENTS

Student Solutions Manual The Student Solutions Manual contains detailed solutions to the odd-numbered exercises in the text.

Instructor's Resource Manual The Instructor's Resource Manual contains detailed solutions to the even-numbered exercises in the book and a bank of test questions. It also contains special supplements for the statistical packages SAS, SYSTAT, and SPSS. These supplements cover the same material that is discussed in the book using Minitab.

Computerized Test Bank A computerized version of the testbank from the Instructor's Resource Manual is available in both IBM and Macintosh formats.

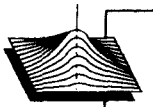
McGraw-Hill's Statistics Discovery Series: A Guide to Learning Statistics This supplement is intended to help students enhance their understanding of introductory statistics. Each section of this study guide contains study objectives, an overview of the topics covered, key terms and definitions, worked-out examples, helpful hints to the student, and new exercises and their solutions.

McGraw-Hill's Statistics Discovery Series: A Guide to Minitab This supplement helps the student gain a better understanding of statistics through the use of the statistics software Minitab. Worked-out examples and new exercises for use with Minitab are presented, along with a data disk containing data sets ready for use with Minitab. The supplement contains command information for DOS, Windows, and Macintosh platforms, and is packaged with either an IBM or a Macintosh disk.

McGraw-Hill's Statistics Discovery Series: A Guide to TI Graphing Calculators for Statistics This supplement contains instructions for the student using a graphing calculator in an introductory statistics course through worked-out examples and new exercises. Appendixes for TI81, TI82, and TI85 graphing calculators are also included.

ACKNOWLEDGMENTS

We wish to thank Jack Shira, Maggie Lanzillo Rogers, and Caroline Jumper for their encouragement and help in the production of this book, Carlotta Eaton for help with the Minitab material in the textbook and for preparing the testbank, Gary Ford for help with computer graphics in the solutions manual, Michele Riley for writing the SYSTAT supplement, and Stephanie Meadows for preparing the SPSS supplement in addition to testing all of the instructions for the TI calculators. A special thanks goes to Jo Ann Fisher for her expertise and patience in the preparation of the many versions of the manuscript. We also acknowledge the following reviewers for their helpful suggestions: Robert Heckard, Pennsylvania State University; Robert Lacher, South Dakota State University; Ronald Pierce, Eastern Kentucky University; C. Bradley Russell, Clemson University; Sam C. Saunders, Washington State University.



CONTENTS

Preface xi

Introduction: Statistics—The Tool of Decision Makers **1**

Experimental Exercise I: Gaining Experience in Designing a Study and Posing Research Questions 9

Chapter 1 Descriptive Statistics **11**

1.1 Some Detective Work: Introducing Exploratory Data Analysis 12

Experimental Exercise II: Conducting a Study and Displaying Data in a Stem-and-Leaf Diagram 21

1.2 Picturing the Distribution: Histograms 22

1.3 Measures of Location 33

Experimental Exercise III: Understanding the Shape and Location of a Distribution—And Beginning to Think About Association Between Two Variables 34

1.4 Measures of Variability 43

1.5 Box Plots 52

Experimental Exercise IV: Gaining Experience in Constructing Box Plots and Computing Basic Statistics 60

Experimental Exercise V: Using Box Plots to Compare Multiple Data Sets 60

Technology Tools **64**

Chapter 2 Probability and Counting **79**

2.1 Intuitive Probability 80

Experimental Exercise VI: Understanding the Difference Between Classical Probability and Relative Frequency Probability 86

2.2 Tree Diagrams and Elementary Genetics 87
Experimental Exercise VII: Simulating a Genetics Experiment 97
2.3 Sample Spaces and Events 97
2.4 Some Rules of Probability 106
2.5 Conditional Probability 114
2.6 Independence and the Multiplication Rule 119
2.7 Bayes' Rule 131
2.8 Counting Sample Points 136
2.9 Counting Arrangements of Objects: Permutations 142
2.10 Counting Selections of Objects: Combinations 147
Technology Tools **154**

Chapter 3 Random Variables: Discrete Distributions **155**

3.1 The Basic Definitions 156
3.2 Computing Probabilities: The Discrete Case 159
Experimental Exercise VIII: Investigating the Behavior of a Pair of "Fair" Dice Via the Game of Craps 164
Experimental Exercise IX: Choosing Between Two Reasonable Proposed Probability Functions for a Discrete Random Variable 165
3.3 Measures of Location and Variability 166
3.4 The Binomial Distribution 177
3.5 Computing Binomial Probabilities 185
3.6 The Poisson Distribution 190

Chapter 4 Continuous Distributions: The Normal Curve **199**

4.1 Computing Probabilities: The Continuous Case 200
4.2 Measures of Location and Variability 205
4.3 The Normal Distribution 208
4.4 Computing Normal Probabilities 214
4.5 Approximating the Binomial Distribution 224
Experimental Exercise X: Investigating the Behavior of One Discrete and One Continuous Random Variable 233

Chapter 5 Inferences on a Single Mean and a Single Variance 235

5.1 The General Statistical Problem: Sampling 236
5.2 Point Estimation of the Mean 242
5.3 The Central Limit Theorem and Interval Estimation of the Mean with the Variance Known 248
Experimental Exercise XI: Illustrating the Central Limit Theorem and the Concept of a Confidence Interval on the Mean 257
5.4 The t Distribution and Interval Estimation of the Mean with the Variance Unknown 258
Experimental Exercise XII: Conducting a Study to Estimate Lambda, the Parameter in a Poisson Process 269
5.5 Introduction to Hypothesis Testing 270
5.6 Testing a Hypothesis on the Mean 275
5.7 The Chi-Squared Distribution and Interval Estimation of the Variance 290
5.8 Testing a Hypothesis on the Variance 298
Technology Tools 306

Chapter 6 Inferences on Proportions 311

6.1 Estimating a Proportion 312
6.2 Testing a Hypothesis on a Proportion 322
Experimental Exercise XIII: Designing and Conducting a Hypothesis Testing Experiment 330
6.3 Estimating the Difference in Proportions 330
Experimental Exercise XIV: Conducting a Study to Compare Two Proportions p_1 and p_2 336
6.4 Comparing Two Proportions: Hypothesis Testing 337
Technology Tools 345

Chapter 7 Statistical Quality Control 349

7.1 Properties of Control Charts: $\overline{X}$ Charts 350
7.2 The Geometric Distribution and Control Charts 357
7.3 The Shewhart Control Chart for the Range: R Charts 364
7.4 Shewhart P Charts and C Charts 366

Experimental Exercise XV: Gaining Experience in Constructing and Using a Control Chart 372
Technology Tools **375**

Chapter 8 Comparison of Two Variances and Two Means **379**

8.1 Comparison of Two Variances and the F Distribution 380
8.2 Comparing Two Means: Pooled Estimation 391
8.3 Comparing Two Means: Pooled-T Tests 400
8.4 Comparing Two Means: Variances Unequal 405
Experimental Exercise XVI: Using the Two-Sample T Test and the Preliminary F Test 412
8.5 Comparison of Two Means: Paired-T Tests 413
Experimental Exercise XVII: Conducting a Study of Hand Dominance and Coordination via a Paired Design 421
Technology Tools **426**

Chapter 9 Regression and Correlation **435**

9.1 Introduction to Simple Linear Regression 436
9.2 The Method of Least Squares 443
Experimental Exercise XVIII: Designing and Carrying Out a Simple Regression Study 451
9.3 Interpreting Computer-Generated Output 451
9.4 The Simple Linear Regression Model 457
9.5 Confidence Interval Estimation of $\mu_{Y|x}$ and Prediction Intervals on $Y|x$ 465
9.6 Inferences on the Slope 472
9.7 Multiple Regression 478
9.8 Introduction to Correlation 488
Experimental Exercise XIX: Reviewing the Concept of Correlation 497
Experimental Exercise XX: Performing a "Hands On" Regression Experiment 498
Experimental Exercise XXI: Designing and Carrying Out a Multiple Regression Study 500
Technology Tools **505**

Chapter 10 Categorical Data **511**

10.1 2×2 Contingency Tables: Notation 512
10.2 Testing for Association 516
10.3 $r \times c$ Contingency Tables 525
Experimental Exercise XXII: Testing for Association Between Two Random Variables 530
Technology Tools **533**

Chapter 11 Analysis of Variance **535**

11.1 One-Way Classification 536
11.2 Paired and Multiple Comparisons 552
11.3 The Randomized Complete Block Design 572
Technology Tools **589**

Chapter 12 Some Distribution-Free Alternatives **593**

12.1 One-Sample Procedures 594
12.2 Two-Sample Procedures for Matched Data 602
12.3 A Two-Sample Procedure for Unmatched Data 609
12.4 Correlation Procedures 612
12.5 k-Sample Procedures 614

Appendix A Statistical Tables **A1**

Table I Cumulative Binomial Distribution A2
Table II Poisson Distribution A7
Table III Cumulative Standard Normal Distribution A8
Table IV A Table of Random Digits A10
Table V Blood Pressure Data A12
Table VI T Distribution A14
Table VII Sample Size for Estimating the Mean (T tests) A16
Table VIII Cumulative Chi-Squared Distribution A18

Table IX	Control Chart Constants	A20
Table X	*F* Distribution	A21
Table XI	Least Significant Studentized Ranges r_p	A42
Table XII	Wilcoxon Signed-Rank Test	A43
Table XIII	Wilcoxon Rank-Sum Test	A44

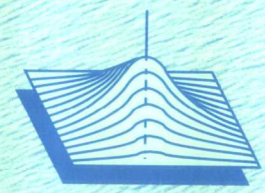
Appendix B Answers to Selected Exercises **A49**

References **A95**

Index **A97**

INTRODUCTION

Statistics—The Tool of Decision Makers



EXPERIMENTAL EXERCISE I: GAINING EXPERIENCE IN DESIGNING A STUDY AND POSING RESEARCH QUESTIONS

In this day and age we are deluged with figures. Many of these figures are commonly referred to as statistics. Figures are published on a huge range of topics: We read government reports on the state of the Social Security system, the arms race, and the cost of medical care. We are forced to deal daily with the effects of changes in the prime-interest rate, the cost of living, and the inflation and unemployment rates. We are influenced to change our driving habits when we hear news reports that "statistics indicate that a child not protected by a child's safety seat is five times more likely to be killed in an automobile crash than is one who uses such a restraint." To most of us, these and other "statistical" reports are mysterious and somewhat threatening. We are never quite sure how they arose, how accurate they are, or what they really mean.

The purpose of this book is to introduce the methods used in the field of applied statistics. It is convenient to break these methods into two broad categories called **descriptive methods** and **inferential methods**. Descriptive methods are techniques, both analytic and graphical, that are used to simply describe or paint a picture of a data set. These techniques are presented in Chapter 1 and are used throughout the text. Inferential methods are techniques that are used to draw conclusions or to make inferences about a large group of objects based on the observation of only a portion of the members of the group. These methods make use of the concepts of probability theory. They are discussed in Chapters 5 through 12, after a brief survey of the fundamentals of probability. The methods presented in this text are those that are used to prepare many of the statistical reports that you read about in the newspaper and hear about on news reports. By gaining some insight into the manner in which statistical reports are prepared, it is hoped that they will appear less mysterious: You will be able to better interpret the figures and assess the meaning of the findings that are placed before you.

When studying any subject for the first time, you must learn its basic vocabulary. Fortunately, there are only a few concepts that must be understood before we begin our detailed study of statistical methods. We introduce these concepts here on an intuitive level via some typical problems that can be approached statistically. The terms presented will be defined more precisely in later chapters as the need arises.

Statistical studies have one important characteristic in common: Interest always centers on a target group of objects. We want to describe the characteristics of this group or draw some conclusions concerning its behavior. Unfortunately, the group is usually too large to study in its entirety. For this reason, the conclusions that we reach must be based on the observation of only a portion or a subset of its members. The terms used to describe this situation are *population* and *sample*.

The **population** in a statistical study is the group of objects about which conclusions are to be drawn.

A **sample** is a portion or subset of objects drawn from the population.

The following examples should clarify the meaning of the two terms.

EXAMPLE 1 The use of alcohol among teenagers is of concern to all of us. A study is to be conducted to help answer questions concerning the drinking habits of young persons in the United States in the 12- to 19-year-old age group. Questions to be answered via the data gathered are as follows: What proportion of the individuals in this age group drink on a regular basis? and If an individual is a regular drinker, what is the average age at which drinking began? The population here consists of *all* persons in the United States who are at least 12 years old but who have not yet reached their twentieth birthday. This group does exist but it is so large that it is physically impossible to study it in its entirety. We must try to answer the questions posed by selecting and interviewing only a sample of those young people who constitute the population.

EXAMPLE 2 Industrial robots are being used more and more frequently in American industry. They are of particular value in the automotive industry. A study of the characteristics of a robot designed to spray-paint the hoods of automobiles is to be conducted. The primary question to be answered is, On the average, how many hoods can such a robot paint before it requires cleaning and other maintenance? In this example, the population is somewhat hypothetical. It includes *all* robots of the type described, both those that currently exist, and those that will be produced in the future. Since we cannot study objects that do not yet exist, we must draw conclusions about these robots based on a sample of those currently in use.

EXAMPLE 3 A city council is considering the possibility of beginning a limited bus service for its citizens. Before taking this step, the opinion of its 25,000 adult citizens is to be sought. In particular, council members want to know whether a majority of the adult citizens favors the proposal. They also want to determine the average number of blocks that an individual is willing to walk to catch the bus so that effective routes can be developed. The population here consists of the 25,000 individuals living in this city who are considered to be adults. Since it is impossible to contact each of these persons individually, we must sample to obtain an idea of public opinion.

In a statistical study, interest centers on some particular variables associated with the population. The values assumed by these variables can change from one member of the population to another. The change is due to random or chance influences and is therefore somewhat unpredictable. The term *random variable* is used to describe such a variable. That is,

A random variable is a variable whose numerical value is determined by the outcome of some chance experiment.

In Example 1, the use of alcohol among teenagers is being studied. To determine the proportion of young people in the population who drink on a regular basis, we might ask the question, Do you usually drink on at least three separate occasions per month? We can define a random variable X by agreeing to record a 1 for X if the

answer to this question is yes; otherwise, we will record a 0. The variable X is random because its value can change from person to person, and the particular value that it assumes for a given individual depends on chance. We also want to determine the average age at which regular drinkers began to drink. To do this, we need to record the value of the random variable Y, the age at which drinking began, for the appropriate individuals.

Example 2 entails trying to determine the characteristics of a particular type of industrial robot. To help answer the question, On the average, how many hoods can such a robot paint before it requires cleaning and other maintenance? we will study the random variable Z, the number of hoods painted per robot. The values assumed by Z will vary from robot to robot due to the differences in workmanship, differences in environmental conditions, and other chance factors.

We leave the identification of the random variables to be studied in Example 3 as an exercise.

Random variables studied in this text fall into two broad categories. They are either *continuous* or *discrete*.

A continuous random variable is a random variable that, prior to the experiment, can conceivably assume any value in some interval or continuous span of real numbers.

The random variable Y, the age at which a regular drinker began to drink, is continuous. It can conceivably lie anywhere between perhaps 10 and 20 years of age, excluding the value 20 itself.

A discrete random variable is a random variable that can assume at most a finite or a countably infinite number of possible values.

To say that a set of values is *finite* means that the number of values can be counted. To say that it is *countably infinite* means that we can begin to count the values but then realize that there is no end to them. The random variable X, which assumes only the values 0 or 1, depending on whether or not the individual sampled answers yes or no to the question, Do you usually drink on at least three separate occasions per month?, is discrete because the number of possible values is finite. The random variable Z, the number of hoods painted by a robot before maintenance is necessary, is discrete. Its set of possible values is 0, 1, 2, ... In this case the number of possible values is countably infinite. Whether finite or countably infinite, a discrete random variable assumes its values only at isolated points.

We will use uppercase letters such as X, Y, and Z to denote random variables; we will use lowercase letters x, y, and z to denote the observed values of these variables. For example, if a teenager reports that she usually drinks on at least three separate occasions per month and that she started drinking at age 14 years and 6 months, then we write $x = 1$ and $y = 14.5$.

As you will soon discover, statistical studies can be thought of as studies of the behavior of one or more random variables. Associated with these random variables are certain constants or numerical measures which are descriptive in nature. The average value of the variable over the entire population is such a measure; it describes the value about which the values of the random variable tend to cluster. The difference between the smallest and largest value assumed by the random variable over the entire population is another such measure; it gives us a rough idea of the degree of dispersion or spread exhibited by the random variable. Measures such as these are called *population parameters.*

A population parameter is a descriptive measure associated with a random variable when the variable is considered over the entire population.

Unfortunately, since we do not usually study the entire population, the actual numerical values of a specific population parameter are seldom known. We must attempt to approximate their true values based on information obtained from a sample. Before we can use information gained from a sample to answer questions concerning the population from which the sample is drawn, we must be able to describe our sample in a logical way. To do so, we make use of what are commonly called *statistics.*

A statistic is a descriptive measure associated with a random variable when the variable is considered only over a sample.

EXAMPLE 4 In Example 1, we want to determine the average age at which a regular drinker in the 12- to 19-year-old age group began to drink. This average age is a population parameter. It is the value that we would obtain if we could do the following: Locate and interview every person in the United States who is at least 12 years old but who has not yet reached his or her twentieth birthday, determine which of these individuals are regular drinkers, record the age at which each regular drinker began to drink, and then average these values. This is obviously an impossible task! Nevertheless, we can approximate the value of this population parameter in a logical way. We simply interview a sample of young people and record the age at which the regular drinkers in the sample began to drink. We then average these values. This sample average is a statistic. Its value probably is not exactly the same as that which would be obtained if we could interview everyone in the population. However, for sufficiently large samples it usually will be fairly accurate. Similarly, the proportion of drinkers in the 12- to 19-year-old age group is a population parameter; the proportion of members of the sample who drink is a statistic.

As you can see, statisticians and users of statistics must learn to think on two levels. The theoretical or population level is an ideal world. It is the world that we would like to study but usually cannot. Its characteristics are described by parameters. The world of reality is the sample. This is the level at which we operate. Its characteristics