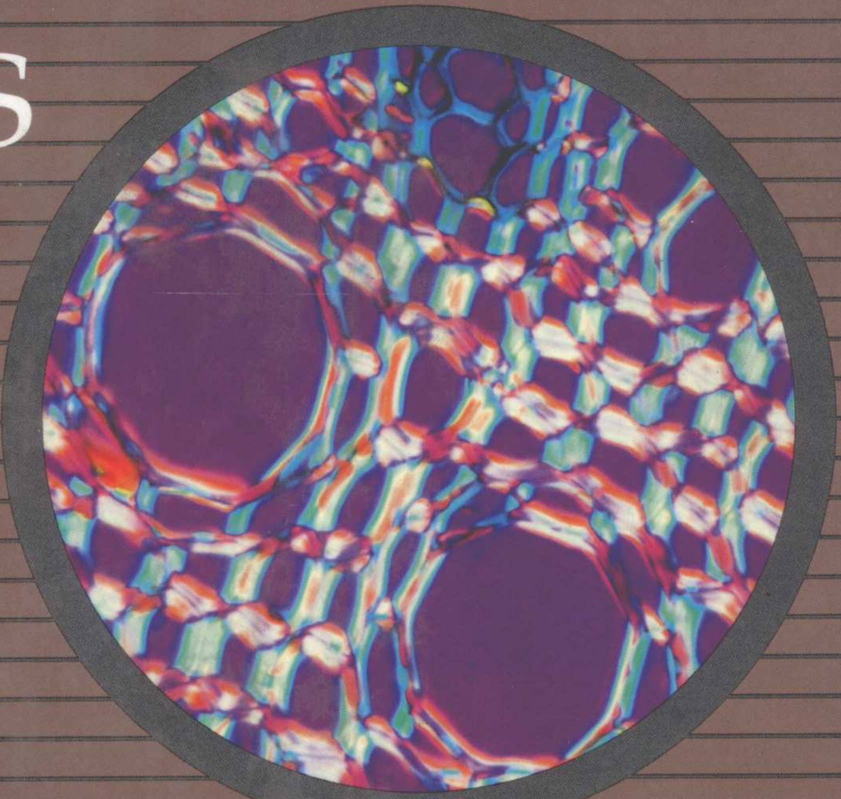
Statistical Research Methods in the Life Sciences



P. V. Rao

Statistical Research Methods in the Life Sciences

P. V. Rao

University of Florida

An Alexander Kugushev Book

Duxbury Press
An Imprint of Brooks/Cole Publishing Company
I(T)P® An International Thomson Publishing Company

Pacific Grove • Albany • Belmont • Bonn • Boston • Cincinnati • Detroit • Johannesburg • London
Madrid • Melbourne • Mexico City • New York • Paris • Singapore • Tokyo • Toronto • Washington

Sponsoring Editor: *Alex Kugushev*
Assistant Editor: *Cynthia Mazow*
Marketing Team: *Carolyn Crockett/Romy Taormina*
Editorial Assistants: *Martha O'Connor/Rita Jaramillo*
Production Editor: *Tessa A. McGlasson*
Manuscript Editor: *Bernard Gilbert*
Cover Design: *Harry Voigt*
Cover Image: *Spike Walker/Tony Stone Images*

Interior Design: *Cloyce J. Wall*
Interior Illustration: *Suffolk Technical Illustrators*
Art Editor: *Lisa Torri*
Indexer: *Ted Laux*
Typesetting: *SuperScript*
Cover Printing: *Phoenix Color Corp.*
Printing and Binding: *Quebecor/Fairfield*

For more information, contact Duxbury Press at Brooks/Cole Publishing Company:

BROOKS/COLE PUBLISHING COMPANY
511 Forest Lodge Road
Pacific Grove, CA 93950
USA

International Thomson Publishing Europe
Berkshire House 168-173
High Holborn
London WC1V 7AA
England

Thomas Nelson Australia
102 Dodds Street
South Melbourne, 3205
Victoria, Australia

Nelson Canada
1120 Birchmount Road
Scarborough, Ontario
Canada M1K 5G4

International Thomson Editores
Seneca 53
Col. Polanco
11560 México, D.F., México

International Thomson Publishing GmbH
Königswinterer Strasse 418
53227 Bonn
Germany

International Thomson Publishing Asia
221 Henderson Road
#05-10 Henderson Building
Singapore 0315

International Thomson Publishing Japan
Hirakawacho Kyowa Building, 3F
2-2-1 Hirakawacho
Chiyoda-ku, Tokyo 102
Japan

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Rao, P. V. (Pejaver Vishwamber), [date]-
Statistical research methods in the life sciences / P.V. Rao.
p. cm.
Includes bibliographical references (p.) and index.
ISBN 0–534–93141-3
1. Life sciences--Statistical methods. I. Title.
QH323.5.R34 1997
570'.72--dc21 97–15435
CIP
AC

THIS BOOK IS PRINTED ON ACID-FREE RECYCLED PAPER

To the memory of my parents, *Janaki and Shridhar Rao*
To the love and friendship of my wife, *Premila*

PREFACE

This book is primarily intended as a text for a one- or two-semester course on statistical research methods for graduate students in biology, agriculture, and related life sciences. The book contains a variety of real examples and exercises drawn from these areas, including many from the author's consulting experience.

Features

The book introduces statistical models early and uses a model based approach in the development of the statistical methods. The main focus is on planning and analyzing designed experiments. Point and interval estimation take priority over statistical hypothesis testing. Particular attention is paid to methods of constructing and interpreting one-at-a-time and simultaneous confidence and prediction intervals (one- and two-sided). Methods and guidelines for determining sample sizes get more emphasis in this text than is normally found in books written for similar audiences. The important difference between hypothesis testing and confidence interval approaches to sample size calculation is emphasized, and guidelines are given on how to use SAS to determine sample sizes.

Even though the book emphasizes interpretation of results over computational details, computational details are given when it is felt that such details help in the interpretation of the computed quantities. Printouts from popular statistical software—SAS and StatXact—are frequently used to display the results in the worked examples in the text.

Also available from the publisher is the *SAS Companion for Statistical Research Methods in the Life Sciences*, authored by Mary Sue Younger (1997), which shows how to use the statistical computing software SAS to perform the calculations described in this book.

Level

The book, with the exception of Appendix B, is written at a mathematical level typical among first-year graduate students in life sciences. The ability to manipulate simple mathematical formulas with symbols and interpret graphs of simple functions is expected, but knowledge of calculus or a background in statistics is not necessary. Some familiarity with upper division linear algebra and calculus will be needed if covering Appendix B.

Organization

The book can be divided into seven parts. The first part, consisting of the first three chapters, is devoted to a discussion of some basic concepts and definitions central to the study of statistics. Throughout this initial part, the focus is on the theme that statistics deals with methods of collecting and using information obtained in samples to draw conclusions about populations. The notion of population, sample, sampling distribution, estimation, hypothesis testing, and prediction are introduced in this part. In Part 2 (Chapters 4, 5, and 6) the most commonly occurring situations—in which the researcher is interested in making inferences about one or two populations—are used to describe statistical methods of estimation, hypothesis testing, and prediction. Part 3 (Chapters 7, 8, and 9) discusses some general issues pertaining to designing research studies and presents methods based on one-way analysis of variance (ANOVA) for the design and analysis of comparative experiments; that is, experiments in which the objective is to compare several treatments with each other. Part 4 contains Chapters 10 and 11 and is devoted to a discussion of regression methods. A detailed discussion of simple and multiple linear regression is included in these chapters. Chapter 12 is the fifth part, in which ANOVA models and regression models are treated as special cases of the general linear models. Analysis of covariance is treated as an example of the use of general linear models. Chapters 13, 14, and 15 constitute Part 6, in which the one-way ANOVA is extended to cover experiments involving multiple factors with fixed and random effects. Finally, Part 7 (Chapter 16) provides an introduction to analysis of repeated measures designs, an important application of the methods based on general linear models involving random and fixed effects. In this text, split-plot experiments are treated as special cases of repeated measures studies.

Suggested Use

Most of the material in the book has been class tested in graduate-level applied statistics courses taught by the author to graduate students in the life sciences and in statistics. The book contains sufficient material to cover a two-semester sequence of courses. Depending on the emphasis of the course, some material may be omitted or used as extra reading assignments. For example, for students with some statistics background, Chapters 1, 2, and 3, may be covered very quickly; class time might be spent on only the main definitions with the remaining material used as reading assignments. In some courses, one or more of the chapters and sections dealing with ordinal data, categorical data, and sample sizes may be skipped. One possible division of the text material into two semesters will cover the ANOVA and regression topics in the first semester with the second semester devoted to ordinal data, categorical data, and factorial (fixed and mixed models) analyses. Such a division will correspond to the following coverage of the topics in the text:

Semester 1: Chs. 1, 2, 3, 4, 7, 8 (skip 8.8 and 8.9), 9, 10 (skip 10.10), and 11
Semester 2: Chs. 5, 6, Secs. 8.8, 8.9, 10.10, Chs. 12, 13, 14, 15, and 16

The book, supplemented with selected topics from Appendix B, can also be used as the text for a one- or two-semester applied statistics course for first-year graduate students majoring in statistics. Appendix B collects a number of theoretical results that a masters-level statistician should be familiar with.

Supplements

In addition to Mary Sue Younger's *SAS Companion*, which provides SAS software instruction for the examples in *Statistical Research Methods in the Life Sciences*, the following supplementary materials are available: the *Solutions Manual* contains complete solutions for all the problems in the text; the *Student Solutions Manual* contains complete solutions for all the odd-numbered problems in the text; and a data disk (attached to the inside back cover of the text) contains data sets for the problems in the text. The data sets are formatted for SAS, StataQuest, Minitab, and in ASCII.

Acknowledgments

I wish to thank many colleagues and students who contributed directly and indirectly to the development of this book. In particular, I would like to thank my colleague Dennis D. Wackerly, whose many insightful comments have helped me a great deal in the arduous task of completing this writing project. Thanks are also due to Randy L. Carter, for his careful review of an early draft of the first four chapters. I am indebted to Victor Chew, for providing me with data that were used as examples in this book, and to Geoff Vining, for class-testing parts of the book. Thanks are also due to Yoko Tanaka for helping me with the artwork in the text, and to several reviewers for their valuable feedback: Dale O. Everson, University of Idaho; Marvin Lentner, Virginia Tech; Frank G. Margin, University of Florida; Michael Martin, Stanford University; Deborah Rumsey-Johnson, Kansas State University; Mack C. Shelley, Iowa State University; and Mary Sue Younger, University of Tennessee. Finally, and most importantly, I wish to thank my wife, Premila, my daughter Anita and her husband, Ralph, and my son Anil, for their warmth, love, and understanding during the several years when this book was being written.

P. V. Rao

CONTENTS

CHAPTER 7 Designing Research Studies 243

CHAPTER 8 Single-Factor Studies: One-Way ANOVA 279

CHAPTER 9 Single Factor Studies: Comparing Means and Determining Sample Sizes 326

CHAPTER 10 Simple Linear Regression 378

CHAPTER 11 Multiple Linear Regression 463

CHAPTER 12 The General Linear Model 540

1 Statistics: Its Objectives and Scope

1.1 Introduction

The word *statistics* means different things to different people. In everyday usage, statistics are numbers used to summarize information about objects or phenomena. We have statistics pertaining to the performance of a football team, statistics to describe physical aspects of a human being, statistics to summarize characteristics of groups of human beings, statistics to describe weather conditions, and statistics to measure the effectiveness of a drug, for example. In this book, we will use the term in a broader sense: Statistics refers to a body of scientific principles and methodologies that are useful for obtaining information about a phenomenon or a large collection of items. Statistical methods are techniques for using limited amounts of information to arrive at conclusions—called statistical inferences—about the phenomenon or the collection of items of interest. The use of statistics to make inferences is best illustrated by means of some examples.

EXAMPLE 1.1 For residents of West-Central Florida, many of whom live in homes built on reclaimed phosphate mine lands, the possibility of above-normal indoor radiation is a matter of great concern. A regulatory agency investigating the possible health implications of living in these homes—for example, the Environmental Protection Agency (EPA) or the Florida Department of Health and Rehabilitative Services (HRS)—seeks answers to the following questions.

1 What is the average indoor radiation level in homes built on reclaimed phosphate mine lands? How do indoor radiation levels vary between homes within a region?

2 What is the average level of indoor radiation exposure for residents of reclaimed phosphate mine lands? How do these exposure levels vary among residents?

3 What proportion of the residents of these homes are being exposed to radiation levels considered hazardous to human health?

Obviously, the answers to these questions will lead to some inferences about the radiation levels in all homes built on reclaimed lands. As is often the case in scientific investigations, it is virtually impossible to obtain all of the data required for a complete and precise answer to these questions. Measuring indoor radiation levels and determining the exposure levels of occupants in every home would be impractical in terms of time, equipment, and personnel. Consequently, the needed information is obtained on the basis of measurements from relatively few homes and residents; that is, the required inference is based on information contained in a selected subset of homes and residents in the region of interest. ■

EXAMPLE 1.2 The effect of harsh environmental conditions in tropical and subtropical areas is such that many commonly used tropical grasses (for example, Pensacola Bahia grass) are unable to meet the maintenance requirement of grazing livestock. Studies based on chemical analyses and digestibility (quantity digested as a percent of total intake) in laboratory experiments have shown that Mott dwarf elephant grass has the potential to support high levels of animal performance throughout the grazing season. In a recent study, Caceres (1990) compared, for sheep, the digestibility of Mott dwarf elephant grass harvested in June and September with the digestibility of Pensacola Bahia grass harvested in June and September. The questions addressed by Caceres include the following.

1 What is the average digestibility of Pensacola harvested in June? How do digestibility values for Pensacola harvested in June vary from sheep to sheep?

2 Can the average digestibility for Mott harvested in June be expected to be higher than that for Pensacola harvested in June? How much higher?

3 What can be said about the differential digestibility between June- and September-harvested Pensacola compared to the differential digestibility between June- and September-harvested Mott?

Answers to these questions require inferences about the future performance of two varieties of grass harvested at two different times. These inferences must be based on measured digestibility values for a selected number of animals fed each of the four types of grass—June-harvested Pensacola (JP), September-harvested Pensacola (SP), June-harvested Mott (JM), and September-harvested Mott (SM). On that basis, inferences are made about four sets of measurements: (1) the future digestibility values for all sheep fed JP; (2) values for all sheep fed SP; (3) values for all sheep fed JM; (4) values for all sheep fed SM. ■

EXAMPLE 1.3 An investigator is interested in evaluating the relationship between age, blood sugar level, and blood cholesterol level of insulin-dependent diabetics who are on a special experimental diet. The investigator wants to answer the following questions, among others:

1 How does the blood cholesterol level change with age and blood sugar level?

2 Are higher cholesterol levels associated with higher sugar levels?

3 Do older diabetics tend to have higher sugar and cholesterol levels?

These questions relate to a collection of measurements for all insulin-dependent diabetics who are on the special diet. Each measurement in this collection consists of three values—the age, blood sugar level, and blood cholesterol level of a patient. ■

As Examples 1.1–1.3 illustrate, scientific inferences involve two distinct types of collections of measurements: the collection about which information is desired; and the collection from which information is derived. These collections of measurements are called the population and sample, respectively, and are described in more detail in the next section.

1.2 Population and sample

Population

A *population* (sometimes referred to as a statistical population) is a collection (or aggregate) of measurements about which an inference is desired. Example 1.4 contains descriptions of several populations.

EXAMPLE 1.4 The regulatory agencies in Example 1.1 would like to obtain information about indoor radiation levels in homes built on reclaimed phosphate mine lands.

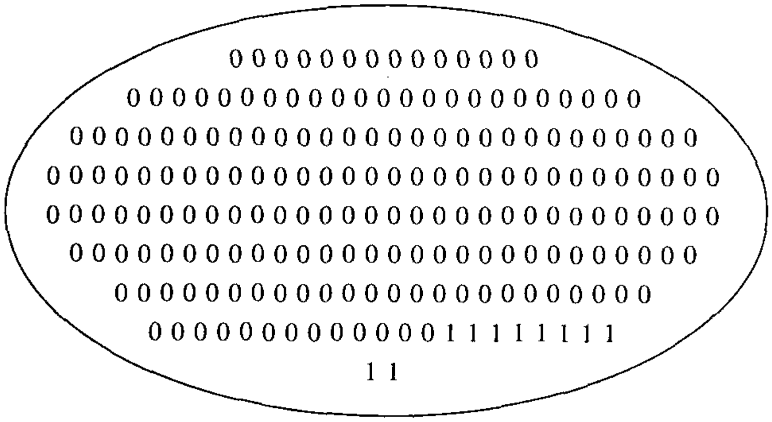
Suppose that there are 4000 homes built on reclaimed lands and that a total of 15,200 persons reside in them. Then, the first question concerning the indoor radiation levels pertains to a population of 4000 radiation-level measurements, one for each home in the group of homes under consideration. The second question in Example 1.1 refers to a population formed by the collection of 15,200 measurements of exposure levels of the residents in the 4000 homes. The third question asks for information about the proportion of residents of the 4000 homes who are being exposed to a hazardous level of radiation. What population does this question refer to? Suppose that, for every individual residing in one of these homes, we place the measurement 1 in the population if the resident is exposed to an unsafe level of radiation; otherwise, we place the measurement 0 in the population. The resulting population is a collection of 15,200 measurements consisting of 0s and 1s. Figure 1.1 shows a 0–1 population of 200 individuals. In this population 1 and 0 represent, respectively, an individual who was and was not exposed to an unsafe level of radiation. Notice that since the population contains 10 measurements equal to 1, 10% of the population have been exposed to unsafe radiation levels. The third question in Example 1.1, then, refers to the proportion of 1s in the 0–1 population.

Now consider Example 1.2. In this example, the scientist is interested in making inferences about the digestibility of four types of forage—JP, SP, JM, and SM—in sheep. The first question refers to the population of digestibility measurements for all sheep who are fed JP. Unlike the three populations in Example 1.1, this population exists only conceptually, because it consists of a set of measurements to be observed in the future. The second question in Example 1.2 refers to two conceptual populations: the potential digestibility measurements for sheep fed JP; and the digestibility measurements for sheep fed JM. The third question in Example 1.2 inquires about the averages of four conceptual populations representing the future digestibility measurements from sheep fed each of the four forage grasses.

Finally, in Example 1.3, the population of interest is a collection of measurements—each of which consists of three values (age, blood sugar level, and blood cholesterol level)—for an insulin-dependent diabetic who is on the experimental diet. ■

FIGURE 1.1

Statistical population representing exposure levels of 200 individuals



0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1
1 1

Note that, in statistics, a measurement is one of the elements that form the population. In certain populations, each measurement may consist of several values. Populations in which each measurement is a single value are called *univariate*