BIOSTATISTICAL ANALYSIS

THIRD EDITION

JERROLD H. ZAR

BIOSTATISTICAL ANALYSIS

THIRD EDITION

JERROLD H. ZAR

Department of Biological Sciences
Northern Illinois University

江苏工业学院图书馆
藏书章

PRENTICE HALL
Upper Saddle River, New Jersey 07458

Library of Congress Cataloging in Publication Data

Zar, Jerrold H.
Biostatistical analysis.

Bibliography.
Includes index.
1. Biometry. I. Title
QH323.5.Z37 1996 574′.01′5195 83-13884
ISBN 0-13-084542-6

Editorial/production supervision: *Electronic Technical Publishing*
Production editor: *Susan Fisher*
Cover director: *Jayne Conte*
Manufacturing manager: *Trudy Pisciotti*
Senior editor: *Sheri L. Snavely*
Editorial assistants: *Nancy Bauer* and *Lisa Tarabokjia*

© 1996, 1984, 1974 by Prentice-Hall, Inc.
Simon & Schuster/A Viacom Company
Upper Saddle River, New Jersey 07458

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

Printed in the United States of America

10 9 8 7 6 5 4 3 2

ISBN 0-13-084542-6

90000>

9 780130 845429

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S. A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Simon & Schuster Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

PREFACE

A great portion of contemporary biological inquiry requires a basic appreciation and knowledge of statistical techniques, as has become apparent to biological researchers, journal editors, and college curriculum planners. Reflecting the magnificent diversity of scientific endeavors that can be found within the biological sciences, this book presents a broad collection of data-analysis techniques, which will address the statistical needs of the majority of biological investigators.

Now in its third edition, this book has been called upon to fulfill two purposes. First, it has served as an introductory textbook, assuming no prior knowledge of statistics. Secondly, it has functioned as a reference work, covering a sufficient variety of concepts and procedures to satisfy a large portion of the biological disciplines that require statistical analysis, and being consulted long after formal instruction has ended.

Colleges and universities have long offered a diverse array of introductory statistics courses, some without emphasis on particular fields in which data might be collected and some—like those for which this book will be explicitly useful—focusing on statistical methods of utility to a specific field. Walker (1929: 148, 151–163) reported that, although the teaching of probability has a much longer history, the first statistics course at a U.S. university or college probably was at Columbia, in the economics department, in 1880; followed in 1887 by the first in psychology, at Pennsylvania; in 1889 by the first in anthropology, at Clark; in 1897 by the first in biology, at Harvard; in 1898 by the first in mathematics, at Illinois; and in 1900 by the first in education, at Columbia. By the end of the nineteenth century only about a dozen institutions offered courses dealing with statistical methods, but by 1929 all universities and most large colleges in the country offered such instruction. Specifically in biology, the first courses with statistical content were probably those taught by Charles B. Davenport at Harvard (1887–1899) and at Chicago (1889–1904), and his *Statistical Methods in Biological Variation* may have been the first American book focused on statistics (ibid.: 159)

In order to be useful as a reference, as well as to allow for differences in content among courses for which it might be used, this book contains much more material than would be included in a one-academic-term course. Therefore, I have been asked to recommend what I consider the basic topics for an introductory treatment. With no authoritarian intent, I suggest these book sections as such a core treatment of biostatistical methods, to be augmented by (or substituted by) others of the instructor's preference: 1.1–1.4, 2.1–2.4, 3.1, 3.2, 3.4, 4.1, 4.4–4.6, 6.1–6.4, 7.1–7.6, 8.1–8.5, 8.9, 8.10, 9.1–9.3, 9.5, 10.1–10.4, 10.6, 11.1–11.7, 12.1–12.7, 14.1, 15.1, 16.1–16.7, 17.1–17.3, 18.1–18.3, 18.9, 19.1–19.4, 21.1, 21.2, 21.4, 21.5, 22.1, 22.3, 22.6.

The material in this book requires no previous mathematical competence beyond very elementary algebra, although the discussions include some topics that appear seldom if at all, in other general texts. Also, cognizance is taken of the increased use of computer capability in academic and nonacademic institutions of all sizes. There are statistical procedures that are of importance but which involve computations so demanding that they practically, if not actually, preclude noncomputer execution. The principles of some of these are presented with the assumption that computer programs (software) will perform the laborious computations, but with the realization that the biologist must enter into the interpretation of the results of the computer's calculations. The data in the examples and exercises are largely fictional and are intended to demonstrate statistical procedures, not biological principles.

A final contribution toward achieving a book with self-sufficiency for most biostatistical needs is the inclusion of a thorough set of statistical tables, the majority of which are more extensive than those found in other introductory or advanced texts, and including many not found in any other texts.

A book of this nature requires and benefits from the assistance of many people. For the preparation of three editions I have been indebted to the library services of the University of Illinois (Urbana), Northern Illinois University, and the latter's collections networks. I also gratefully acknowledge the cooperation of the computer services at Northern Illinois University, which assisted in running many of the computer programs I prepared to generate some of the statistical appendix tables. For the tables taken from previously published sources, thanks are here given for the permission to reprint them; full acknowledgement of each source is found immediately following the appearance of the reprinted material. Additionally, I am pleased to recognize the editorial and production staff at Prentice Hall and Electronic Technical Publishing for their valued professional assistance in transforming my manuscripts into the published product.

Over many years, my teachers, students, and colleagues have aided in guiding me to the material that is presented in this volume. Available space precludes mention of all those providing input and influence to this writing endeavor. However, special recognition must be made of the late S. Charles Kendeigh, University of Illinois (Urbana), who, through considerate mentorship, first alerted me to the need for quantitative analysis of biological data that led me to produce the first edition; the late Edward Batschelet, University of Zurich, who, with enthusiasm, patience, and kindness, provided me with encouragement and inspiration on statistical matters throughout the preparation of much of the first two editions; and the ever supportive and stimulating Arthur W. Ghent, University of Illinois (Urbana), who—from pre-publishing days through the current book edition—has offered statistical and biological commentary both enlightening and challenging. This edition benefited substantially from the manuscript commentary provided by Arthur W. Ghent, Mikel Aickin (Arizona State University), Peter D. Macdonald (McMaster University), Emilia P. Martins (University of Oregon), Daniel M. Pavuk (Bowling Green State University), Trevor Price (University of California, San Diego), and others. Finally, I acknowledge my wife, Carol, for her prolonged patience during the preparation of the three editions of this book over a period of more than twenty years.

J. H. Z.
DeKalb, Illinois

此为试读，需要完整PDF请访问： www.ertongbook.com

CONTENTS

12 TWO-FACTOR ANALYSIS OF VARIANCE 235

13 DATA TRANSFORMATIONS 277

14 MULTIWAY FACTORIAL ANALYSIS OF VARIANCE 285

15 NESTED (HIERARCHICAL) ANALYSIS OF VARIANCE 307

16 SIMPLE LINEAR REGRESSION 317

17 COMPARING SIMPLE LINEAR REGRESSION EQUATIONS 353

18 SIMPLE LINEAR CORRELATION 371

19 MULTIPLE REGRESSION AND CORRELATION 407

20 POLYNOMIAL REGRESSION 447

21 TESTING FOR GOODNESS OF FIT 457

22 CONTINGENCY TABLES 483

23 MORE ON DICHOTOMOUS VARIABLES 513

24 TESTING FOR RANDOMNESS 569

25 CIRCULAR DISTRIBUTIONS: DESCRIPTIVE STATISTICS 591

26 CIRCULAR DISTRIBUTIONS: HYPOTHESIS TESTING 615

APPENDIX A *ANALYSIS OF VARIANCE HYPOTHESIS TESTING* *App1*

APPENDIX B *STATISTICAL TABLES AND GRAPHS* *App11*

ANSWERS TO EXERCISES *Ans1*

LITERATURE CITED *L1*

INDEX *I1*

此为试读,需要完整PDF请访问: www.ertongbook.com

1

INTRODUCTION

Many investigations in the biological sciences are quantitative, with observations consisting of numerical facts called *data*. (One numerical fact is a *datum*.) As biological entities are counted or measured, it becomes apparent that some objective methods are necessary to aid the investigator in presenting and analyzing research data.

The word "statistics" is derived from the Latin for "state," indicating the historical importance of governmental data gathering, which related principally to demographic information (including census data and "vital statistics"), and often to their use in military recruitment and tax collecting.*

This term is often encountered as a synonym for "data": One hears of college enrollment statistics (how many senior students, how many students from each geographic location, etc.), statistics of a baseball game (how many runs scored, how many strike-outs, etc.), labor statistics (numbers of workers unemployed, numbers employed in various occupations), and so on. Hereafter, this use of the word "statistics" will not appear in this book. Instead, "statistics" will be used to refer to the *analysis and interpretation of data with a view toward objective evaluation of the reliability of the conclusions based on the data*. Statistics applied to biological problems is simply called *biostatistics* or, sometimes, *biometry*† (the latter term literally meaning "biological measurement"). Although the field of statistics has roots extending back hundreds of years, its development began in earnest

*Peters (1987: 79) and Walker (1929: 32) attribute the first use of the term "statistics" to a German professor, Gottfried Achenwall (1719–1772), who used the German word *Statistik* in 1749, and the first published use of the English word to John Sinclair (1754–1835) in 1791.

†The term "biometry" apparently was conceived between 1892 and 1901 by Karl Pearson, along with the name, *Biometrika*, for the still-important English journal he helped found, and this term was first published in the inaugural issue of that journal in 1901 (Snedecor, 1954).

in the late nineteenth century, and a major impetus from early in this development has been the need to examine biological data.

Before data can be analyzed, they must be collected, and statistical considerations can aid in the design of experiments and in the setting up of hypotheses to be tested. Many biologists attempt the analysis of their research data only to find that too few data were collected to enable reliable conclusions to be drawn, or that much extra effort was expended in collecting data that cannot be of ready aid in the analysis of the experiment. Thus, a knowledge of basic statistical principles and procedures is important even before an experiment is begun.

Once the data have been obtained, we may organize and summarize them in such a way as to arrive at their orderly presentation. Such procedures are often termed *descriptive statistics*. For example, a tabulation might be made of the heights of all members of a senior English class, indicating an average height for each sex, or for each age. However, it might be desired to make some generalizations from these data. We might, for example, wish to make a reasonable estimate of the heights of all seniors in the university. Or we might wish to conclude whether the males in the university are on the average taller than the females. The ability to make such generalized conclusions, inferring characteristics of the whole from characteristics of its parts, lies within the realm of *inferential statistics*.

1.1 Types of Biological Data

A characteristic that may differ from one biological entity to another is termed a *variable* (or *variate*). Different kinds of variables may be encountered by biologists, and it is desirable to be able to distinguish among them. The classification used here is that which is standardly employed (Senders, 1958; Siegel, 1956; Stevens, 1946, 1968).

However, slavish adherence to this taxonomy can be misleading (e.g., see Velleman and Wilkinson, 1993, and the references cited therein), for not all data fit neatly into these categories and some may be treated differently depending upon the questions asked of them.

Data on a Ratio Scale. Consider that the heights of a group of plants constitute a variable of interest, and perhaps the number of leaves per plant is another variable under consideration. Thanks to measuring devices at the biologist's disposal, it is possible to assign a numerical value to the height of each plant, and counting the leaves allows a numerical value to be assigned to the number of leaves on each plant. Regardless of whether the height measurements are recorded in centimeters, inches, or other units, and regardless of whether the leaves are counted in a number system using base 10 or any other base, there are two fundamentally important characteristics of these data.

First, there is a constant size interval between any adjacent units on the measurement scale. That is, the difference in height between a 36 cm and a 37 cm plant is the same as the difference between a 39 cm and a 40 cm plant, and the difference between eight and ten leaves is equal to the difference between nine and eleven leaves. (This may seem simpleminded, but it is very important, as we shall see on examining the other scales of measurement.)

Second, it is important that there exists a zero point on the measurement scale and that there is a physical significance to this zero. This enables us to say something meaningful about the ratio of measurements. We can say that a 30 cm (11.8 in.) tall plant is half as tall as a 60 cm (23.6 in.) plant, and that a plant with forty-five leaves has three times as many leaves as a plant with fifteen.

Measurement scales having a constant interval size and a true zero point are said to be *ratio scales* of measurement. Besides lengths and numbers of items, ratio scales include weights (mg, lb, etc.), volumes (cc, cu ft, etc.), capacities (ml, qt, etc.), rates (cm/sec, mph, mg/min, etc.), and lengths of time (hr, yr, etc.).

Data on an Interval Scale. Some measurement scales possess a constant interval size but not a true zero; they are called *interval scales*. An outstanding example is that of the two common temperature scales: Celsius (C) and Fahrenheit (F). We can see that the same difference exists between 20°C (68°F) and 25°C (77°F) as between 5°C (41°F) and 10°C (50°F); i.e., the measurement scale is composed of equal-sized intervals. But it cannot be said that a temperature of 40°C (104°F) is twice as hot as a temperature of 20°C (68°F); i.e., the zero point is arbitrary. (Temperature measurements on the absolute, or Kelvin [K], scale can be referred to a physically meaningful zero and thus constitute a ratio scale.)

Some interval scales encountered in biological data collection are *circular scales*. Time of day and time of the year are examples of such scales. The interval between 2:00 PM (i.e., 1400 hr) and 3:30 PM (1530 hr) is the same as the interval between 8:00 AM (0800 hr) and 9:30 AM (0930 hr). But one cannot speak of ratios of times of day because the zero point (midnight) on the scale is arbitrary, in that one could just as well set up a scale for time of day which would have noon, or 3:00 PM, or any other time as the zero point. Circular biological data are occasionally compass points, as if one records the compass direction in which an animal or plant is oriented. Since the designation of north as 0° is arbitrary, this circular scale is a form of interval scale of measurement. Some special statistical procedures are available for circular data; these are discussed in chapters 25 and 26.

Data on an Ordinal Scale. The preceding paragraphs on ratio and interval scales of measurement discussed data between which we know numerical differences. For example, if man A weighs 80 kg and man B weighs 70 kg, then man A is known to weigh 10 kg more than B. But our data may, instead, be a record only of the fact that man A weighs more than man B (with no indication of how much more). Thus, we may be dealing with relative differences rather than with quantitative differences. Such data consist of an ordering or ranking of measurements and are said to be on an *ordinal* scale of measurement ("ordinal" being from the Latin word for "order"). One may speak of one biological entity being shorter, darker, faster, or more active than another; the sizes of five cell types might be labeled 1, 2, 3, 4, and 5, to denote their magnitudes relative to each other; or success in learning to run a maze may be recorded as A, B, or C.

It is often true that biological data expressed on the ordinal scale could have been expressed on the interval or ratio scale had exact measurements been obtained (or obtainable). Sometimes data that were originally on interval or ratio scales will be changed to ranks; for example, examination grades of 99, 85, 73, and 66% (ratio scale) might be recorded as A, B, C, and D (ordinal scale), respectively.

Ordinal scale data contain and convey less information than ratio or interval data, for only relative magnitudes are known. Consequently, quantitative comparisons are impossible (e.g., we cannot speak of a grade of C being half as good as a grade of A, or of the difference between cell sizes 1 and 2 being the same as the difference between sizes 3 and 4). However, we will see that many useful statistical procedures are, in fact, applicable to ordinal data.

Data on a Nominal Scale. Sometimes the variable under study is classified by some quality it possesses rather than by a numerical measurement. In such cases the variable is called an *attribute*, and we are said to be using a *nominal scale* of measurement. Genetic phenotypes are commonly encountered biological attributes; the possible manifestation of an animal's eye color may be blue or brown, and if human hair color were the attribute of interest, we might record black, brown, blonde, or red. On a nominal scale ("nominal" is from the Latin word for "name"), animals might be classified as male or female, or as left- or right-handed. Or plants might be classified as dead or alive, or as with or without thorns. Taxonomic categories also form a nominal classification scheme (e.g., a plant might be classified as pine, spruce, or fir). Sometimes data from an ordinal, interval, or ratio scale of measurement may be recorded in nominal-scale categories. For example, heights may be recorded as tall or short, or performance on an examination as pass or fail.

As will be seen, statistical methods useful with ratio, interval, or ordinal data generally are not applicable to nominal data, and we must, therefore, be able to identify such situations when they occur.

Continuous and Discrete Data. When we spoke above of plant heights, we were dealing with a variable that could be any conceivable value within any observed range; this is referred to as a *continuous variable*. That is, if we measure a height of 35 cm and a height of 36 cm, an infinite number of heights is possible in the range from 35 to 36 cm: a plant might be 35.07 cm tall or 35.988 cm tall, or 35.3263 cm tall, etc., although, of course, we do not have devices sensitive enough to detect this infinity of heights. A continuous variable is one for which there is a possible value between any other two possible values.

However, when speaking of the number of leaves on a plant, we are dealing with a variable that can take on only certain values. It might be possible to observe 27 leaves, or 28 leaves, but 27.43 leaves and 27.9 leaves are values of the variable that are impossible to obtain. Such a variable is termed a *discrete* or *discontinuous variable* (also known as a *meristic variable*). The number of white blood cells in 1 mm^3 of blood, the number of giraffes visiting a water hole, and the number of eggs laid by a grasshopper are all discrete variables. The possible values of a discrete variable generally are consecutive integers, but this is not necessarily so. If the leaves on our plants are always formed in pairs, then only even integers are possible values of the variable. And the ratio of number of wings to number of legs of insects is a discrete variable that may only have the value of 0, 0.3333..., or 0.6666... (i.e., $\frac{0}{6}$, $\frac{2}{6}$, or $\frac{4}{6}$, respectively).*

Ratio-, interval-, and ordinal-scale data may be either continuous or discrete. Nominal-scale data by their nature are discrete.

1.2 Accuracy and Significant Figures

Accuracy is the nearness of a measurement to the actual value of the variable being measured. *Precision* is not a synonymous term, but refers to the closeness to each other of repeated measurements of the same quantity.

If we report that the hind leg of a frog is 8 cm long, we are stating the number 8 (a value of a continuous variable) as an estimate of the frog's true leg length. This estimate was made using some sort of a measuring device. Had the device been capable of more accuracy, we might have concluded that the leg was 8.3 cm long, or perhaps 8.32 cm long. When recording values of continuous variables, it is important to designate the accuracy with which the measurements have been made. By convention, the value 8 denotes a measurement in the range of 7.50000 ... to 8.49999..., the value 8.3 designates a range of 8.25000 ... to 8.34999..., and the value 8.32 implies that the true value lies within the range of 8.31500... to 8.32499.... That is, the reported value is the midpoint of the implied range, and the size of this range is designated by the last decimal place in the measurement. The value of 8 cm implies a range of accuracy of 1 cm, 8.3 cm implies a range of 0.1 cm, and 8.32 cm implies a range of 0.01 cm. Thus, to record a value of 8.0 implies greater accuracy of measurement than does the recording of a value of 8, for in the first instance the true value is said to lie between 7.95000 ... and 8.049999... (i.e., within a range of 0.1 cm), whereas 8 implies a value between 7.50000 ... and 8.49999... (i.e., within a range of 1 cm). To state 8.00 cm implies an accuracy in measurement which ascertains the frog's limb length to be between 7.99500 ... and 8.00499 ... cm (i.e., within a range of 0.01 cm). Those digits in a number that denote the accuracy of the measurement are referred to as *significant figures*. Thus, 8 has one significant figure, 8.0 and 8.3 each have two significant figures, and 8.00 and 8.32 each have three.

In working with exact values of discrete variables, the preceding considerations do not apply. That is, it is sufficient to state that our frog has four limbs or that its left lung contains thirteen flukes. The use of 4.0 or 13.00 would be inappropriate, for since the numbers involved are exactly 4 and 13, there is no question of accuracy or significant figures.

But there are instances where significant figures and implied accuracy come into play with discrete data. An entomologist may report that there are 72,000 moths in a particular forest area. In doing so, it is probably not being claimed that this is the exact

---

*The ellipses (...) may be read as "and so on." Here, they indicate that $\frac{2}{6}$ and $\frac{4}{6}$ are repeating decimal fractions, which could just as well have been written as 0.3333333333333 ... and 0.6666666666666..., respectively.

number but an estimate of the exact number, perhaps accurate to two significant figures. In such a case, 72,000 would imply a range of accuracy of 1000, so that the true value might lie anywhere from 71,500 to 72,500. If the entomologist wished to convey the fact that this estimate is believed to be accurate to the nearest 100 (i.e., to three significant figures), rather than to the nearest 1000, it would be better to present the data in the form of *scientific notation*, as follows: If the number 7.2×10^4 (= 72,000) is written, a range of accuracy of 0.1×10^4 (= 1000) is implied, and the true value is assumed to lie between 71,500 and 72,500. But if 7.20×10^4 were written, a range of accuracy of 0.01×10^4 (= 100) would be implied, and the true value would be assumed to be in the range of 71,950 to 72,050. Thus, the accuracy of large values (and this applies to continuous as well as discrete variables) can be expressed succinctly using scientific notation.

Calculators and computers typically yield results with more significant figures than are justified by the data. However, it is good practice—to avoid rounding error—to retain many significant figures for all steps until the last in a sequence of calculations, and on attaining the result of the final step to round off to the appropriate number of figures.

1.3 Frequency Distributions

When collecting and summarizing large amounts of data, it is often helpful to record the data in the form of a *frequency table*. Such a table simply involves a listing of all the observed values of the variable being studied and how many times each value is observed. Consider the tabulation of the frequency of occurrence of sparrow nests in each of several different locations. This is illustrated in Example 1.1, where the observed nest sites are listed, and for each site the number of nests observed is recorded. The distribution of the total number of observations among the various categories is termed a *frequency distribution*. Example 1.1 is a frequency table for nominal data, and these data may also be presented graphically by means of a *bar graph* (Fig. 1.1), where the height of each bar is proportional to the frequency in the class represented. The widths of all bars in a bar graph should be equal so that the eye of the reader is not distracted from the differences in bar heights; this also makes the area of each bar proportional to the frequency it represents. Also, the frequency scale on the vertical axis should begin at zero to avoid the apparent differences among bars. If, for example, a bar graph of the data of Example 1.1 were constructed with the vertical axis representing frequencies

EXAMPLE 1.1 The location of sparrow nests. A frequency table of nominal data.

Nest site	*Number of nests observed*
A. Vines	56
B. Building eaves	60
C. Low tree branches	46
D. Tree and building cavities	49

此为试读，需要完整PDF请访问：www.ertongbook.com