



SECOND EDITION

# INTRODUCTORY **BIOSTATISTICS**

Chap T. Le • Lynn E. Eberly



WILEY



# **INTRODUCTORY BIOSTATISTICS**



# INTRODUCTORY BIostatISTICS

---

Second Edition

**CHAP T. LE**

Distinguished Professor of Biostatistics  
Director of Biostatistics and Bioinformatics  
Masonic Cancer Center  
University of Minnesota

**LYNN E. EBERLY**

Associate Professor of Biostatistics  
School of Public Health  
University of Minnesota

**WILEY**

Copyright © 2016 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.  
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at [www.copyright.com](http://www.copyright.com). Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

**Limit of Liability/Disclaimer of Warranty:** While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at [www.wiley.com](http://www.wiley.com).

### ***Library of Congress Cataloging-in-Publication Data***

Names: Le, Chap T., 1948– | Eberly, Lynn E.

Title: Introductory biostatistics.

Description: Second edition / Chap T. Le, Lynn E. Eberly. | Hoboken, New Jersey : John Wiley & Sons, Inc., 2016. | Includes bibliographical references and index.

Identifiers: LCCN 2015043758 (print) | LCCN 2015045759 (ebook) | ISBN 9780470905401 (cloth) | ISBN 9781118595985 (Adobe PDF) | ISBN 9781118596074 (ePub)

Subjects: LCSH: Biometry. | Medical sciences—Statistical methods.

Classification: LCC QH323.5 .L373 2016 (print) | LCC QH323.5 (ebook) | DDC 570.1/5195—dc23

LC record available at <http://lccn.loc.gov/2015043758>

Set in 10/12pt Times by SPi Global, Pondicherry, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

*To my wife, Minhha, and my daughters, Mina and Jenna with love*

C.T.L.

*To my husband, Andy, and my sons, Evan, Jason, and Colin, with love;  
you bring joy to my life*

L.E.E.





# CONTENTS

<b>Preface to the Second Edition</b>	<b>xiii</b>
<b>Preface to the First Edition</b>	<b>xv</b>
<b>About the Companion Website</b>	<b>xix</b>
<b>1 Descriptive Methods for Categorical Data</b>	<b>1</b>
1.1 Proportions	1
1.1.1 Comparative Studies	2
1.1.2 Screening Tests	5
1.1.3 Displaying Proportions	7
1.2 Rates	10
1.2.1 Changes	11
1.2.2 Measures of Morbidity and Mortality	13
1.2.3 Standardization of Rates	15
1.3 Ratios	18
1.3.1 Relative Risk	18
1.3.2 Odds and Odds Ratio	18
1.3.3 Generalized Odds for Ordered $2 \times k$ Tables	21
1.3.4 Mantel–Haenszel Method	25
1.3.5 Standardized Mortality Ratio	28
1.4 Notes on Computations	30
Exercises	32
<b>2 Descriptive Methods for Continuous Data</b>	<b>55</b>
2.1 Tabular and Graphical Methods	55
2.1.1 One-Way Scatter Plots	55
2.1.2 Frequency Distribution	56
2.1.3 Histogram and Frequency Polygon	60

2.1.4	Cumulative Frequency Graph and Percentiles	64
2.1.5	Stem and Leaf Diagrams	68
2.2	Numerical Methods	69
2.2.1	Mean	69
2.2.2	Other Measures of Location	72
2.2.3	Measures of Dispersion	73
2.2.4	Box Plots	76
2.3	Special Case of Binary Data	77
2.4	Coefficients of Correlation	78
2.4.1	Pearson's Correlation Coefficient	80
2.4.2	Nonparametric Correlation Coefficients	83
2.5	Notes on Computations	85
	Exercises	87
<b>3</b>	<b>Probability and Probability Models</b>	<b>103</b>
3.1	Probability	103
3.1.1	Certainty of Uncertainty	104
3.1.2	Probability	104
3.1.3	Statistical Relationship	106
3.1.4	Using Screening Tests	109
3.1.5	Measuring Agreement	112
3.2	Normal Distribution	114
3.2.1	Shape of the Normal Curve	114
3.2.2	Areas Under the Standard Normal Curve	116
3.2.3	Normal Distribution as a Probability Model	122
3.3	Probability Models for Continuous Data	124
3.4	Probability Models for Discrete Data	125
3.4.1	Binomial Distribution	126
3.4.2	Poisson Distribution	128
3.5	Brief Notes on the Fundamentals	130
3.5.1	Mean and Variance	130
3.5.2	Pair-Matched Case-Control Study	130
3.6	Notes on Computations	132
	Exercises	134
<b>4</b>	<b>Estimation of Parameters</b>	<b>141</b>
4.1	Basic Concepts	142
4.1.1	Statistics as Variables	143
4.1.2	Sampling Distributions	143
4.1.3	Introduction to Confidence Estimation	145
4.2	Estimation of Means	146
4.2.1	Confidence Intervals for a Mean	147
4.2.2	Uses of Small Samples	149
4.2.3	Evaluation of Interventions	151
4.3	Estimation of Proportions	153

4.4	Estimation of Odds Ratios	157
4.5	Estimation of Correlation Coefficients	160
4.6	Brief Notes on the Fundamentals	163
4.7	Notes on Computations	165
	Exercises	166
<b>5</b>	<b>Introduction to Statistical Tests of Significance</b>	<b>179</b>
5.1	Basic Concepts	180
5.1.1	Hypothesis Tests	181
5.1.2	Statistical Evidence	182
5.1.3	Errors	182
5.2	Analogies	185
5.2.1	Trials by Jury	185
5.2.2	Medical Screening Tests	186
5.2.3	Common Expectations	186
5.3	Summaries and Conclusions	187
5.3.1	Rejection Region	187
5.3.2	$p$ Values	189
5.3.3	Relationship to Confidence Intervals	191
5.4	Brief Notes on the Fundamentals	193
5.4.1	Type I and Type II Errors	193
5.4.2	More about Errors and $p$ Values	194
	Exercises	194
<b>6</b>	<b>Comparison of Population Proportions</b>	<b>197</b>
6.1	One-Sample Problem with Binary Data	197
6.2	Analysis of Pair-Matched Data	199
6.3	Comparison of Two Proportions	202
6.4	Mantel–Haenszel Method	206
6.5	Inferences for General Two-Way Tables	211
6.6	Fisher’s Exact Test	217
6.7	Ordered $2 \times K$ Contingency Tables	219
6.8	Notes on Computations	222
	Exercises	222
<b>7</b>	<b>Comparison of Population Means</b>	<b>235</b>
7.1	One-Sample Problem with Continuous Data	235
7.2	Analysis of Pair-Matched Data	237
7.3	Comparison of Two Means	242
7.4	Nonparametric Methods	246
7.4.1	Wilcoxon Rank-Sum Test	246
7.4.2	Wilcoxon Signed-Rank Test	250
7.5	One-Way Analysis of Variance	252
7.5.1	One-Way Analysis of Variance Model	253
7.5.2	Group Comparisons	258

7.6	Brief Notes on the Fundamentals	259
7.7	Notes on Computations	260
	Exercises	260
<b>8</b>	<b>Analysis of Variance</b>	<b>273</b>
8.1	Factorial Studies	273
8.1.1	Two Crossed Factors	273
8.1.2	Extensions to More Than Two Factors	278
8.2	Block Designs	280
8.2.1	Purpose	280
8.2.2	Fixed Block Designs	281
8.2.3	Random Block Designs	284
8.3	Diagnostics	287
	Exercises	291
<b>9</b>	<b>Regression Analysis</b>	<b>297</b>
9.1	Simple Regression Analysis	298
9.1.1	Correlation and Regression	298
9.1.2	Simple Linear Regression Model	301
9.1.3	Scatter Diagram	302
9.1.4	Meaning of Regression Parameters	302
9.1.5	Estimation of Parameters and Prediction	303
9.1.6	Testing for Independence	307
9.1.7	Analysis of Variance Approach	309
9.1.8	Some Biomedical Applications	311
9.2	Multiple Regression Analysis	317
9.2.1	Regression Model with Several Independent Variables	318
9.2.2	Meaning of Regression Parameters	318
9.2.3	Effect Modifications	319
9.2.4	Polynomial Regression	319
9.2.5	Estimation of Parameters and Prediction	320
9.2.6	Analysis of Variance Approach	321
9.2.7	Testing Hypotheses in Multiple Linear Regression	322
9.2.8	Some Biomedical Applications	330
9.3	Graphical and Computational Aids	334
	Exercises	336
<b>10</b>	<b>Logistic Regression</b>	<b>351</b>
10.1	Simple Regression Analysis	353
10.1.1	Simple Logistic Regression Model	353
10.1.2	Measure of Association	355
10.1.3	Effect of Measurement Scale	356
10.1.4	Tests of Association	358
10.1.5	Use of the Logistic Model for Different Designs	358
10.1.6	Overdispersion	359

10.2	Multiple Regression Analysis	362
10.2.1	Logistic Regression Model with Several Covariates	363
10.2.2	Effect Modifications	364
10.2.3	Polynomial Regression	365
10.2.4	Testing Hypotheses in Multiple Logistic Regression	365
10.2.5	Receiver Operating Characteristic Curve	372
10.2.6	ROC Curve and Logistic Regression	374
10.3	Brief Notes on the Fundamentals	375
10.4	Notes on Computing	377
	Exercises	377
<b>11</b>	<b>Methods for Count Data</b>	<b>383</b>
11.1	Poisson Distribution	383
11.2	Testing Goodness of Fit	387
11.3	Poisson Regression Model	389
11.3.1	Simple Regression Analysis	389
11.3.2	Multiple Regression Analysis	393
11.3.3	Overdispersion	402
11.3.4	Stepwise Regression	404
	Exercises	406
<b>12</b>	<b>Methods for Repeatedly Measured Responses</b>	<b>409</b>
12.1	Extending Regression Methods Beyond Independent Data	409
12.2	Continuous Responses	410
12.2.1	Extending Regression using the Linear Mixed Model	410
12.2.2	Testing and Inference	414
12.2.3	Comparing Models	417
12.2.4	Special Cases: Random Block Designs and Multi-level Sampling	418
12.3	Binary Responses	423
12.3.1	Extending Logistic Regression using Generalized Estimating Equations	423
12.3.2	Testing and Inference	425
12.4	Count Responses	427
12.4.1	Extending Poisson Regression using Generalized Estimating Equations	427
12.4.2	Testing and Inference	428
12.5	Computational Notes	431
	Exercises	432
<b>13</b>	<b>Analysis of Survival Data and Data from Matched Studies</b>	<b>439</b>
13.1	Survival Data	440
13.2	Introductory Survival Analyses	443
13.2.1	Kaplan–Meier Curve	444
13.2.2	Comparison of Survival Distributions	446

13.3	Simple Regression and Correlation	450
13.3.1	Model and Approach	451
13.3.2	Measures of Association	452
13.3.3	Tests of Association	455
13.4	Multiple Regression and Correlation	456
13.4.1	Proportional Hazards Model with Several Covariates	456
13.4.2	Testing Hypotheses in Multiple Regression	457
13.4.3	Time-Dependent Covariates and Applications	461
13.5	Pair-Matched Case–Control Studies	464
13.5.1	Model	465
13.5.2	Analysis	466
13.6	Multiple Matching	468
13.6.1	Conditional Approach	469
13.6.2	Estimation of the Odds Ratio	469
13.6.3	Testing for Exposure Effect	470
13.7	Conditional Logistic Regression	472
13.7.1	Simple Regression Analysis	473
13.7.2	Multiple Regression Analysis	478
	Exercises	484
<b>14</b>	<b>Study Designs</b>	<b>493</b>
14.1	Types of Study Designs	494
14.2	Classification of Clinical Trials	495
14.3	Designing Phase I Cancer Trials	497
14.4	Sample Size Determination for Phase II Trials and Surveys	499
14.5	Sample Sizes for Other Phase II Trials	501
14.5.1	Continuous Endpoints	501
14.5.2	Correlation Endpoints	502
14.6	About Simon’s Two-Stage Phase II Design	503
14.7	Phase II Designs for Selection	504
14.7.1	Continuous Endpoints	505
14.7.2	Binary Endpoints	505
14.8	Toxicity Monitoring in Phase II Trials	506
14.9	Sample Size Determination for Phase III Trials	508
14.9.1	Comparison of Two Means	509
14.9.2	Comparison of Two Proportions	511
14.9.3	Survival Time as the Endpoint	513
14.10	Sample Size Determination for Case–Control Studies	515
14.10.1	Unmatched Designs for a Binary Exposure	516
14.10.2	Matched Designs for a Binary Exposure	518
14.10.3	Unmatched Designs for a Continuous Exposure	520
	Exercises	522
	<b>References</b>	<b>529</b>
	<b>Appendices</b>	<b>535</b>
	<b>Answers to Selected Exercises</b>	<b>541</b>
	<b>Index</b>	<b>585</b>

# PREFACE TO THE SECOND EDITION

This second edition of the book adds several new features:

- An expanded treatment of one-way ANOVA including multiple testing procedures;
- A new chapter on two-way, three-way, and higher level ANOVAs, including both fixed, random, and mixed effects ANOVAs;
- A substantially revised chapter on regression;
- A new chapter on models for repeated measurements using linear mixed models and generalized estimating equations;
- Examples worked throughout the book in R in addition to SAS software;
- Additional end of chapter exercises in several chapters.

These features have been added with the help of a new second author. As in the first edition, data sets used in the in-chapter examples and end of chapter exercises are largely based on real studies on which we collaborated. The very large data tables referred to throughout this book are too large for inclusion in the printed text; they are available at [www.wiley.com/go/Le/Biostatistics](http://www.wiley.com/go/Le/Biostatistics).

We thank previous users of the book for feedback on the first edition, which led to many of the improvements in this second edition. We also thank Megan Schlick, Division of Biostatistics at the University of Minnesota, for her assistance with preparation of several files and the index for this edition.

Chap T. Le  
Lynn E. Eberly  
*Minneapolis, MN*  
*September 2015*





# **PREFACE TO THE FIRST EDITION**

A course in introductory biostatistics is often required for professional students in public health, dentistry, nursing, and medicine, and for graduate students in nursing and other biomedical sciences, a requirement that is often considered a roadblock, causing anxiety in many quarters. These feelings are expressed in many ways and in many different settings, but all lead to the same conclusion: that students need help, in the form of a user-friendly and real data-based text, in order to provide enough motivation to learn a subject that is perceived to be difficult and dry. This introductory text is written for professionals and beginning graduate students in human health disciplines who need help to pass and benefit from the basic biostatistics requirement of a one-term course or a full-year sequence of two courses. Our main objective is to avoid the perception that statistics is just a series of formulas that students need to “get over with,” but to present it as a way of thinking – thinking about ways to gather and analyze data so as to benefit from taking the required course. There is no better way to do that than to base a book on real data, so many real data sets in various fields are provided in the form of examples and exercises as aids to learning how to use statistical procedures, still the nuts and bolts of elementary applied statistics.

The first five chapters start slowly in a user-friendly style to nurture interest and motivate learning. Sections called “Brief Notes on the Fundamentals” are added here and there to gradually strengthen the background and the concepts. Then the pace is picked up in the remaining seven chapters to make sure that those who take a full-year sequence of two courses learn enough of the nuts and bolts of the subject. Our basic strategy is that most students would need only one course, which would end at about the middle of Chapter 9, after covering simple linear regression; instructors may add a few sections of Chapter 14. For students who take only one course, other chapters would serve as references to supplement class discussions as well as for

their future needs. A subgroup of students with a stronger background in mathematics would go on to a second course, and with the help of the brief notes on the fundamentals would be able to handle the remaining chapters. A special feature of the book is the sections “Notes on Computations” at the end of most chapters. These notes cover the uses of Microsoft’s Excel, but samples of SAS computer programs are also included at the end of many examples, especially the advanced topics in the last several chapters.

The way of thinking called *statistics* has become important to all professionals, not only those in science or business, but also caring people who want to help to make the world a better place. But what is biostatistics, and what can it do? There are popular definitions and perceptions of statistics. We see “vital statistics” in the newspaper: announcements of life events such as births, marriages, and deaths. Motorists are warned to drive carefully, to avoid “becoming a statistic.” Public use of the word is widely varied, most often indicating lists of numbers, or data. We have also heard people use the word *data* to describe a verbal report, a believable anecdote. For this book, especially in the first few chapters, we do not emphasize statistics as things, but instead, offer an active concept of “doing statistics.” The doing of statistics is a way of thinking about numbers (collection, analysis, presentation), with emphasis on relating their interpretation and meaning to the manner in which they are collected. Formulas are only a part of that thinking, simply tools of the trade; they are needed but not as the only things one needs to know.

To illustrate statistics as a way of thinking, let us begin with a familiar scenario: criminal court procedures. A crime has been discovered and a suspect has been identified. After a police investigation to collect evidence against the suspect, a prosecutor presents summarized evidence to a jury. The jurors are given the rules regarding convicting beyond a reasonable doubt and about a unanimous decision, and then they debate. After the debate, the jurors vote and a verdict is reached: guilty or not guilty. Why do we need to have this time-consuming, cost-consuming process of trial by jury? One reason is that the truth is often unknown, at least uncertain. Perhaps only the suspect knows but he or she does not talk. It is uncertain because of variability (every case is different) and because of possibly incomplete information. Trial by jury is the way our society deals with uncertainties; its goal is to minimize mistakes.

How does society deal with uncertainties? We go through a process called *trial by jury*, consisting of these steps: (1) we form an assumption or hypothesis (that every person is innocent until proved guilty), (2) we gather data (evidence against the suspect), and (3) we decide whether the hypothesis should be rejected (guilty) or should not be rejected (not guilty). With such a well-established procedure, sometimes we do well, sometimes we do not. Basically, a successful trial should consist of these elements: (1) a probable cause (with a crime and a suspect), (2) a thorough investigation by police, (3) an efficient presentation by a prosecutor, and (4) a fair and impartial jury.

In the context of a trial by jury, let us consider a few specific examples: (1) the *crime* is lung cancer and the *suspect* is cigarette smoking, or (2) the *crime* is leukemia and the *suspect* is pesticides, or (3) the *crime* is breast cancer and the *suspect* is a defective gene. The process is now called *research* and the tool to carry out that research is biostatistics. In a simple way, biostatistics serves as the biomedical

version of the trial by jury process. It is the *science of dealing with uncertainties using incomplete information*. Yes, even science is uncertain; scientists arrive at different conclusions in many different areas at different times; many studies are inconclusive (hung jury). The reasons for uncertainties remain the same. Nature is complex and full of unexplained biological variability. But most important, we always have to deal with incomplete information. It is often not practical to study an entire population; we have to rely on information gained from a *sample*.

How does science deal with uncertainties? We learn how society deals with uncertainties; we go through a process called *biostatistics*, consisting of these steps: (1) we form an assumption or hypothesis (from the research question), (2) we gather data (from clinical trials, surveys, medical record abstractions), and (3) we make decision(s) (by doing statistical analysis/inference; a guilty verdict is referred to as *statistical significance*). Basically, a successful research should consist of these elements: (1) a good research question (with well-defined objectives and endpoints), (2) a thorough investigation (by experiments or surveys), (3) an efficient presentation of data (organizing data, summarizing, and presenting data: an area called *descriptive statistics*), and (4) proper statistical inference. This book is a problem-based introduction to the last three elements; together they form a field called *biostatistics*. The coverage is rather brief on data collection but very extensive on descriptive statistics (Chapters 1, 2), especially on methods of statistical inference (Chapters 4–12). Chapter 3, on probability and probability models, serves as the link between the descriptive and inferential parts. Notes on computations and samples of SAS computer programs are incorporated throughout the book. About 60% of the material in the first eight chapters overlaps with chapters from *Health and Numbers: A Problems-Based Introduction to Biostatistics* (another book by Wiley), but new topics have been added and others rewritten at a somewhat higher level. In general, compared to *Health and Numbers*, this book is aimed at a different audience – those who need a whole year of statistics and who are more mathematically prepared for advanced algebra and precalculus subjects.

I would like to express my sincere appreciation to colleagues, teaching assistants, and many generations of students for their help and feedback. I have learned very much from my former students, I hope that some of what they have taught me is reflected well in many sections of this book. Finally, my family bore patiently the pressures caused by my long-term commitment to the book; to my wife and daughters, I am always most grateful.

Chap T. Le  
Edina, Minnesota



# ABOUT THE COMPANION WEBSITE

This book is accompanied by a companion website:

**[www.wiley.com/go/Le/Biostatistics](http://www.wiley.com/go/Le/Biostatistics)**

The website includes:

- Electronic copy of the larger data sets used in Examples and Exercises



---

# 1

---

## DESCRIPTIVE METHODS FOR CATEGORICAL DATA

Most introductory textbooks in statistics and biostatistics start with methods for summarizing and presenting continuous data. We have decided, however, to adopt a different starting point because our focused areas are in the biomedical sciences, and health decisions are frequently based on proportions, ratios, or rates. In this first chapter we will see how these concepts appeal to common sense, and learn their meaning and uses.

### 1.1 PROPORTIONS

Many outcomes can be classified as belonging to one of two possible categories: presence and absence, nonwhite and white, male and female, improved and nonimproved. Of course, one of these two categories is usually identified as of primary interest: for example, presence in the presence and absence classification, nonwhite in the white and nonwhite classification. We can, in general, relabel the two outcome categories as positive (+) and negative (−). An outcome is *positive* if the primary category is observed and is *negative* if the other category is observed.

It is obvious that, in the summary to characterize observations made on a group of people, the number  $x$  of positive outcomes is not sufficient; the group size  $n$ , or total number of observations, should also be recorded. The number  $x$  tells us very little and becomes meaningful only after adjusting for the size  $n$  of the group; in other words, the two figures  $x$  and  $n$  are often combined into a *statistic*, called a *proportion*:

$$p = \frac{x}{n}.$$

The term *statistic* means a summarized quantity from observed data. Clearly,  $0 \leq p \leq 1$ . This proportion  $p$  is sometimes expressed as a percentage and is calculated as follows:

$$\text{percentage}(\%) = \frac{x}{n}(100).$$

### Example 1.1

A study published by the Urban Coalition of Minneapolis and the University of Minnesota Adolescent Health Program surveyed 12915 students in grades 7–12 in Minneapolis and St. Paul public schools. The report stated that minority students, about one-third of the group, were much less likely to have had a recent routine physical checkup. Among Asian students, 25.4% said that they had not seen a doctor or a dentist in the last two years, followed by 17.7% of Native Americans, 16.1% of blacks, and 10% of Hispanics. Among whites, it was 6.5%.

*Proportion* is a number used to describe a group of people according to a *dichotomous*, or *binary*, *characteristic* under investigation. It is noted that characteristics with multiple categories can have a proportion calculated per category, or can be dichotomized by pooling some categories to form a new one, and the concept of proportion applies. The following are a few illustrations of the use of proportions in the health sciences.

#### 1.1.1 Comparative Studies

Comparative studies are intended to show possible differences between two or more groups; Example 1.1 is such a typical comparative study. The survey cited in Example 1.1 also provided the following figures concerning boys in the group who use tobacco at least weekly. Among Asians, it was 9.7%, followed by 11.6% of blacks, 20.6% of Hispanics, 25.4% of whites, and 38.3% of Native Americans.

In addition to surveys that are cross-sectional, as seen in Example 1.1, data for comparative studies may come from different sources; the two fundamental designs being retrospective and prospective. *Retrospective studies* gather past data from selected cases and controls to determine differences, if any, in *exposure* to a suspected *risk factor*. These are commonly referred to as *case-control studies*; each such study is focused on a particular disease. In a typical case-control study, cases of a specific disease are ascertained as they arise from population-based registers or lists of hospital admissions, and controls are sampled either as disease-free persons from the population at risk or as hospitalized patients having a diagnosis other than the one under study. The advantages of a retrospective study are that it is economical and provides answers to research questions relatively quickly because the cases are already available. Major limitations are due to the inaccuracy of the exposure histories and uncertainty about the appropriateness of the control sample; these problems sometimes hinder retrospective studies and make them less preferred than prospective studies. The following is an example of a retrospective study in the field of occupational health.



**Example 1.2**

A case–control study was undertaken to identify reasons for the exceptionally high rate of lung cancer among male residents of coastal Georgia. Cases were identified from these sources:

1. Diagnoses since 1970 at the single large hospital in Brunswick;
2. Diagnoses during 1975–1976 at three major hospitals in Savannah;
3. Death certificates for the period 1970–1974 in the area.

Controls were selected from admissions to the four hospitals and from death certificates in the same period for diagnoses other than lung cancer, bladder cancer, or chronic lung cancer. Data are tabulated separately for smokers and nonsmokers in Table 1.1. The exposure under investigation, “shipbuilding,” refers to employment in shipyards during World War II. By using a separate tabulation, with the first half of the table for nonsmokers and the second half for smokers, we treat *smoking* as a potential confounder. A *confounder* is a factor, an exposure by itself, not under investigation but related to the disease (in this case, lung cancer) and the exposure (shipbuilding); previous studies have linked smoking to lung cancer, and construction workers are more likely to be smokers. The term *exposure* is used here to emphasize that employment in shipyards is a suspected *risk* factor; however, the term is also used in studies where the factor under investigation has beneficial effects.

In an examination of the smokers in the data set in Example 1.2, the numbers of people employed in shipyards, 84 and 45, tell us little because the sizes of the two groups, cases and controls, are different. Adjusting these absolute numbers for the group sizes (397 cases and 315 controls), we have:

1. For the smoking controls,

$$\begin{aligned}\text{proportion with exposure} &= \frac{45}{315} \\ &= 0.143 \text{ or } 14.3\%.\end{aligned}$$

2. For the smoking cases,

$$\begin{aligned}\text{proportion with exposure} &= \frac{84}{397} \\ &= 0.212 \text{ or } 21.2\%.\end{aligned}$$

**TABLE 1.1**

Smoking	Shipbuilding	Cases	Controls
No	Yes	11	35
	No	50	203
Yes	Yes	84	45
	No	313	270

The results reveal different exposure histories: the proportion in shipbuilding among cases was higher than that among controls. It is *not* in any way conclusive proof, but it is a good *clue*, indicating a possible *relationship* between the disease (lung cancer) and the exposure (shipbuilding).

Similar examination of the data for nonsmokers shows that, by taking into consideration the numbers of cases and controls, we have the following figures for shipbuilding employment:

1. For the non-smoking controls,

$$\begin{aligned}\text{proportion with exposure} &= \frac{35}{238} \\ &= 0.147 \quad \text{or} \quad 14.7\%.\end{aligned}$$

2. For the non-smoking cases,

$$\begin{aligned}\text{proportion with exposure} &= \frac{11}{61} \\ &= 0.180 \quad \text{or} \quad 18.0\%.\end{aligned}$$

The results for non-smokers also reveal different exposure histories: the proportion in shipbuilding among cases was again higher than that among controls.

The analyses above also show that the case-control difference in the proportions with the exposure among smokers, that is,

$$21.2 - 14.3 = 6.9\%,$$

is different from the case-control difference in the proportions with the exposure among nonsmokers, which is:

$$18.0 - 14.7 = 3.3\%.$$

The differences, 6.9% and 3.3%, are *measures* of the strength of the relationship between the disease and the exposure, one for each of the two strata: the two groups of smokers and nonsmokers, respectively. The calculation above shows that the possible effects of employment in shipyards (as a suspected risk factor) are different for smokers and nonsmokers. This difference of differences, if confirmed, is called a *three-term interaction* or *effect modification*, where smoking alters the effect of employment in shipyards as a risk for lung cancer. In that case, *smoking* is not only a confounder, it is an *effect modifier*, which modifies the effects of shipbuilding (on the possibility of having lung cancer).

Another illustration is provided in the following example concerning glaucomatous blindness.

TABLE 1.2

	Population	Cases	Cases per 100 000
White	32 930 233	2832	8.6
Nonwhite	3 933 333	3227	82.0

Example 1.3

Counts of persons registered blind from glaucoma are listed in Table 1.2.

For these *disease registry data*, direct calculation of a proportion results in a very tiny fraction, that is, the number of cases of the disease per person at risk. For convenience, in Table 1.2, this is multiplied by 100 000, and hence the result expresses the number of cases per 100 000 people. This data set also provides an example of the use of proportions as disease *prevalence*, which is defined as:

$$\text{prevalence} = \frac{\text{number of diseased persons at the time of investigation}}{\text{total number of persons examined}}.$$

*Disease prevalence* and related concepts are discussed in more detail in Section 1.2.2.

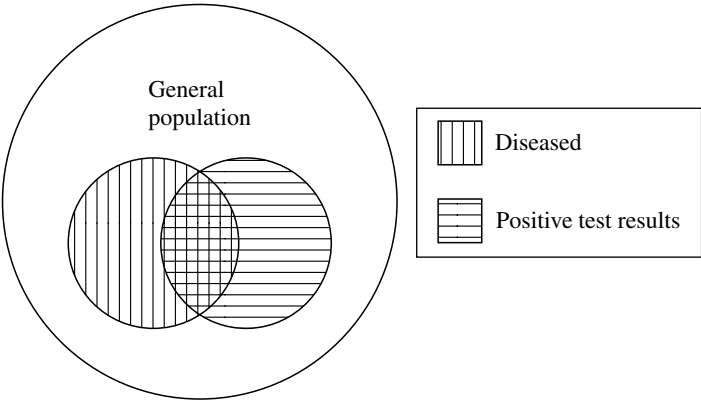
For blindness from glaucoma, calculations in Example 1.3 reveal a striking difference between the races: The blindness prevalence among nonwhites was over eight times that among whites. The number “100 000” was selected arbitrarily; any power of 10 would be suitable so as to obtain a result between 1 and 100, sometimes between 1 and 1000; it is easier to state the result “82 cases per 100 000” than to say that the prevalence is 0.00082.

1.1.2 Screening Tests

Other uses of proportions can be found in the evaluation of *screening tests* or *diagnostic procedures*. Following these procedures, using clinical observations or laboratory techniques, people are classified as healthy or as falling into one of a number of disease categories. Such tests are important in medicine and epidemiologic studies and may form the basis of early interventions. Almost all such tests are imperfect, in the sense that healthy persons will occasionally be classified wrongly as being ill, while some people who are really ill may fail to be detected. That is, *misclassification* is unavoidable. Suppose that each person in a large population can be classified as truly positive or negative for a particular disease; this true diagnosis may be based on more refined methods than are used in the test, or it may be based on evidence that emerges after the passage of time (e.g., at autopsy). For each class of people, diseased and healthy, the test is applied, with the results depicted in Figure 1.1.

The two proportions fundamental to evaluating diagnostic procedures are sensitivity and specificity. *Sensitivity* is the proportion of diseased people detected as positive by the test:

$$\text{sensitivity} = \frac{\text{number of diseased persons who test positive}}{\text{total number of diseased persons}}.$$



**FIGURE 1.1** Graphical display of a screening test.

The corresponding errors are *false negatives*. *Specificity* is the proportion of healthy people detected as negative by the test:

$$\text{specificity} = \frac{\text{number of healthy persons who test negative}}{\text{total number of healthy persons}}.$$

The corresponding errors are *false positives*.

Clearly, it is desirable that a test or screening procedure be highly sensitive and highly specific. However, the two types of errors go in opposite directions; for example, an effort to increase sensitivity may lead to more false positives, and vice versa.

**Example 1.4**

A cytological test was undertaken to screen women for cervical cancer. Consider a group of 24 103 women consisting of 379 women whose cervixes are abnormal (to an extent sufficient to justify concern with respect to possible cancer) and 23 724 women whose cervixes are acceptably healthy. A test was applied and results are tabulated in Table 1.3. (This study was performed with a rather old test and is used here only for illustration.)

**TABLE 1.3**

True	Test		Total
	–	+	
–	23 362	362	23 724
+	225	154	379

The calculations

$$\begin{aligned}\text{sensitivity} &= \frac{154}{379} \\ &= 0.406 \text{ or } 40.6\% \\ \text{specificity} &= \frac{23\,362}{23\,724} \\ &= 0.985 \text{ or } 98.5\%\end{aligned}$$

show that the test is highly specific (98.5%) but not very sensitive (40.6%); among the 379 women with the disease, more than half (59.4%) had false negatives. The implications of the use of this test are:

1. If a woman without cervical cancer is tested, the result would almost surely be negative, *but*
2. If a woman with cervical cancer is tested, the chance is that the disease would go undetected because 59.4% of these cases would result in false negatives.

Finally, it is important to note that throughout this section, proportions have been defined so that both the numerator and the denominator are counts or frequencies, and the numerator corresponds to a subgroup of the larger group involved in the denominator, resulting in a number between 0 and 1 (or between 0 and 100%). It is straightforward to generalize this concept for use with characteristics having more than two outcome categories; for each category we can define a proportion, and these category-specific proportions add up to 1 (or 100%).

### Example 1.5

An examination of the 668 children reported living in crack/cocaine households shows 70% blacks, followed by 18% whites, 8% Native Americans, and 4% other or unknown.

#### 1.1.3 Displaying Proportions

Perhaps the most effective and most convenient way of presenting data, especially discrete data, is through the use of graphs. Graphs convey the information, the general patterns in a set of data, at a single glance. Therefore, graphs are often easier to read than tables; the most informative graphs are simple and self-explanatory. Of course, to achieve that objective, graphs should be constructed carefully. Like tables, they should be clearly labeled and units of measurement and/or magnitude of quantities should be included. Remember that graphs must tell their own story; they should be complete in themselves and require little or no additional explanation.

**Bar Charts** Bar charts are a very popular type of graph used to display several proportions for quick comparison. In applications suitable for bar charts, there are several groups and we investigate one binary characteristic. In a bar chart, the various

groups are represented along the horizontal axis; they may be arranged alphabetically, by the size of their proportions, or on some other rational basis. A vertical bar is drawn above each group such that the height of the bar is the proportion associated with that group. The bars should be of equal width and should be separated from one another so as not to imply continuity.

**Example 1.6**

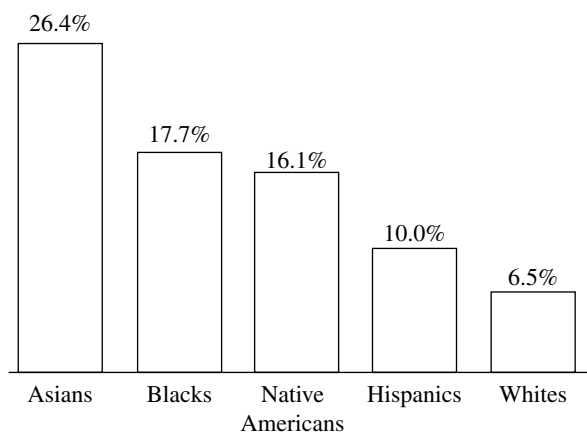
We can present the data set on children without a recent physical checkup (Example 1.1) by a bar chart, as shown in Figure 1.2.

**Pie Charts** Pie charts are another popular type of graph. In applications suitable for pie charts, there is only one group but we want to decompose it into several categories. A pie chart consists of a circle; the circle is divided into wedges that correspond to the magnitude of the proportions for various categories. A pie chart shows the differences between the sizes of various categories or subgroups as a decomposition of the total. It is suitable, for example, for use in presenting a budget, where we can easily see the difference between United States expenditures on health care and defense. In other words, a bar chart is a suitable graphic device when we have several groups, each associated with a different proportion; whereas a pie chart is more suitable when we have one group that is divided into several categories. The proportions of various categories in a pie chart should add up to 100%. Like bar charts, the categories in a pie chart are usually arranged by the size of the proportions. They may also be arranged alphabetically or on some other rational basis.

**Example 1.7**

We can present the data set on children living in crack households (Example 1.5) by a pie chart as shown in Figure 1.3.

Another example of the pie chart’s use is for presenting the proportions of deaths due to different causes.



**FIGURE 1.2** Children without a recent physical checkup.