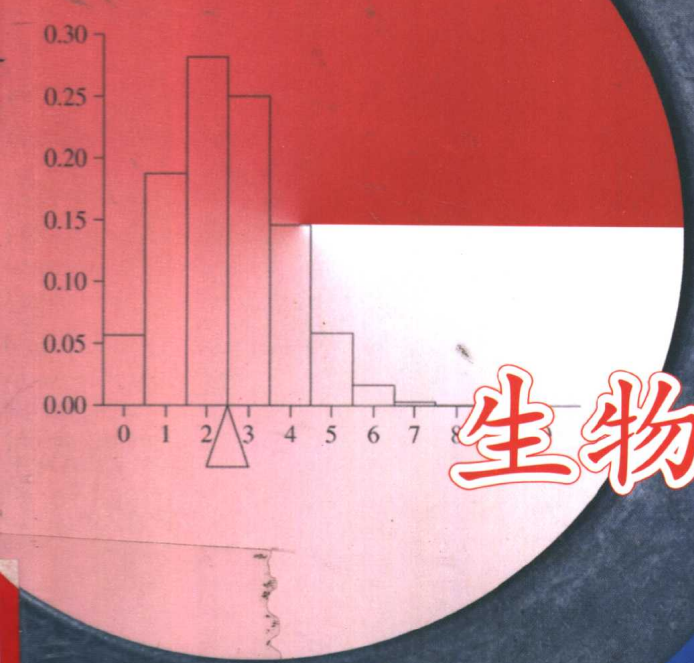
国外大学生物学优秀教材（影印版）

An Introduction To Biostatistics



GLOVER & MITCHELL

生物统计学导论

Mc Graw Hill

清华大学出版社

Mc Graw Hill 麦格劳-希尔教育出版集团

An Introduction To
Biostatistics

生物统计学导论

Thomas Glover
Hobart and William Smith Colleges

Kevin Mitchell
Hobart and William Smith Colleges

清华大学出版社
http://www.tup.tsinghua.edu.cn

麦格劳-希尔教育出版集团

(京)新登字 158 号

北京市版权局著作权合同登记号:图字 01-2001-4063

书　名:生物统计学导论
作　者:Glover&Mitchell
出版者:清华大学出版社(北京清华大学学研大厦,邮编 100084)
　　http://www.tup.tsinghua.edu.cn
印刷者:北京牛山世兴印刷厂
发行者:新华书店总店北京发行所
开　本:787×960 1/16　印张:27
版　次:2001 年 10 月第 1 版　2001 年 10 月第 1 次印刷
书　号:ISBN 7-302-04819-3/Q·15
印　数:0001~3000
定　价:38.00 元

出 版 前 言

为了使生物学教学适应21世纪生命科学发展的需要，同时也为了提高学生阅读专业文献和获取信息的能力，结合当前生物学在高等院校中教学的实际情况，我们精选了一些国外优秀的生物教材，组织专家进行了评阅和审核，组成国外大学生物学优秀教材系列(影印版)。该系列反映了国外大学生物学教材的最新内容和编写特色，多数教材经过教学实践，被国外很多大学广泛采用，并获得好评，因而不断再版。本书即是其中的一册。

希望这套教材能对高等院校师生和广大科技人员有所帮助，同时对我国的生命科学赶超世界先进水平起到一定的推动作用。

欢迎广大读者将使用本系列教材后的意见反馈给我们，更欢迎国内外专家、教授积极向我社推荐国外的优秀生物学教材，以便我们将国外大学生物学优秀教材系列做得更好。

清华大学出版社

2001年8月

To our families:

Emily Glover

Ellen, Dianne, and Dorothy Mitchell

and to Professor Sydney S. Y. Young for his mentorship of T. Glover and clear exposition of biostatistics so many years ago.

PREFACE

Our goal in writing this book was to generate an accessible and relatively complete introduction for undergraduates to the use of statistics in the biological sciences. The text is designed for a one-quarter or one-semester class in introductory statistics for the life sciences. The target audience is sophomore and junior biology, environmental studies, biochemistry, and health sciences majors. Appropriate prerequisites include some coursework in biology as well as a foundation in algebra but not calculus. Examples are taken from many areas in the life sciences including genetics, physiology, ecology, agriculture, and medicine.

This text emphasizes the relationships between probability, probability distributions, and hypothesis testing. We have tried to highlight the expected value of various test statistics under the null and research hypotheses as a way to understand the methodology of hypothesis testing. In addition, we have incorporated nonparametric alternatives to many situations along with the standard parametric analysis. These nonparametric techniques are included because undergraduate student projects often involve sampling populations where the underlying distribution is unknown, and because the development of the nonparametric tests is usually readily understandable for students with modest math backgrounds. The nonparametrics can be skipped or skimmed without any loss of continuity.

We have tried to include interesting and easily understandable examples with each concept. The problems at the end of each chapter have a range of difficulty and come from a variety of disciplines. Most are not real-life examples but are realistic in their design and data values. The end-of-chapter problems are randomized within each chapter to require the student to choose the appropriate analysis. Many undergraduate texts present a concept or test and immediately give all the problems that can be solved by that technique. This approach prevents students from having to make the real-life decision about the appropriate analysis. We believe this decision making is a critical skill in the introduction of statistical analysis and have provided a large number of opportunities to practice and develop this skill.

The material for this text derives principally from a required biostatistics course one of us (Glover) has taught to undergraduates for more than 20 years and from a second course in nonparametric statistics and field data analysis that the other of us (Mitchell) has taught more recently during several term abroad programs to Queensland, Australia. Recent shifts in undergraduate curricula have deemphasized calculus for biology

students and are now highlighting statistical analysis as a fundamental quantitative skill. Hopefully our text will make teaching and learning that skill somewhat less arduous.

Supplemental Materials

The material in this textbook can be supported by a wide variety of statistical packages and ancillary materials. The selection of these support materials is usually dictated by personal interests and cost considerations. Here we wish to highlight four items that we have found quite useful in teaching biostatistics to undergraduates. Presently we use Statview Software developed by SAS Institute Inc. in the laboratory sessions of our course. This software is easy to use, relatively flexible and can complete nearly all the statistical techniques presented in our text. Information regarding this program can be found at **www.Statview.com**. A second very useful and accessible statistical package is MINITAB by Minitab Inc. Information regarding this statistical software can be found at **www.minitab.com**.

For student purchase we have used Texas Instrument calculators ranging from the TI-35 model to the TI-83 model. The price range for those calculators is considerable and might clearly be a factor in choosing a required calculator for a particular course. Although calculators such as the TI-35 do less automatically, they sometimes give the student clearer insights into the statistical tests by requiring a few more computational steps. The ease of computation afforded by computer programs or sophisticated calculators sometimes leads to a "black box" mentality about statistics and their calculation.

Finally, for both students and instructors we recommend D. J. Hand et al., editors, 1994, *A Handbook of Small Data Sets*, Chapman & Hall, London. This book contains 510 small data sets ranging from the numbers of Prussian military personel killed by horse kicks from 1875–1894 (data set #283) to the shape of bead work on leather goods of Shoshoni Indians (data set #150). The data sets are interesting, manageable, and amenable to statistical analysis using techniques presented in our text. While 16 of the data sets from the Handbook were utilized as examples or problems in our text, there are many others that could serve as engaging and useful practice problems.

The Instructor's Answer Manual is available on this text's website at **www.mhhe.com/zoology** (click on this book's cover). Instructors can access a printable version with both questions and answers that correlate to this text, or a printable version with answers only. For students, the book's website offers additional practice questions, as well as web links to pertinent papers and information.

PageOut® is the solution for professors who need to build a course website. Features of this service include the following:

- The PageOut Library offers instant access to fully loaded course websites with no work required on the instructor's part.
- Courses can now be password protected.
- Professors can now upload, store, and manage up to 10 MB of data.
- Professors can copy their course and share it with colleagues, or use it as a foundation for next semester.

Short on time? Let us do the work. Our McGraw-Hill service team is ready to build your PageOut website, and provide content and any necessary training. Learn more about PageOut and other McGraw-Hill digital solutions at **www.mhhe.com/solutions**.

Acknowledgments

Thanks are due to the following people at McGraw-Hill for their support and guidance throughout the preparation of this text: Marge Kemp, Sponsoring Editor; Donna Nemmers, Developmental Editor; Gloria Schiesl, Senior Project Manager; James W. Bradley, Copy Editor; and David W. Hash, design coordinator for the cover.

We further thank our colleagues Joel T. Kerlan and James M. Ryan for their encouragement, Ann Warner for meticulously word processing the manuscript, and the students of Hobart and William Smith Colleges for their many comments and suggestions, particularly Aline Gadue for her careful scrutiny of early drafts.

Finally, we gratefully thank the following reviewers of the manuscript for this first edition: Tyler Haynes, *Boston University;* William A. Hayes, *Delta State University;* Andrew Jay Tierman, *Saginaw Valley State University;* John E. Weinstein, *Texas A & M University, Commerce;* Brenda L. Young, *Daemen College*. Their insights and suggestions have been most helpful.

Thomas J. Glover
Kevin J. Mitchell

BRIEF CONTENTS

CONTENTS

1

Introduction to Data Analysis

Concepts in Chapter 1:

- Scientific Method and Statistical Analysis
- Parameters: Descriptive Characteristics of Populations
- Statistics: Descriptive Characteristics of Samples
- Variable Types: Continuous, Discrete, Ranked, and Categorical
- Measures of Central Tendency: Mean, Median, and Mode
- Measures of Dispersion: Range, Variance, and Standard Deviation
- Descriptive Statistics for Frequency Data
- Effects of Coding on Descriptive Statistics
- Tables and Graphs
- Quartiles and Box Plots
- Accuracy, Precision, and the 30–300 Rule

1.1 Introduction

The modern study of the life sciences includes experimentation, data gathering, and interpretation. The following text offers an introduction to the methods used to perform these fundamental activities.

The design and evaluation of experiments, formally known as the **scientific method,** is utilized in all scientific fields and is often implied rather than explicitly outlined in many investigations. The components of the scientific method include observation, formulation of a potential question or problem, construction of a hypothesis, followed by a prediction, and the design of an experiment to test the prediction. Let's consider these components briefly.

Observation of a Particular Event

Generally an observation can be classified as either quantitative or qualitative. Quantitative observations are based on some sort of measurement, e.g., length, weight,

temperature, and pH. Qualitative observations are based on categories reflecting a quality or characteristic of the observed event, e.g., male versus female, diseased versus healthy, and mutant versus wild type.

Statement of the Problem

A series of observations often leads to the formulation of a particular problem or unanswered question. This usually takes the form of a "why" question and implies a cause and effect relationship. For example, suppose upon investigating a remote Fijian island community you realized that the vast majority of the adults suffer from hypertension (abnormally elevated blood pressures with the systolic over 165 mmHg and the diastolic over 95 mmHg). Note that the individual observations here are quantitative while the percentage that are hypertensive is based on a qualitative evaluation of the sample. From these preliminary observations one might formulate the question: Why are so many adults in this population hypertensive?

Formulation of a Hypothesis

A hypothesis is a tentative explanation for the observations made. A good hypothesis suggests a cause and effect relationship and is testable.

The Fijian community may demonstrate hypertension because of diet, life style, genetic makeup, or combinations of these factors. Because we've noticed extraordinary consumption of octopi in their diet and knowing octopods have a very high cholesterol content, we might hypothesize that the high level of hypertension is caused by diet.

Making a Prediction

If the hypothesis is properly constructed, it can and should be used to make predictions. Predictions are based on deductive reasoning and take the form of an "if-then" statement. For example, a good prediction based on the hypothesis above would be: If the hypertension is caused by a high cholesterol diet, then changing the diet to a low cholesterol one should lower the incidence of hypertension.

The criteria for a valid (properly stated) prediction are:

1. An "if" clause stating the hypothesis.
2. A "then" clause that
   (a) suggests altering a causative factor in the hypothesis (change of diet);
   (b) predicts the outcome (lower level of hypertension);
   (c) provides the basis for an experiment.

Design of the Experiment

The entire purpose and design of an experiment is to accomplish one goal, that is, to test the hypothesis. An experiment tests the hypothesis by testing the correctness or incorrectness of the predictions that came from it. Theoretically, an experiment should alter or test only the factor suggested by the prediction, while all other factors remain constant.

How would you design an experiment to test the diet hypothesis in the hypertensive population?

The best way to test the hypothesis above is by setting up a controlled experiment. This might involve using two randomly chosen groups of adults from the community and treating both identically with the exception of the one factor being tested. The control group represents the "normal" situation, has all factors present, and is used as a standard or basis for comparison. The experimental group represents the "test" situation and includes all factors except the variable that has been altered, in this case the diet. If the group with the low cholesterol diet exhibits *significantly* lower levels of hypertension, the hypothesis is supported by the data. On the other hand, if the change in diet has no effect on hypertension, then a new or revised hypothesis should be formulated and the experimental procedure redesigned. Finally, the generalizations that are drawn by relating the data to the hypothesis can be stated as conclusions.

While these steps outlined above may seem straightforward, they often require considerable insight and sophistication to apply properly.

In our example how the groups are chosen is not a trivial problem. They must be constructed without bias and must be large enough to give the researcher an acceptable level of confidence in the results. Further, how large a change is significant enough to support the hypothesis? What is *statistically significant* may not be *biologically significant*. How can one be objective and make decisions in the face of uncertainty?

A foundation in statistical methods will help you design and interpret experiments properly. The field of statistics is broadly defined as the methods and procedures for collecting, classifying, summarizing, and analyzing data, and utilizing the data to test scientific hypotheses. The term *statistics* is derived from the Latin for state, and originally referred to information gathered in various censuses that could be numerically summarized to describe aspects of the state, e.g., bushels of wheat per year, number of military aged men, etc. Over time statistics has come to mean the scientific study of numerical data based on natural phenomena. Statistics applied to the life sciences is often called **biostatistics** or **biometry**. The foundations of biostatistics go back several hundred years, but statistical analysis of biological systems began in earnest in the late nineteenth century as biology became more quantitative and experimental.

1.2 Populations and Samples

Today we use statistics as a means of informing the decision-making processes in the face of the uncertainties that most real world problems present. Often we wish

to make generalizations about populations that are too large or too difficult to survey completely. In these cases we sample the population and use characteristics of the sample to extrapolate to characteristics of the larger population. See Figure 1.1.

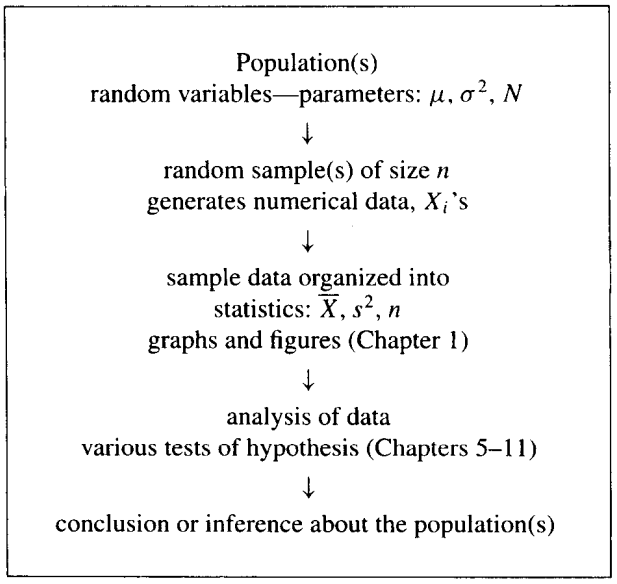


FIGURE 1.1
The general approach to statistical analysis.

Real-world problems concern large groups or **populations** about which inferences must be made. (Is there a size difference between two color morphs of the same species of sea star? Are the offspring of a certain cross of fruit flies in a 3:1 ratio of normal to eyeless?) Certain characteristics of the population are of particular interest (systolic blood pressure, weight in grams, resting body temperature). The values of these characteristics will vary from individual to individual within the population. These characteristics are called **random variables** because they vary in an unpredictable way or in a way that appears or is assumed to depend on chance. The different types of variables are described in Section 1.3.

A descriptive measure associated with a random variable when it is considered over the *entire population* is called a **parameter**. Examples are the mean weight of all green turtles, *Chelonia mydas*, or the variance in clutch size of the tiger snake, *Notechis scutatus*. In general, such parameters are difficult, if not impossible, to determine because the population is too large or expensive to study in its entirety. Consequently, one is forced to examine a subset or **sample** of the population and make inferences about the entire population based on this sample. A descriptive measure associated with a random variable of a *sample* is called a **statistic**. The mean weight of 25 female green turtles laying eggs on Heron Island or the variability in clutch size of 50 clutches of tiger snake eggs collected in southeastern Queensland are examples of statistics.