

# Statistical Inference

*John W. Tukey*

**George Casella**

*Cornell University*

**Roger L. Berger**

*North Carolina State University*



**Duxbury Press**

*An Imprint of Wadsworth Publishing Company*

Belmont, California

**Duxbury Press**  
*An Imprint of Wadsworth Publishing Company*  
A division of Wadsworth, Inc.

© 1990 by Wadsworth, Inc., Belmont, California 94002.  
All rights reserved. No part of this book may be reproduced,  
stored in a retrieval system, or transcribed, in any form or by  
any means—electronic, mechanical, photocopying, recording,  
or otherwise—without the prior written permission of the  
publisher, Wadsworth Publishing Company, Belmont,  
California 94002.

Printed in the United States of America

10 9 8

Library of Congress Cataloging in Publication Data

Casella, George.

Statistical inference / George Casella, Roger L. Berger.

p. cm.

Includes bibliographical references.

ISBN 0-534-11958-1

1. Mathematical statistics. 2. Probabilities. I. Berger, Roger  
L. II. Title.

QA276.C37 1990

519.5—dc20

89-27287

CIP

Sponsoring Editor: John Kimmel  
Marketing Representative: Charlie Delmar  
Editorial Assistant: Jenny Greenwood  
Production Coordinator: Joan Marsh  
Interior Design: Vernon Boes  
Cover Design: Lisa Thompson  
Interior Illustration: Lori Heckelman  
Typesetting: Electronic Technical Publishing  
Cover Printing: Phoenix Color Corporation  
Printing and Binding: Arcata Graphics/Fairfield

*To Anne and Vicki*



# Preface

When someone discovers that you are writing a textbook, one (or both) of two questions will be asked. The first is “Why are you writing a book?” and the second is “How is your book different from what’s out there?” The first question is fairly easy to answer. You are writing a book because you are not entirely satisfied with the available texts. The second question is harder to answer. The answer can’t be put in a few sentences so, in order not to bore your audience (who may be asking the question only out of politeness), you try to say something quick and witty. It usually doesn’t work.

The purpose of this book is to build theoretical statistics (as different from mathematical statistics) from the first principles of probability theory. Logical development, proofs, ideas, themes, etc., evolve through statistical arguments. Thus, starting from the basics of probability, we develop the theory of statistical inference using techniques, definitions, and concepts that are statistical and are natural extensions and consequences of previous concepts. When this endeavor was started, we were not sure how well it would work. The final judgment of our success is, of course, left to the reader.

The book is intended for first-year graduate students majoring in statistics or in a field where a statistics concentration is desirable. The prerequisite is one year of calculus. (Some familiarity with matrix manipulations would be useful, but is not essential.) The book can be used for a two-semester, or three-quarter, introductory course in statistics.

The first four chapters cover basics of probability theory and introduce many fundamentals that are later necessary. Chapters 5 and 6 are the first statistical chapters. Chapter 5 is transitional (between probability and statistics) and can be the starting point for a course in statistical theory for students with some probability background. Chapter 6 is somewhat unique, detailing three statistical principles (sufficiency, likelihood, and invariance) and showing how these principles are important in modeling data. Not all instructors will cover this chapter in detail, although we strongly recommend spending some time here. In particular, the likelihood and invariance principles are treated in detail. Along with the sufficiency principle, these principles, and the thinking behind them, are fundamental to total statistical understanding.

Chapters 7–9 represent the central core of statistical inference, estimation (point and interval) and hypothesis testing. A major feature of these chapters is the division into methods of *finding* appropriate statistical techniques and methods of *evaluating* these techniques. Finding and evaluating are of interest to both the theorist and the practitioner, but we feel that it is important to separate these endeavors. Different concerns are important, and different rules are invoked. Of further interest may be the sections of these chapters titled Other Considerations. Here, we indicate how the rules of statistical inference may be relaxed (as is done every day) and still produce meaningful inferences. Many of the techniques covered in these sections are ones that are used in consulting and are helpful in analyzing and inferring from actual problems.

The final three chapters can be thought of as special topics, although we feel that some familiarity with the material is important in anyone's statistical education. Chapter 10 is a thorough introduction to decision theory and contains the most modern material we could include. Chapter 11 deals with the analysis of variance (oneway and randomized block), building the theory of the complete analysis from the more simple theory of treatment contrasts. Our experience has been that experimenters are most interested in inferences from contrasts, and using principles developed earlier, most tests and intervals can be derived from contrasts. Finally, Chapter 12 treats the theory of regression, dealing first with simple linear regression and then covering regression with "errors in variables." This latter topic is quite important, not only to show its own usefulness and inherent difficulties, but also to illustrate the limitations of inferences from ordinary regression.

As more concrete guidelines for basing a one-year course on this book, we offer the following suggestions. There can be two distinct types of courses taught from this book. One kind we might label "more mathematical," being a course appropriate for students majoring in statistics and having a solid mathematics background (at least  $1\frac{1}{2}$  years of calculus, some matrix algebra, and perhaps a real analysis course). For such students we recommend covering Chapters 1–9 in their entirety (which should take approximately 22 weeks) and spend the remaining time customizing the course with selected topics from Chapters 10–12. Once the first nine chapters are covered, the material in each of the last three chapters is self-contained, and can be covered in any order.

Another type of course is "more practical." Such a course may also be a first course for mathematically sophisticated students, but is aimed at students with one year of calculus who may not be majoring in statistics. It stresses the more practical uses of statistical theory, being more concerned with understanding basic statistical concepts and deriving reasonable statistical procedures for a variety of situations, and less concerned with formal optimality investigations. Such a course will necessarily omit a certain amount of material, but the following list of sections can be covered in a one-year course:

Chapter	Sections
1	All
2	2.1, 2.2, 2.3
3	3.1, 3.2

4	4.1, 4.2, 4.3, 4.5
5	5.1, 5.2, 5.3.1, 5.4
6	6.1.1, 6.2.1
7	7.1, 7.2.1, 7.2.2, 7.2.3, 7.3.1, 7.3.3, 7.4
8	8.1, 8.2.1, 8.2.3, 8.2.4, 8.3.1, 8.3.2, 8.4
9	9.1, 9.2.1, 9.2.2, 9.2.4, 9.3.1, 9.4
11	11.1, 11.2
12	12.1, 12.2

If time permits, there can be some discussion (with little emphasis on details) of the material in Sections 4.4, 5.5, and 6.1.2, 6.1.3, 6.1.4. The material in Sections 11.3 and 12.3 may also be considered.

The exercises have been gathered from many sources and are quite plentiful. We feel that, perhaps, the only way to master this material is through practice, and thus we have included much opportunity to do so. The exercises are as varied as we could make them, and many of them illustrate points that are either new or complementary to the material in the text. Some exercises are even taken from research papers. (It makes you feel old when you can include exercises based on papers that were new research during your own student days!) Although the exercises are not subdivided like the chapters, their ordering roughly follows that of the chapter. (Subdivisions often give too many hints.) Furthermore, the exercises become (again, roughly) more challenging as their numbers become higher.

As this is an introductory book with a relatively broad scope, the topics are not covered in great depth. However, we felt some obligation to guide the reader one step further in the topics that may be of interest. Thus, we have included many references, pointing to the path to deeper understanding of any particular topic. (*The Encyclopedia of Statistical Sciences*, edited by Kotz, Johnson, and Read, provides a fine introduction to many topics.)

To write this book, we have drawn on both our past teachings and current work. We have also drawn on many people, to whom we are extremely grateful. We thank our colleagues at Cornell, North Carolina State, and Purdue—in particular, Jim Berger, Larry Brown, Sir David Cox, Ziding Feng, Janet Johnson, Leon Gleser, Costas Goutis, Dave Lansky, George McCabe, Chuck McCulloch, Myra Samuels, Steve Schwager, and Shayle Searle, who have given their time and expertise in reading parts of this manuscript, offered assistance, and taken part in many conversations leading to constructive suggestions. We also thank Shanti Gupta for his hospitality, and the library at Purdue, which was essential. We are grateful for the detailed reading and helpful suggestions of Shayle Searle and of our reviewers, both anonymous and non-anonymous (Jim Albert, Dan Coster, and Tom Wehrly). We also thank David Moore and George McCabe for allowing us to use their tables, and Steve Hirdt for supplying us with data. Since this book was written by two people who, for most of the time, were at least 600 miles apart, we lastly thank Bitnet for making this entire thing possible.

George Casella  
Roger L. Berger



*"I can see nothing," said I, handing it back to my friend.  
"On the contrary, Watson, you can see everything. You fail,  
however, to reason from what you see. You are too timid  
in drawing your inferences."*

**Dr. Watson and Sherlock Holmes**  
*The Adventure of the Blue Carbuncle*



# Contents

## 1 Probability Theory 1

- 1.1 Set Theory 1
- 1.2 Probability Theory 5
  - 1.2.1 Axiomatic Foundations 6
  - 1.2.2 The Calculus of Probabilities 9
  - 1.2.3 Counting 13
  - 1.2.4 Equally Likely Outcomes 16
- 1.3 Conditional Probability and Independence 18
- 1.4 Random Variables 26
- 1.5 Distribution Functions 29
- 1.6 Density and Mass Functions 34
  - Exercises* 37

## 2 Transformations and Expectations 45

- 2.1 Distributions of Functions of a Random Variable 45
- 2.2 Expected Values 54
- 2.3 Moments and Moment Generating Functions 58
- 2.4 Differentiating Under an Integral Sign 68
  - Exercises* 76
  - Miscellanea* 82

<b>3 Common Families of Distributions</b>	<b>85</b>
3.1 Discrete Distributions	85
3.2 Continuous Distributions	99
3.3 Exponential Families	112
3.4 Location and Scale Families	115
<i>Exercises</i>	121
<i>Miscellanea</i>	126
<b>4 Multiple Random Variables</b>	<b>128</b>
4.1 Joint and Marginal Distributions	128
4.2 Conditional Distributions and Independence	137
4.3 Bivariate Transformations	146
4.4 Hierarchical Models and Mixture Distributions	153
4.5 Covariance and Correlation	160
4.6 Multivariate Distributions	168
4.7 Inequalities and Identities	178
4.7.1 Numerical Inequalities	178
4.7.2 Functional Inequalities	181
4.7.3 Probability Inequalities	184
4.7.4 Identities	186
<i>Exercises</i>	190
<i>Miscellanea</i>	199
<b>5 Properties of a Random Sample</b>	<b>201</b>
5.1 Basic Concepts of Random Samples	201
5.2 Sums of Random Variables from a Random Sample	205
5.3 Convergence Concepts	213
5.3.1 Convergence in Probability	213

5.3.2 Almost Sure Convergence	214
5.3.3 Convergence in Distribution	216
<b>5.4 Sampling from the Normal Distribution</b>	<b>220</b>
5.4.1 Properties of the Sample Mean and Variance	220
5.4.2 The Derived Distributions: Student's <i>t</i> and Snedecor's <i>F</i>	225
<b>5.5 Order Statistics</b>	<b>228</b>
<i>Exercises</i>	236
<i>Miscellanea</i>	243

## **6 Principles of Data Reduction** **246**

<b>6.1 The Sufficiency Principle</b>	<b>247</b>
6.1.1 Sufficient Statistics	247
6.1.2 Minimal Sufficient Statistics	254
6.1.3 Ancillary Statistics	257
6.1.4 Sufficient, Ancillary, and Complete Statistics	259
<b>6.2 The Likelihood Principle</b>	<b>264</b>
6.2.1 The Likelihood Function	265
6.2.2 The Formal Likelihood Principle	267
<b>6.3 The Invariance Principle</b>	<b>273</b>
<i>Exercises</i>	280
<i>Miscellanea</i>	282

## **7 Point Estimation** **284**

<b>7.1 Introduction</b>	<b>284</b>
<b>7.2 Methods of Finding Estimators</b>	<b>285</b>
7.2.1 Method of Moments	285
7.2.2 Maximum Likelihood Estimators	289
7.2.3 Bayes Estimators	297
7.2.4 Invariant Estimators	300
<b>7.3 Methods of Evaluating Estimators</b>	<b>303</b>
7.3.1 Mean Squared Error	303
7.3.2 Best Unbiased Estimators	307
7.3.3 Sufficiency and Unbiasedness	316
7.3.4 Consistency	322
<b>7.4 Other Considerations</b>	<b>325</b>

7.4.1 Asymptotic Variance of Maximum Likelihood Estimators	325
7.4.2 Taylor Series Approximations	328
<i>Exercises</i>	331
<i>Miscellanea</i>	342

## 8 Hypothesis Testing

345

8.1 Introduction	345
8.2 Methods of Finding Tests	346
8.2.1 Likelihood Ratio Tests	346
8.2.2 Invariant Tests	351
8.2.3 Bayesian Tests	354
8.2.4 Union–Intersection and Intersection–Union Tests	356
8.3 Methods of Evaluating Tests	358
8.3.1 Error Probabilities and the Power Function	358
8.3.2 Most Powerful Tests	365
8.3.3 Unbiased and Invariant Tests	370
8.3.4 Locally Most Powerful Tests	376
8.3.5 Sizes of Union–Intersection and Intersection–Union Tests	378
8.4 Other Considerations	381
8.4.1 Asymptotic Distribution of LRTs	381
8.4.2 Other Large-Sample Tests	383
<i>Exercises</i>	385
<i>Miscellanea</i>	400

## 9 Interval Estimation

403

9.1 Introduction	403
9.2 Methods of Finding Interval Estimators	406
9.2.1 Inverting a Test Statistic	406
9.2.2 Pivotal Quantities	413
9.2.3 Guaranteeing an Interval	416
9.2.4 Bayesian Intervals	422
9.2.5 Invariant Intervals	426
9.3 Methods of Evaluating Interval Estimators	429
9.3.1 Size and Coverage Probability	429
9.3.2 Test-Related Optimality	433
9.3.3 Invariant Optimality	437

9.4 Other Considerations	440
9.4.1 Approximate Maximum Likelihood Intervals	441
9.4.2 Other Approximate Intervals	443
<i>Exercises</i>	446
<i>Miscellanea</i>	458

## **10 Decision Theory** 461

10.1 Introduction	461
10.2 Common Decision Theoretic Analyses	464
10.2.1 Point Estimation	464
10.2.2 Hypothesis Testing	467
10.2.3 Interval Estimation	470
10.3 Decision Theoretic Bayes Rules	472
10.3.1 Bayesian Decision Problems	473
10.3.2 Finding Bayes Rules	474
10.4 Admissibility of Decision Rules	479
10.4.1 Comparing Decision Rules	479
10.4.2 Finding Admissible Rules and Complete Classes	481
10.4.3 Admissibility of the Sample Mean Under Normality	485
10.5 Minimax Rules	487
10.6 Invariant Decision Problems	492
10.7 Stein's Paradox	495
<i>Exercises</i>	500
<i>Miscellanea</i>	507

## **11 The Analysis of Variance** 509

11.1 Introduction	509
11.2 The Oneway Analysis of Variance	509
11.2.1 Model and Distribution Assumptions	511
11.2.2 The Classic ANOVA Hypothesis	512
11.2.3 Inferences Regarding Linear Combinations of Means	515
11.2.4 The ANOVA <i>F</i> Test	518
11.2.5 Simultaneous Estimation of Contrasts	522
11.2.6 Partitioning Sums of Squares	525

11.3 Randomized Complete Block Designs 528

11.3.1 Model and Distribution Assumptions 530

11.3.2 Treatment Contrasts 532

11.3.3 Simultaneous Estimation and Testing 537

11.3.4 Partitioning Sums of Squares 539

11.3.5 Implications of Random Blocking 541

*Exercises* 543

*Miscellanea* 551

## 12 Linear Regression 554

12.1 Introduction 554

12.2 Simple Linear Regression 557

12.2.1 Least Squares: A Mathematical Solution 557

12.2.2 Best Linear Unbiased Estimation: A Statistical Solution 560

12.2.3 Models and Distribution Assumptions 564

12.2.4 Estimation and Testing with Normal Errors 567

12.2.5 Estimation and Prediction at a Specified  $x = x_0$  574

12.2.6 Simultaneous Estimation and Confidence Bands 577

12.3 Regression with Errors in Variables 581

12.3.1 Functional and Structural Relationships 583

12.3.2 A Least Squares Solution 584

12.3.3 Maximum Likelihood Estimation 586

12.3.4 Confidence Sets 592

*Exercises* 595

*Miscellanea* 603

## Tables 607

1 Normal 608

2 Student's  $t$  609

3 Chi Squared 610

4 Snedecor's  $F$  612

Distributions 624

1 Catalog 624

2 Relationships 630

References 631

Author Index 641

Subject Index 643

# 1 Probability Theory

*“Take time to consider. The smallest point may be the most essential.”*

**Sherlock Holmes**

*The Adventure of the Red Circle*

The subject of probability theory is the foundation upon which all of statistics is built, providing a means for modeling populations, experiments, or almost anything else that could be considered a random phenomenon. Through these models, statisticians are able to draw inferences about populations, inferences based on examination of only a part of the whole.

The theory of probability has a long and rich history, dating back at least to the seventeenth century when, at the request of their friend, the Chevalier de Meré, Pascal and Fermat developed a mathematical formulation of gambling odds.

The aim of this chapter is not to give a thorough introduction to probability theory; such an attempt would be foolhardy in so short a space. Rather, we attempt to outline some of the basic ideas of probability theory that are fundamental to the study of statistics.

Just as statistics builds upon the foundation of probability theory, probability theory in turn builds upon set theory, which is where we begin.

## 1.1 Set Theory

One of the main objectives of a statistician is to draw conclusions about a population of objects by conducting an experiment. The first step in this endeavor is to identify the possible outcomes or, in statistical terminology, the sample space.

**DEFINITION 1.1.1:** The set,  $S$ , of all possible outcomes of a particular experiment is called the *sample space* for the experiment.

If the experiment consists of tossing a coin, the sample space contains two outcomes, heads and tails; thus,

$$S = \{H, T\}.$$

If, on the other hand, the experiment consists of observing the reported SAT scores of randomly selected students at a certain university, the sample space would be

the set of positive integers between 200 and 800 that are multiples of ten—that is,  $S = \{200, 210, 220, \dots, 780, 790, 800\}$ . Finally, consider an experiment where the observation is reaction time to a certain stimulus. Here, the sample space would consist of all positive numbers, that is,  $S = (0, \infty)$ .

We can classify sample spaces into two types according to the number of elements they contain. Sample spaces can be either countable or uncountable; if the elements of a sample space can be put into 1–1 correspondence with a subset of the integers, the sample space is countable. Of course, if the sample space contains only a finite number of elements, it is countable. Thus, the coin-toss and SAT score sample spaces are both countable (in fact, finite), whereas the reaction time sample space is uncountable, since the positive real numbers cannot be put into 1–1 correspondence with the integers. If, however, we measured reaction time to the nearest second, then the sample space would be (in seconds)  $S = \{0, 1, 2, 3, \dots\}$ , which is then countable.

This distinction between countable and uncountable sample spaces is important only in that it dictates the way in which probabilities can be assigned. For the most part, this causes no problems, although the mathematical treatment of the situations is different. On a philosophical level, it might be argued that there can only be countable sample spaces, since measurements cannot be made with infinite accuracy. (A sample space consisting of, say, all ten-digit numbers is a countable sample space.) While in practice this is true, probabilistic and statistical methods associated with uncountable sample spaces are, in general, less cumbersome than those for countable sample spaces, and provide a close approximation to the true (countable) situation.

Once the sample space has been defined, we are in a position to consider collections of possible outcomes of an experiment.

**DEFINITION 1.1.2:** An *event* is any collection of possible outcomes of an experiment, that is, any subset of  $S$  (including  $S$  itself).

Let  $A$  be an event, a subset of  $S$ . We say the event  $A$  occurs if the outcome of the experiment is in the set  $A$ . When speaking of probabilities, we generally speak of the probability of an event, rather than a set. But we may use the terms interchangeably.

We first need to define formally the following two relationships, which allow us to order and equate sets:

$$A \subset B \Leftrightarrow x \in A \Rightarrow x \in B; \quad (\text{containment})$$

$$A = B \Leftrightarrow A \subset B \text{ and } B \subset A. \quad (\text{equality})$$

Given any two events (or sets)  $A$  and  $B$ , we have the following elementary set operations:

**Union:** The union of  $A$  and  $B$ , written  $A \cup B$ , is the set of elements that belong to either  $A$  or  $B$  or both:

$$A \cup B = \{x: x \in A \text{ or } x \in B\}.$$

*Intersection:* The intersection of  $A$  and  $B$ , written  $A \cap B$ , is the set of elements that belong to both  $A$  and  $B$ :

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

*Complementation:* The complement of  $A$ , written  $A^c$ , is the set of all elements that are not in  $A$ :

$$A^c = \{x : x \notin A\}.$$

**Example 1.1.1:** Consider the experiment of selecting a card at random from a standard deck, and noting its suit: clubs (C), diamonds (D), hearts (H), or spades (S). The sample space is

$$S = \{C, D, H, S\},$$

and some possible events are

$$A = \{C, D\} \text{ and } B = \{D, H, S\}.$$

From these events we can form

$$A \cup B = \{C, D, H, S\}, A \cap B = \{D\}, \text{ and } A^c = \{H, S\}.$$

Furthermore, notice that  $A \cup B = S$  (the event  $S$ ), and  $(A \cup B)^c = \emptyset$ , where  $\emptyset$  denotes the *empty set* (the set consisting of no elements). ||

The elementary set operations can be combined, somewhat akin to the way addition and multiplication can be combined. As long as we are careful, we can treat sets as if they were numbers. We can now state the following useful properties of set operations.

**THEOREM 1.1.1:** For any three events  $A$ ,  $B$ , and  $C$  defined on a sample space  $S$ ,

**1. Commutativity**

$$A \cup B = B \cup A,$$

$$A \cap B = B \cap A;$$

**2. Associativity**

$$A \cup (B \cup C) = (A \cup B) \cup C,$$

$$A \cap (B \cap C) = (A \cap B) \cap C;$$

**3. Distributive Laws**

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C);$$

**4. DeMorgan's Laws**

$$(A \cup B)^c = A^c \cap B^c,$$

$$(A \cap B)^c = A^c \cup B^c.$$

*Proof:* The proof of much of this theorem is left as Exercise 1.3. Also, Exercises 1.4 and 1.5 generalize the theorem. To illustrate the technique, however, we will prove the Distributive Law

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

(You might be familiar with the use of Venn diagrams to “prove” theorems in set theory. We caution that although Venn diagrams are sometimes helpful in visualizing a situation, they do not constitute a formal proof.) To prove that two sets are equal, it must be demonstrated that each set contains the other. Formally, then

$$\begin{aligned} A \cap (B \cup C) &= \{x \in S : x \in A \text{ and } x \in (B \cup C)\}; \\ (A \cap B) \cup (A \cap C) &= \{x \in S : x \in (A \cap B) \text{ or } x \in (A \cap C)\}. \end{aligned}$$

We first show that  $A \cap (B \cup C) \subset (A \cap B) \cup (A \cap C)$ . Let  $x \in (A \cap (B \cup C))$ . By the definition of intersection, it must be that  $x \in (B \cup C)$ , that is, either  $x \in B$  or  $x \in C$ . Since  $x$  also must be in  $A$ , we have that either  $x \in (A \cap B)$  or  $x \in (A \cap C)$ ; therefore,

$$x \in ((A \cap B) \cup (A \cap C)),$$

and the containment is established.

Now assume  $x \in ((A \cap B) \cup (A \cap C))$ . This implies that  $x \in (A \cap B)$  or  $x \in (A \cap C)$ . If  $x \in (A \cap B)$  then  $x$  is in both  $A$  and  $B$ . Since  $x \in B$ ,  $x \in (B \cup C)$  and thus  $x \in (A \cap (B \cup C))$ . If, on the other hand,  $x \in (A \cap C)$ , the argument is similar, and we again conclude that  $x \in (A \cap (B \cup C))$ . Thus, we have established  $(A \cap B) \cup (A \cap C) \subset A \cap (B \cup C)$ , showing containment in the other direction and, hence, proving the Distributive Law.  $\square$

The operations of union and intersection can be extended to infinite collections of sets as well. If  $A_1, A_2, A_3, \dots$  is a collection of sets, all defined on a sample space  $S$ , then

$$\begin{aligned} \bigcup_{i=1}^{\infty} A_i &= \{x \in S : x \in A_i \text{ for some } i\}, \\ \bigcap_{i=1}^{\infty} A_i &= \{x \in S : x \in A_i \text{ for all } i\}. \end{aligned}$$

For example, let  $S = (0, 1]$  and define  $A_i = [(1/i), 1]$ . Then

$$\begin{aligned} \bigcup_{i=1}^{\infty} A_i &= \bigcup_{i=1}^{\infty} [(1/i), 1] = \{x \in (0, 1] : x \in [(1/i), 1] \text{ for some } i\} \\ &= \{x \in (0, 1]\} = (0, 1]; \end{aligned}$$

$$\begin{aligned}
 \bigcap_{i=1}^{\infty} A_i &= \bigcap_{i=1}^{\infty} [(1/i), 1] = \{x \in (0, 1] : x \in [(1/i), 1] \text{ for all } i\} \\
 &= \{x \in (0, 1] : x \in [1, 1]\} \\
 &= \{1\} \quad (\text{the point } 1).
 \end{aligned}$$

It is also possible to define unions and intersections over uncountable collections of sets. If  $\Gamma$  is an index set (a set of elements to be used as indices) then

$$\begin{aligned}
 \bigcup_{a \in \Gamma} A_a &= \{x \in S : x \in A_a \text{ for some } a\}, \\
 \bigcap_{a \in \Gamma} A_a &= \{x \in S : x \in A_a \text{ for all } a\}.
 \end{aligned}$$

If, for example, we take  $\Gamma = \{\text{all positive real numbers}\}$  and  $A_a = (0, a]$ , then  $\bigcup_{a \in \Gamma} A_a = (0, \infty)$  is an uncountable union. While uncountable unions and intersections do not play a major role in statistics, they sometimes provide a useful mechanism for obtaining an answer (see Section 8.2.4).

Finally, we discuss the idea of a partition of the sample space.

 **DEFINITION 1.1.3:** Two events  $A$  and  $B$  are *disjoint* (or *mutually exclusive*) if  $A \cap B = \emptyset$ . The events  $A_1, A_2, \dots$  are *pairwise disjoint* (or *mutually exclusive*) if  $A_i \cap A_j = \emptyset$  for all  $i \neq j$ .

Disjoint sets are sets with no points in common. If we draw a Venn diagram for two disjoint sets, the sets do not overlap. The collection

$$A_i = [i, i + 1), \quad i = 0, 1, 2, \dots$$

consists of pairwise disjoint sets. Note further that  $\bigcup_{i=0}^{\infty} A_i = [0, \infty)$ .

**DEFINITION 1.1.4:** If  $A_1, A_2, \dots$  are pairwise disjoint and  $\bigcup_{i=1}^{\infty} A_i = S$ , then the collection  $A_1, A_2, \dots$  forms a *partition* of  $S$ .

The sets  $A_i = [i, i + 1)$  form a partition of  $[0, \infty)$ . In general, partitions are very useful, allowing us to divide the sample space into small, nonoverlapping pieces.

## 1.2 Probability Theory

When an experiment is performed, the realization of the experiment is an outcome in the sample space. If the experiment is performed a number of times, different outcomes may occur each time or some outcomes may repeat. This “frequency of occurrence” of an outcome can be thought of as a probability. More probable outcomes occur more frequently. If the outcomes of an experiment can be described probabilistically, we are on our way to analyzing the experiment statistically.

In this section we describe some of the basics of probability theory. We do not define probabilities in terms of frequencies but instead take the mathematically simpler axiomatic approach. As will be seen, the axiomatic approach is not concerned with the interpretations of probabilities, but is concerned only that the probabilities are defined by a function satisfying the axioms. Interpretations of the probabilities are quite another matter. The “frequency of occurrence” of an event is one example of a particular interpretation of probability. Another possible interpretation is a subjective one, where rather than thinking of probability as frequency, we can think of it as a belief in the chance of an event occurring.

### 1.2.1 Axiomatic Foundations

For each event  $A$  in the sample space  $S$  we want to associate with  $A$  a number between zero and one which will be called the probability of  $A$ , denoted by  $P(A)$ . It would seem natural to define the domain of  $P$  (the set where the arguments of the function  $P(\cdot)$  are defined) as all subsets of  $S$ ; that is, for each  $A \subset S$  we define  $P(A)$  as the probability that  $A$  occurs. Unfortunately, matters are not that simple. There are some technical difficulties to overcome. We will not dwell on these technicalities; although they are of importance, they are usually of more interest to probabilists than to statisticians. However, a firm understanding of statistics requires at least a passing familiarity with the following.

**DEFINITION 1.2.1:** A collection of subsets of  $S$  is called a *Borel field* (or *sigma-algebra*), denoted by  $\mathcal{B}$ , if it satisfies the following three properties:

1.  $\emptyset \in \mathcal{B}$  (the empty set is contained in  $\mathcal{B}$ ).
2. If  $A \in \mathcal{B}$  then  $A^c \in \mathcal{B}$  ( $\mathcal{B}$  is closed under complementation).
3. If  $A_1, A_2, \dots \in \mathcal{B}$  then  $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$  ( $\mathcal{B}$  is closed under countable unions).

The empty set  $\emptyset$  is a subset of any set. Thus,  $\emptyset \subset S$ . Property (1) states that this subset is always in a Borel field. Since  $S = \emptyset^c$ , properties (1) and (2) imply that  $S$  is always in  $\mathcal{B}$  also. In addition, from DeMorgan’s Laws it follows that  $\mathcal{B}$  is closed under countable intersections. If  $A_1, A_2, \dots \in \mathcal{B}$ , then  $A_1^c, A_2^c, \dots \in \mathcal{B}$  by property (2), and therefore  $\cap_{i=1}^{\infty} A_i^c \in \mathcal{B}$ . However, using DeMorgan’s Law (as in Exercise 1.4), we have

$$(1.2.1) \quad \left( \bigcup_{i=1}^{\infty} A_i^c \right)^c = \bigcap_{i=1}^{\infty} A_i.$$

Thus, again by property (2),  $\bigcap_{i=1}^{\infty} A_i \in \mathcal{B}$ .

Associated with sample space  $S$  we can have many different Borel fields. For example, the collection of the two sets  $\{\emptyset, S\}$  is a Borel field, usually called the trivial Borel field. The only Borel field we will be concerned with is the smallest one that contains all of the open sets in a given sample space  $S$ .

**Example 1.2.1:** If  $S$  is finite or countable, then these technicalities really do not arise, for we define for a given sample space  $S$

$\mathcal{B} = \{\text{all subsets of } S, \text{ including } S \text{ itself}\}.$

If  $S$  has  $n$  elements, there are  $2^n$  sets in  $\mathcal{B}$  (see Exercise 1.13). For example, if  $S = \{1, 2, 3\}$ , then  $\mathcal{B}$  is the following collection of  $2^3 = 8$  sets:

$$\begin{array}{lll} \{1\} & \{1, 2\} & \{1, 2, 3\} \\ \{2\} & \{1, 3\} & \emptyset \\ \{3\} & \{2, 3\} & \end{array} \quad ||$$

In general, if  $S$  is uncountable, it is not an easy task to describe  $\mathcal{B}$ . However,  $\mathcal{B}$  is chosen to contain any set of interest.

**Example 1.2.2:** Let  $S = (-\infty, \infty)$ , the real line. Then  $\mathcal{B}$  is chosen to contain all sets of the form

$$[a, b], \quad (a, b], \quad (a, b), \quad \text{and} \quad [a, b)$$

for all real numbers  $a$  and  $b$ . Also, from the properties of  $\mathcal{B}$ , it follows that  $\mathcal{B}$  contains all sets that can be formed by taking (possibly countably infinite) unions and intersections of sets of the above varieties. ||

We are now in a position to define a probability function.

**DEFINITION 1.2.2:** Given a sample space  $S$  and an associated Borel field  $\mathcal{B}$ , a *probability function* is a function  $P$  with domain  $\mathcal{B}$  that satisfies

1.  $P(A) \geq 0$  for all  $A \in \mathcal{B}$ .
2.  $P(S) = 1$ .
3. If  $A_1, A_2, \dots \in \mathcal{B}$  are pairwise disjoint, then  $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ .

The three properties given in Definition 1.2.2 are usually referred to as the Axioms of Probability (or the Kolmogorov Axioms, after A. Kolmogorov, one of the fathers of probability theory). Any function  $P$  that satisfies the Axioms of Probability is called a probability function. The axiomatic definition makes no attempt to tell what particular function  $P$  to choose; it merely requires  $P$  to satisfy the axioms. For any sample space many different probability functions can be defined. Which one(s) reflect what is likely to be observed in a particular experiment is still to be discussed.

**Example 1.2.3:** Consider the simple experiment of tossing a fair coin, so  $S = \{H, T\}$ . By a “fair” coin we mean a balanced coin that is equally as likely to land heads up as tails up, and hence the reasonable probability function is the one that assigns equal probabilities to heads and tails, that is,

$$(1.2.2) \quad P(\{H\}) = P(\{T\}).$$

Note that (1.2.2) does not follow from the axioms of probability but rather is outside of the axioms. We have used a symmetry interpretation of probability (or just intuition) to impose the requirement that heads and tails be equally probable. Since  $S = \{\text{H}\} \cup \{\text{T}\}$ , we have, from axiom (2),  $P(\{\text{H}\} \cup \{\text{T}\}) = 1$ . Also,  $\{\text{H}\}$  and  $\{\text{T}\}$  are disjoint, so  $P(\{\text{H}\} \cup \{\text{T}\}) = P(\{\text{H}\}) + P(\{\text{T}\})$  and

$$(1.2.3) \quad P(\{\text{H}\}) + P(\{\text{T}\}) = 1.$$

Simultaneously solving (1.2.2) and (1.2.3) shows that  $P(\{\text{H}\}) = P(\{\text{T}\}) = \frac{1}{2}$ .

Since (1.2.2) is based on our knowledge of the particular experiment, not the axioms, any nonnegative values for  $P(\{\text{H}\})$  and  $P(\{\text{T}\})$  that satisfy (1.2.3) define a legitimate probability function. For example, we might choose  $P(\{\text{H}\}) = \frac{1}{9}$  and  $P(\{\text{T}\}) = \frac{8}{9}$ . ||

The physical reality of the experiment might dictate the probability assignment, as the next example illustrates. Of course, the assignment must satisfy the Kolmogorov Axioms.

**Example 1.2.4:** The game of darts is played by throwing a dart at a board, and receiving a score corresponding to the number assigned to the region in which the dart lands. For a novice player, it seems reasonable to assume that the probability of the dart hitting a particular region is proportional to the area of the region. Thus, a bigger region has a higher probability of being hit.

Referring to Figure 1.2.1, the dart board has radius  $r$ , and the distance between rings is  $r/5$ . If we make the assumption that the board is always hit (see Exercise 1.42 for a variation on this), then we have

$$P(\text{scoring } i \text{ points}) = \frac{\text{Area of region } i}{\text{Area of dart board}}.$$

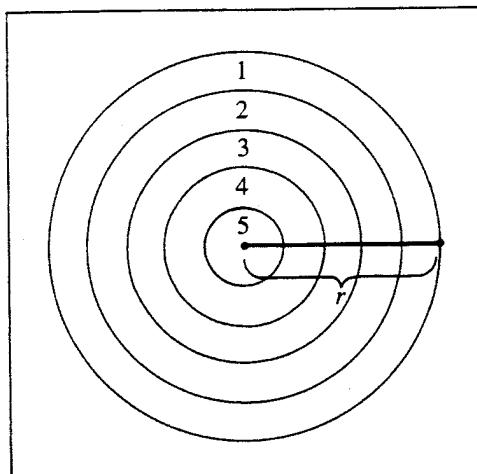


FIGURE 1.2.1 Dart board for Example 1.2.4

For example

$$P(\text{scoring 1 point}) = \frac{\pi r^2 - \pi(4r/5)^2}{\pi r^2} = 1 - \left(\frac{4}{5}\right)^2.$$

It is easy to derive the general formula, and we find that

$$P(\text{scoring } i \text{ points}) = \frac{(6-i)^2 - (5-i)^2}{5^2}, \quad i = 1, \dots, 5$$

independent of  $\pi$  and  $r$ . It is straightforward to verify that this is a probability function. (See Exercise 1.11 for further details.) ||

Before we leave the axiomatic development of probability, there is one further point to consider. Axiom (3) of Definition 1.2.2, which is commonly known as the Axiom of Countable Additivity, is not universally accepted among statisticians. Indeed, it can be argued that axioms should be simple, self-evident statements. Comparing axiom (3) to the other axioms, which are simple and self-evident, may lead us to doubt whether it is reasonable to assume the truth of axiom (3).

The Axiom of Countable Additivity is rejected by a school of statisticians led by De Finetti (1972), who chooses to replace this axiom with the Axiom of Finite Additivity.

*Axiom of Finite Additivity:* If  $A \in \mathcal{B}$  and  $B \in \mathcal{B}$  are disjoint, then

$$P(A \cup B) = P(A) + P(B).$$

While this axiom may not be entirely self-evident, it is certainly simpler than the Axiom of Countable Additivity.

Assuming only finite additivity, while perhaps more plausible, can lead to unexpected complications in statistical theory—complications that, at this level, do not necessarily enhance understanding of the subject. We therefore proceed under the assumption that the Axiom of Countable Additivity holds.

### 1.2.2 The Calculus of Probabilities

From the Axioms of Probability we can build up many properties of the probability function, properties that are quite helpful in the calculation of more complicated probabilities. Some of these manipulations will be discussed in detail in this section; others will be left as exercises.

We start with some (fairly self-evident) properties of the probability function when applied to a single event.

THEOREM 1.2.1: If  $P$  is a probability function and  $A$  is any set in  $\mathcal{B}$ , then

- a.  $P(\emptyset) = 0$ , where  $\emptyset$  is the empty set;
- b.  $P(A) \leq 1$ ;
- c.  $P(A^c) = 1 - P(A)$ .

*Proof:* It is easiest to prove (c) first. The sets  $A$  and  $A^c$  form a partition of the sample space, that is,  $S = A \cup A^c$ . Therefore,

$$(1.2.4) \quad P(A \cup A^c) = P(S) = 1,$$

by the second axiom. Also,  $A$  and  $A^c$  are disjoint, so by the third axiom,

$$(1.2.5) \quad P(A \cup A^c) = P(A) + P(A^c).$$

Combining (1.2.4) and (1.2.5) gives (c).

Since  $P(A^c) \geq 0$ , (b) is immediately implied by (c). To prove (a), we use a similar argument on  $S = S \cup \emptyset$ . (Recall that both  $S$  and  $\emptyset$  are always in  $\mathcal{B}$ .) Since  $S$  and  $\emptyset$  are disjoint, we have

$$1 = P(S) = P(S \cup \emptyset) = P(S) + P(\emptyset),$$

and thus  $P(\emptyset) = 0$ . □

Theorem 1.2.1 contains properties that are so basic that they also have the flavor of axioms, although we have formally proved them using only the original three Kolmogorov Axioms. The next theorem, which is similar in spirit to Theorem 1.2.1, contains statements that are not so self-evident.

✓ **THEOREM 1.2.2:** If  $P$  is a probability function and  $A$  and  $B$  are any sets in  $\mathcal{B}$ , then

- a.  $P(B \cap A^c) = P(B) - P(A \cap B);$
- b.  $P(A \cup B) = P(A) + P(B) - P(A \cap B);$
- c. If  $A \subset B$  then  $P(A) \leq P(B)$ .

*Proof:* To establish (a) note that for any sets  $A$  and  $B$  we have

$$B = \{B \cap A\} \cup \{B \cap A^c\},$$

and therefore

$$(1.2.6) \quad P(B) = P(\{B \cap A\} \cup \{B \cap A^c\}) = P(B \cap A) + P(B \cap A^c),$$

where the last equality in (1.2.6) follows from the fact that  $B \cap A$  and  $B \cap A^c$  are disjoint. Rearranging (1.2.6) gives (a).

To establish (b), we use the identity

$$(1.2.7) \quad A \cup B = A \cup \{B \cap A^c\}.$$

A Venn diagram will show why (1.2.7) holds, although a formal proof is not difficult (see Exercise 1.2). Using (1.2.7) and the fact that  $A$  and  $B \cap A^c$  are disjoint (since  $A$

and  $A^c$  are), we have

$$(1.2.8) \quad P(A \cup B) = P(A) + P(B \cap A^c) = P(A) + P(B) - P(A \cap B),$$

from (a).

If  $A \subset B$  then  $A \cap B = A$ . Therefore, using (a) we have

$$0 \leq P(B \cap A^c) = P(B) - P(A),$$

establishing (c).  $\square$

Formula (b) of Theorem 1.2.2 gives a useful inequality for the probability of an intersection. Since  $P(A \cup B) \leq 1$  we have from (1.2.8), after some rearranging,

$$(1.2.9) \quad P(A \cap B) \geq P(A) + P(B) - 1.$$

This inequality is a special case of what is known as *Bonferroni's Inequality* [Miller (1981) is a good reference]. Bonferroni's Inequality allows us to bound the probability of a simultaneous event (the intersection) in terms of the probabilities of the individual events.

**Example 1.2.5:** Bonferroni's Inequality is particularly useful when it is difficult (or even impossible) to calculate the intersection probability, but some idea of the size of this probability is desired. Suppose  $A$  and  $B$  are two events and each has probability .95. Then the probability that both will occur is bounded below by

$$P(A \cap B) \geq P(A) + P(B) - 1 = .95 + .95 - 1 = .90.$$

Note that unless the probabilities of the individual events are sufficiently large, the Bonferroni bound is a useless (but correct!) negative number.  $\parallel$

We close this section with a theorem that gives some useful results for dealing with a collection of sets.

**THEOREM 1.2.3:** If  $P$  is a probability function, then

- a.  $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$  for any partition  $C_1, C_2, \dots$ ;
- b.  $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$  for any sets  $A_1, A_2, \dots$  (Boole's Inequality)

*Proof:* Since  $C_1, C_2, \dots$  forms a partition we have that  $C_i \cap C_j = \emptyset$  for all  $i \neq j$ , and  $S = \bigcup_{i=1}^{\infty} C_i$ . Hence,

$$A = A \cap S = A \cap \left( \bigcup_{i=1}^{\infty} C_i \right) = \bigcup_{i=1}^{\infty} (A \cap C_i),$$

where the last equality follows from the Distributive Law (Theorem 1.1.1). We therefore have

$$P(A) = P\left(\bigcup_{i=1}^{\infty}(A \cap C_i)\right).$$

Now, since the  $C_i$  are disjoint, the sets  $A \cap C_i$  are also disjoint, and from the properties of a probability function we have

$$P\left(\bigcup_{i=1}^{\infty}(A \cap C_i)\right) = \sum_{i=1}^{\infty} P(A \cap C_i),$$

establishing (a).

To establish (b) we first construct a disjoint collection  $A_1^*, A_2^*, \dots$ , with the property that  $\cup_{i=1}^{\infty} A_i^* = \cup_{i=1}^{\infty} A_i$ . Define  $A_i^*$  by

$$A_1^* = A_1, \quad A_i^* = A_i \setminus \left( \bigcup_{j=1}^{i-1} A_j \right), \quad i = 2, 3, \dots,$$

where the notation  $A \setminus B$  denotes the part of  $A$  that does not intersect with  $B$ . In more familiar symbols,  $A \setminus B = A \cap B^c$ . It should be easy to see that  $\cup_{i=1}^{\infty} A_i^* = \cup_{i=1}^{\infty} A_i$ , and we therefore have

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} A_i^*\right) = \sum_{i=1}^{\infty} P(A_i^*),$$

where the last equality follows since the  $A_i^*$  are disjoint. To see this, write

$$\begin{aligned} A_i^* \cap A_k^* &= \left\{ A_i \setminus \left( \bigcup_{j=1}^{i-1} A_j \right) \right\} \cap \left\{ A_k \setminus \left( \bigcup_{j=1}^{k-1} A_j \right) \right\} && \text{(definition of } A_i^*) \\ &= \left\{ A_i \cap \left( \bigcup_{j=1}^{i-1} A_j \right)^c \right\} \cap \left\{ A_k \cap \left( \bigcup_{j=1}^{k-1} A_j \right)^c \right\} && \text{(definition of " \setminus ")} \\ &= \left\{ A_i \cap \bigcap_{j=1}^{i-1} A_j^c \right\} \cap \left\{ A_k \cap \bigcap_{j=1}^{k-1} A_j^c \right\} && \text{(DeMorgan's Laws)} \end{aligned}$$

Now if  $i > k$  the first intersection above will contain the set  $A_k^c$ , which will have an empty intersection with  $A_k$ . If  $k > i$  the argument is similar. Further, by construction  $A_i^* \subset A_i$ , so  $P(A_i^*) \leq P(A_i)$  and we have

$$\sum_{i=1}^{\infty} P(A_i^*) \leq \sum_{i=1}^{\infty} P(A_i),$$

establishing (b).  $\square$

There is a similarity between Boole's Inequality and Bonferroni's Inequality. In fact, they are essentially the same thing. We could have used Boole's Inequality to derive (1.2.9). If we apply Boole's Inequality to  $A^c$ , we have

$$P\left(\bigcup_{i=1}^n A_i^c\right) \leq \sum_{i=1}^n P(A_i^c),$$

and using the facts that  $\cup A_i^c = (\cap A_i)^c$  and  $P(A_i^c) = 1 - P(A_i)$ , we obtain

$$1 - P\left(\bigcap_{i=1}^n A_i\right) \leq n - \sum_{i=1}^n P(A_i).$$

This becomes, on rearranging terms,

$$(1.2.10) \quad P\left(\bigcap_{i=1}^n A_i\right) \geq \sum_{i=1}^n P(A_i) - (n - 1),$$

which is a more general version of the Bonferroni Inequality of (1.2.9).

### 1.2.3 Counting

The elementary process of counting can become quite sophisticated when placed in the hands of a statistician. Most often, methods of counting are used in order to construct probability assignments on finite sample spaces, although they can be used to answer other questions also.

**Example 1.2.6:** For a number of years the New York state lottery operated according to the following scheme. From the numbers 1, 2, ..., 44, a person may pick any six for her ticket. The winning number is then decided by randomly selecting six numbers from the forty-four. To be able to calculate the probability of winning we first must count how many different groups of six numbers can be chosen from the forty-four.  $\parallel$

**Example 1.2.7:** In a single-elimination tournament, such as the U.S. Open tennis tournament, players advance only if they win (in contrast to double-elimination or round-robin tournaments). If we have 16 entrants, we might be interested in the number of paths a particular player can take to victory, where a path is taken to mean a sequence of opponents.  $\parallel$

Counting problems, in general, sound complicated, and often we must do our counting subject to many restrictions. The way to solve such problems is to break them down into a series of simple tasks that are easy to count, and employ known rules of combining tasks. The following theorem is a first step in such a process, and is sometimes known as the Fundamental Theorem of Counting.

**THEOREM 1.2.4:** If a job consists of  $k$  separate tasks, the  $i$ th of which can be done in  $n_i$  ways,  $i = 1, \dots, k$ , then the entire job can be done in  $n_1 \times n_2 \times \dots \times n_k$  ways.

*Proof:* It suffices to prove the theorem for  $k = 2$  (see Exercise 1.14). The proof is just a matter of careful counting. The first task can be done in  $n_1$  ways, and for each of these ways we have  $n_2$  choices for the second task. Thus, we can do the job in

$$\underbrace{(1 \times n_2) + (1 \times n_2) + \dots + (1 \times n_2)}_{n_1 \text{ terms}} = n_1 \times n_2$$

ways, establishing the theorem for  $k = 2$ . □

✓ **Example 1.2.8:** Although the Fundamental Theorem of Counting is a reasonable place to start, in applications there are usually more aspects of a problem to consider. For example, in the New York state lottery the first number can be chosen in 44 ways, and the second number in 43 ways, making a total of  $44 \times 43 = 1,892$  ways of choosing the first two numbers. However, if a person is allowed to choose the same number twice, then the first two numbers can be chosen in  $44 \times 44 = 1,936$  ways. ||

The distinction being made in Example 1.2.8 is between counting with replacement and counting without replacement. There is a second crucial element in any counting problem, whether or not the ordering of the tasks is important. To illustrate with the lottery example, suppose the winning numbers are selected in the order 12, 37, 35, 9, 13, 22. Does a person who selected 9, 12, 13, 22, 35, 37 qualify as a winner? In other words, does the order in which the task is performed actually matter? Taking all of these considerations into account, we can construct a  $2 \times 2$  table of possibilities:

		Possible methods of counting	
		Without replacement	With replacement
Ordered	Ordered		
	Unordered		

Before we begin to count, the following definition gives us some extremely helpful notation.

**DEFINITION 1.2.3:** For a positive integer  $n$ ,  $n!$  (read  $n$  factorial) is the product of all of the positive integers less than or equal to  $n$ . That is,

$$n! = n \times (n - 1) \times (n - 2) \times \cdots \times 3 \times 2 \times 1.$$

Furthermore, we define  $0! = 1$ . □

Let us now consider counting all of the possible lottery tickets under each of these four cases.

1. *Ordered, without replacement* From the Fundamental Theorem of Counting, the first number can be selected in 44 ways, the second in 43 ways, etc. So there are

$$44 \times 43 \times 42 \times 41 \times 40 \times 39 = \frac{44!}{38!} = 5,082,517,440$$

possible tickets.

2. *Ordered, with replacement* Since each number can now be selected in 44 ways (because the chosen number is replaced) there are

$$44 \times 44 \times 44 \times 44 \times 44 \times 44 = 44^6 = 7,256,313,856$$

possible tickets.

3. *Unordered, without replacement* We know the number of possible tickets when the ordering must be accounted for, so what we must do is divide out the redundant orderings. Again using the Fundamental Theorem, six numbers can be arranged in  $6 \times 5 \times 4 \times 3 \times 2 \times 1$  ways, so the total number of unordered tickets is

$$\frac{44 \times 43 \times 42 \times 41 \times 40 \times 39}{6 \times 5 \times 4 \times 3 \times 2 \times 1} = \frac{44!}{6!38!} = 7,059,052.$$

This form of counting plays a central role in much of statistics—so much, in fact, that it has earned its own notation.

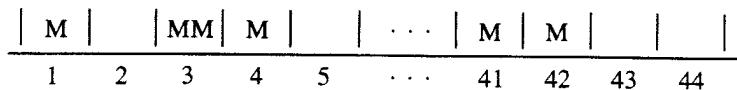
**DEFINITION 1.2.4:** For nonnegative integers  $n$  and  $r$ ,  $n \geq r$ , we define the symbol  $\binom{n}{r}$ , read  $n$  choose  $r$ , as

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

In our lottery example, the number of possible tickets (unordered, without replacement) is  $\binom{44}{6}$ . These numbers are also referred to as *binomial coefficients*, for reasons that will become clear in Chapter 3.

4. *Unordered, with replacement* This is the most difficult case to count. You might first guess that the answer is  $44^6/(6 \times 5 \times 4 \times 3 \times 2 \times 1)$ , but this is not correct (it is too small).

To count in this case, it is easiest to think of placing 6 markers on the 44 numbers. In fact, we can think of the 44 numbers defining bins in which we can place the six markers, M, as shown, for example, in the accompanying figure.



The number of possible tickets is then equal to the number of ways that we can put the 6 markers into the 44 bins. But this can be further reduced by noting that all we need to keep track of is the arrangement of the markers and the walls of the bins. Note further that the two outermost walls play no part. Thus, we have to count all of the arrangements of 43 walls (44 bins yield 45 walls, but we disregard the two end walls) and 6 markers. We therefore have  $43 + 6 = 49$  objects, which can be arranged in  $49!$  ways. However, to eliminate the redundant orderings we must divide by both  $6!$  and  $43!$ , so the total number of arrangements is

$$\frac{49!}{6!43!} = 13,983,816.$$

Although all of the preceding derivations were done in terms of an example, it should be easy to see that they hold in general. For completeness, we can summarize these situations in the following table:

	Number of possible arrangements of size $r$ from $n$ objects	
	Without replacement	With replacement
Ordered	$\frac{n!}{(n-r)!}$	$n^r$
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

### 1.2.4 Equally Likely Outcomes

The counting techniques of the previous section are useful when the sample space  $S$  is a finite set and all the outcomes in  $S$  are equally likely. Then probabilities of events can be calculated by simply counting the number of outcomes in the event. To see this, suppose that  $S = \{s_1, \dots, s_N\}$  is a finite sample space. Saying that all the outcomes are equally likely means that  $P(\{s_i\}) = 1/N$  for every outcome  $s_i$ . Then, using (3) from the definition of a probability function, we have, for any event  $A$ ,

$$P(A) = \sum_{s_i \in A} P(\{s_i\}) = \sum_{s_i \in A} \frac{1}{N} = \frac{\# \text{ of elements in } A}{\# \text{ of elements in } S}.$$

For large sample spaces, the counting techniques might be used to calculate both the numerator and denominator of this expression.

**Example 1.2.9:** Consider choosing a five-card poker hand from a standard deck of 52 playing cards. Obviously, we are sampling without replacement from the deck. But to specify the possible outcomes (possible hands), we must decide whether we think of the hand as being dealt sequentially (ordered) or all at once (unordered). If we wish to calculate probabilities for events that depend on the order, such as the probability of an ace in the first two cards, then we must use the ordered outcomes. But if our events do not depend on the order, we can use the unordered outcomes. For this example we use the unordered outcomes, so the sample space consists of all the five-card hands that can be chosen from the 52-card deck. There are  $\binom{52}{5} = 2,598,960$  possible hands. If the deck is well shuffled and the cards are randomly dealt, it is reasonable to assign probability  $1/2,598,960$  to each possible hand.

We now calculate some probabilities by counting outcomes in events. What is the probability of having four aces? How many different hands are there with four aces? Having specified that four of the cards are aces, there are 48 different ways of specifying the fifth card. Thus,

$$P(\text{four aces}) = \frac{48}{2,598,960},$$

less than 1 chance in 50,000. Only slightly more complicated counting, using Theorem 1.2.4, allows us to calculate the probability of having four of a kind. There are 13 ways to specify which denomination there will be four of. Having specified these four cards, there are 48 ways of specifying the fifth. Thus, the total number of hands with four of a kind is  $(13)(48)$  and

$$P(\text{four of a kind}) = \frac{(13)(48)}{2,598,960} = \frac{624}{2,598,960}.$$

To calculate the probability of exactly one pair (not two pair, no three of a kind, etc.) we combine some of the counting techniques. The number of hands with exactly one pair is

$$(1.2.11) \quad 13 \binom{4}{2} \binom{12}{3} 4^3 = 1,098,240.$$

Expression (1.2.11) comes from Theorem 1.2.4 because

$13 = \#$  of ways to specify the denomination for the pair,

$$\binom{4}{2} = \# \text{ of ways to specify the two cards from that denomination},$$

$$\binom{12}{3} = \# \text{ of ways of specifying the other three denominations},$$

$$4^3 = \# \text{ of ways of specifying the other three cards from those denominations.}$$

Thus,

$$P(\text{exactly one pair}) = \frac{1,098,240}{2,598,960}. \quad ||$$

### 1.3 Conditional Probability and Independence

All of the probabilities that we have dealt with thus far have been unconditional probabilities. A sample space was defined and all probabilities were calculated with respect to that sample space. In many instances, however, we are in a position to update the sample space based on new information. In such cases, we want to be able to update probability calculations or to calculate *conditional probabilities*.

**Example 1.3.1:** Four cards are dealt from the top of a well-shuffled deck. What is the probability that they are the four aces? We can calculate this probability by the methods of the previous section. The number of distinct groups of four cards is

$$\binom{52}{4} = 270,725.$$

Only one of these groups consists of the four aces and every group is equally likely, so the probability of being dealt all four aces is  $1/270,725$ .

We can also calculate this probability by an “updating” argument, as follows. The probability that the first card is an ace is  $4/52$ . *Given that the first card is an ace*, the probability that the second card is an ace is  $3/51$  (there are 3 aces and 51 cards left). Continuing this argument, we get the desired probability as

$$\frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49} = \frac{1}{270,725}. \quad ||$$

In our second method of solving the problem, we updated the sample space after each draw of a card; we calculated conditional probabilities.

DEFINITION 1.3.1: If  $A$  and  $B$  are events in  $S$ , and  $P(B) > 0$ , then the *conditional probability of  $A$  given  $B$* , written  $P(A|B)$ , is

$$(1.3.1) \quad P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad ||$$

Note that what happens in the conditional probability calculation is that  $B$  becomes the sample space:  $P(B|B) = 1$ . The intuition is that our original sample space,  $S$ , has been updated to  $B$ . All further occurrences are then calibrated with respect to their relation to  $B$ . In particular, note what happens to conditional probabilities of disjoint sets. Suppose  $A$  and  $B$  are disjoint, so  $P(A \cap B) = 0$ . It then follows that  $P(A|B) = P(B|A) = 0$ .

**Example 1.3.1 (Continued):** Although the probability of getting all four aces is quite small, let us see how the conditional probabilities change given that some aces have already been obtained. Four cards will again be dealt from a well-shuffled deck, and we now calculate

$$P(4 \text{ aces in 4 cards} | i \text{ aces in } i \text{ cards}), \quad i = 1, 2, 3.$$

The event  $\{4 \text{ aces in 4 cards}\}$  is a subset of the event  $\{i \text{ aces in } i \text{ cards}\}$ . Thus, from the definition of conditional probability, (1.3.1), we know that

$$\begin{aligned} P(4 \text{ aces in 4 cards} | i \text{ aces in } i \text{ cards}) &= \frac{P(\{4 \text{ aces in 4 cards}\} \cap \{i \text{ aces in } i \text{ cards}\})}{P(i \text{ aces in } i \text{ cards})} \\ &= \frac{P(4 \text{ aces in 4 cards})}{P(i \text{ aces in } i \text{ cards})}. \end{aligned}$$

The numerator has already been calculated, and the denominator can be calculated with a similar argument. The number of distinct groups of  $i$  cards is  $\binom{52}{i}$ , and

$$P(i \text{ aces in } i \text{ cards}) = \frac{\binom{4}{i}}{\binom{52}{i}}.$$

Therefore, the conditional probability is given by

$$P(4 \text{ aces in 4 cards} | i \text{ aces in } i \text{ cards}) = \frac{\binom{52}{i}}{\binom{52}{4} \binom{4}{i}} = \frac{(4-i)! 48!}{(52-i)!} = \frac{1}{\binom{52-i}{4-i}}.$$

For  $i = 1, 2$ , and  $3$ , the conditional probabilities are  $.00005$ ,  $.00082$ , and  $.02041$ , respectively. ||

For any  $B$  for which  $P(B) > 0$ , it is straightforward to verify that the probability function  $P(\cdot|B)$  satisfies Kolmogorov's Axioms (see Exercise 1.41). You may suspect that requiring  $P(B) > 0$  is redundant. Who would want to condition on an event of probability zero? Interestingly, sometimes this is a particularly useful way of thinking of things. However, we will defer these considerations until Chapter 4.

Conditional probabilities can be particularly slippery entities and sometimes require careful thought. Consider the following often-told tale.

**Example 1.3.2:** Three prisoners, A, B, and C are on death row. The governor decides to pardon one of the three and chooses at random the prisoner to pardon. He informs the warden of his choice but requests that the name be kept secret for a few days.

The next day, A tries to get the warden to tell him who had been pardoned. The warden refuses. A then asks which of B or C will be executed. The warden thinks for a while, then tells A that B is to be executed.

*Warden's reasoning:* Each prisoner has a  $\frac{1}{3}$  chance of being pardoned. Clearly, either B or C must be executed, so I have given A no information about whether A will be pardoned.

*A's reasoning:* Given that B will be executed, then either A or C will be pardoned. My chance of being pardoned has risen to  $\frac{1}{2}$ .

It should be clear that the warden's reasoning is correct, but let us see why. Let  $A$ ,  $B$ , and  $C$  denote the events that A, B, and C is pardoned, respectively. We know that  $P(A) = P(B) = P(C) = \frac{1}{3}$ . Let  $\mathcal{W}$  denote the event that the warden says B will die. Using (1.3.1), A can update his probability of being pardoned to

$$P(A|\mathcal{W}) = \frac{P(A \cap \mathcal{W})}{P(\mathcal{W})}.$$

What is happening can be summarized in the following table:

Prisoner pardoned	Warden tells A
A	B dies } each with equal
A	C dies } probability
B	C dies
C	B dies

Using this table, we can calculate

$$\begin{aligned} P(\mathcal{W}) &= P(\text{warden says B dies}) \\ &= P(\text{warden says B dies and A pardoned}) \\ &\quad + P(\text{warden says B dies and C pardoned}) \\ &\quad + P(\text{warden says B dies and B pardoned}) \\ &= \frac{1}{6} + \frac{1}{3} + 0 = \frac{1}{2}. \end{aligned}$$

Thus, using the warden's reasoning,

$$\begin{aligned} P(A|\mathcal{W}) &= \frac{P(A \cap \mathcal{W})}{P(\mathcal{W})} \\ (1.3.2) \quad &= \frac{P(\text{warden says B dies and A pardoned})}{P(\text{warden says B dies})} = \frac{1/6}{1/2} = \frac{1}{3}. \end{aligned}$$

However, A falsely interprets the event  $\mathcal{W}$  as equal to the event  $B^c$  and calculates

$$P(A|B^c) = \frac{P(A \cap B^c)}{P(B^c)} = \frac{1/3}{2/3} = \frac{1}{2}.$$

We see that conditional probabilities can be quite slippery and require careful interpretation. For some other variations of this problem, see Exercise 1.47. ||

Re-expressing (1.3.1) gives a useful form for calculating intersection probabilities,

$$(1.3.3) \quad P(A \cap B) = P(A|B)P(B),$$

which is essentially the formula that was used in Example 1.3.1. We can take advantage of the symmetry of (1.3.3) and also write

$$(1.3.4) \quad P(A \cap B) = P(B|A)P(A).$$

When faced with seemingly difficult calculations, we can break up our calculations according to (1.3.3) or (1.3.4), whichever is easier. Furthermore, we can equate the right-hand sides of these equations to obtain (after rearrangement)

$$(1.3.5) \quad P(A|B) = P(B|A) \frac{P(A)}{P(B)},$$

which gives us a formula for “turning around” conditional probabilities. Equation (1.3.5) is often called Bayes’ Rule for its discoverer, Sir Thomas Bayes (although see Stigler, 1983).

Bayes’ Rule has a more general form than (1.3.5), one that applies to partitions of a sample space. We therefore take the following as the definition of Bayes’ Rule.

**BAYES’ RULE:** Let  $A_1, A_2, \dots$  be a partition of the sample space, and let  $B$  be any set. Then, for each  $i = 1, 2, \dots$ ,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)P(A_j)}. \quad \square$$

**Example 1.3.3:** When coded messages are sent, there are sometimes errors in transmission. In particular, Morse code uses “dots” and “dashes,” which are known to occur in the proportion of 3:4. This means that for any given symbol

$$P(\text{dot sent}) = \frac{3}{7} \quad \text{and} \quad P(\text{dash sent}) = \frac{4}{7}.$$

Suppose there is interference on the transmission line, and with probability  $\frac{1}{8}$  a dot is mistakenly received as a dash, and vice versa. If we receive a dot, can we be sure that a dot was sent? Using Bayes' Rule, we can write

$$P(\text{dot sent} \mid \text{dot received}) = P(\text{dot received} \mid \text{dot sent}) \frac{P(\text{dot sent})}{P(\text{dot received})}.$$

Now, from the information given, we know that  $P(\text{dot sent}) = \frac{3}{7}$  and  $P(\text{dot received} \mid \text{dot sent}) = \frac{7}{8}$ . Furthermore, we can also write

$$\begin{aligned} P(\text{dot received}) &= P(\text{dot received} \cap \text{dot sent}) + P(\text{dot received} \cap \text{dash sent}) \\ &= P(\text{dot received} \mid \text{dot sent})P(\text{dot sent}) \\ &\quad + P(\text{dot received} \mid \text{dash sent})P(\text{dash sent}) \\ &= \frac{7}{8} \times \frac{3}{7} + \frac{1}{8} \times \frac{4}{7} = \frac{25}{56}. \end{aligned}$$

Combining these results, we have that the probability of correctly receiving a dot is

$$P(\text{dot sent} \mid \text{dot received}) = \frac{(7/8) \times (3/7)}{25/56} = \frac{21}{25}. \quad ||$$

In some cases it may happen that the occurrence of a particular event,  $B$ , has no effect on the probability of another event,  $A$ . Symbolically, we are saying that

$$(1.3.6) \quad P(A|B) = P(A).$$

If this holds, then by Bayes' Rule (1.3.5) we have

$$\begin{aligned} (1.3.7) \quad P(B|A) &= P(A|B) \frac{P(B)}{P(A)} \\ &= P(A) \frac{P(B)}{P(A)} \quad (\text{from (1.3.6)}) \\ &= P(B), \end{aligned}$$

so the occurrence of  $A$  has no effect on  $B$ . Moreover, since  $P(B|A)P(A) = P(A \cap B)$ , it then follows that

$$P(A \cap B) = P(A)P(B),$$

which we take as the definition of statistical independence.

**DEFINITION 1.3.2:** Two events,  $A$  and  $B$ , are *statistically independent* if

$$(1.3.8) \quad P(A \cap B) = P(A)P(B).$$

Note that independence could have been equivalently defined by either (1.3.6) or (1.3.7) (as long as either  $P(A) > 0$  or  $P(B) > 0$ ). The advantage of (1.3.8) is that it treats the events symmetrically and will be easier to generalize to more than two events.

Many gambling games provide models of independent events. The spins of a roulette wheel and the tosses of a pair of dice are both series of independent events.

**Example 1.3.4:** The gambler introduced at the start of the chapter, the Chevalier de Meré, was particularly interested in the event that he could throw at least 1 six in 4 rolls of a die. We have

$$\begin{aligned} P(\text{at least 1 six in 4 rolls}) &= 1 - P(\text{no six in 4 rolls}) \\ &= 1 - \prod_{i=1}^4 P(\text{no six on roll } i), \end{aligned}$$

where the last equality follows by independence of the rolls. On any roll, the probability of *not* rolling a six is  $\frac{5}{6}$ , so

$$P(\text{at least 1 six in 4 rolls}) = 1 - \left(\frac{5}{6}\right)^4 = .518. \quad \square$$

Independence of  $A$  and  $B$  implies independence of the complements also. In fact, we have the following theorem.

**THEOREM 1.3.1:** If  $A$  and  $B$  are independent events, then the following pairs are also independent:

- a.  $A$  and  $B^c$
- b.  $A^c$  and  $B$
- c.  $A^c$  and  $B^c$

*Proof:* We will prove only (a), leaving the rest as Exercise 1.45. To prove (a) we must show

$$P(A \cap B^c) = P(A)P(B^c).$$

From the definition of conditional probability we have

$$\begin{aligned} P(A \cap B^c) &= P(B^c|A)P(A) \\ &= [1 - P(B|A)]P(A) \quad (\text{since } P(\cdot|A) \text{ is a probability function}) \\ &= [1 - P(B)]P(A) \quad (A \text{ and } B \text{ are independent}) \\ &= P(B^c)P(A). \quad \square \end{aligned}$$

Independence of more than two events can be defined in a manner similar to (1.3.8), but we must be careful. For example, we might think that we could say  $A$ ,  $B$ , and  $C$  are independent if  $P(A \cap B \cap C) = P(A)P(B)P(C)$ . However, this is not the correct condition.

**Example 1.3.5:** Let an experiment consist of tossing two dice. For this experiment the sample space is

$$S = \{(1, 1), (1, 2), \dots, (1, 6), (2, 1), \dots, (2, 6), \dots, (6, 1), \dots, (6, 6)\},$$

that is,  $S$  consists of the 36 ordered pairs formed from the numbers 1 to 6. Define the following events:

$$A = \{\text{doubles appear}\} = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\},$$

$$B = \{\text{the sum is between 7 and 10}\},$$

$$C = \{\text{the sum is 2 or 7 or 8}\}.$$

The probabilities can be calculated by counting among the 36 possible outcomes. We have

$$P(A) = \frac{1}{6}, \quad P(B) = \frac{1}{2}, \quad \text{and} \quad P(C) = \frac{1}{3}.$$

Furthermore,

$$\begin{aligned} P(A \cap B \cap C) &= P(\text{the sum is 8, composed of double fours}) \\ &= \frac{1}{36} \\ &= \frac{1}{6} \times \frac{1}{2} \times \frac{1}{3} \\ &= P(A)P(B)P(C). \end{aligned}$$

However,

$$\begin{aligned} P(B \cap C) &= P(\text{sum equals 7 or 8}) \\ &= \frac{11}{36} \\ &\neq P(B)P(C). \end{aligned}$$

Similarly, it can be shown that  $P(A \cap B) \neq P(A)P(B)$ ; therefore, the requirement  $P(A \cap B \cap C) = P(A)P(B)P(C)$  is not a strong enough condition to guarantee pairwise independence. ||

A second attempt at a general definition of independence, in light of the previous example, might be to define  $A$ ,  $B$ , and  $C$  to be independent if all the pairs are independent. Alas, this condition also fails.

**Example 1.3.6:** Let the sample space  $S$  consist of the  $3!$  permutations of the letters  $a$ ,  $b$ , and  $c$  along with the three triples of each letter. Thus,

$$S = \left\{ \begin{array}{lll} \text{aaa} & \text{bbb} & \text{ccc} \\ \text{abc} & \text{bca} & \text{cba} \\ \text{acb} & \text{bac} & \text{cab} \end{array} \right\}.$$

Furthermore, let each element of  $S$  have probability  $\frac{1}{9}$ . Define

$$A_i = \{i\text{th place in the triple is occupied by } a\}.$$

It is then easy to count that

$$P(A_i) = \frac{1}{3}, \quad i = 1, 2, 3,$$

and

$$P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = \frac{1}{9},$$

so the  $A_i$ s are pairwise independent. But

$$P(A_1 \cap A_2 \cap A_3) = \frac{1}{9} \neq P(A_1)P(A_2)P(A_3),$$

so the  $A_i$ s do not satisfy the probability requirement. ||

The preceding two examples show that simultaneous (or mutual) independence of a collection of events requires an extremely strong definition. The following definition works.

**DEFINITION 1.3.3:** A collection of events  $A_1, \dots, A_n$  are mutually independent if for any subcollection  $A_{i_1}, \dots, A_{i_k}$ , we have

$$P\left(\bigcap_{j=1}^k A_{i_j}\right) = \prod_{j=1}^k P(A_{i_j}).$$

**Example 1.3.7:** Consider the experiment of tossing a coin three times. A sample point for this experiment must indicate the result of each toss. For example, HHT could indicate that two heads, then a tail were observed. The sample space for this experiment has eight points, namely,

$$\{\text{HHH}, \text{HHT}, \text{HTH}, \text{THH}, \text{TTH}, \text{THT}, \text{HTT}, \text{TTT}\}.$$

Let  $H_i$ ,  $i = 1, 2, 3$ , denote the event that the  $i$ th toss is a head. For example,

$$(1.3.9) \quad H_1 = \{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}\}.$$

If we assign probability  $\frac{1}{8}$  to each sample point, then using enumerations such as (1.3.9), we see that  $P(H_1) = P(H_2) = P(H_3) = \frac{1}{2}$ . This says that the coin is fair and has an equal probability of landing heads or tails on each toss.

Under this probability model, the events  $H_1$ ,  $H_2$ , and  $H_3$  are also mutually independent. To verify this we note that

$$P(H_1 \cap H_2 \cap H_3) = P(\{\text{HHH}\}) = \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = P(H_1)P(H_2)P(H_3).$$

To verify the condition in Definition 1.3.3, we also must check each pair. For example,

$$P(H_1 \cap H_2) = P(\{\text{HHH}, \text{HHT}\}) = \frac{2}{8} = \frac{1}{2} \cdot \frac{1}{2} = P(H_1)P(H_2).$$

The equality is also true for the other two pairs. Thus,  $H_1$ ,  $H_2$ , and  $H_3$  are mutually independent. That is, the occurrence of a head on any toss has no effect on any of the other tosses.

It can be verified that the assignment of probability  $\frac{1}{8}$  to each sample point is the only probability model that has  $P(H_1) = P(H_2) = P(H_3) = \frac{1}{2}$  and  $H_1$ ,  $H_2$ , and  $H_3$  mutually independent. ||

## 1.4 Random Variables

In many experiments it is easier to deal with a summary variable than with the original probability structure. For example, in an opinion poll, we might decide to ask 50 people whether they agree or disagree with a certain issue. If we record a “1” for agree and “0” for disagree, the sample space for this experiment has  $2^{50}$  elements, each an ordered string of 1s and 0s of length 50. We should be able to reduce this to a reasonable size! It may be that the only quantity of interest is the number of people who agree (equivalently, disagree) out of 50 and, if we define a variable  $X = \text{number of 1s recorded out of 50}$ , we have captured the essence of the problem. Note that the sample space for  $X$  is the set of integers  $\{0, 1, 2, \dots, 50\}$  and is much easier to deal with than the original sample space.

In defining the quantity  $X$ , we have defined a mapping (a function) from the original sample space to a new sample space, usually a set of real numbers. In general, we have the following definition.

**DEFINITION 1.4.1:** A *random variable* is a function from a sample space  $S$  into the real numbers.

**Example 1.4.1:** In most experiments random variables are implicitly used. Here are some examples.

Experiment	Random variable
Toss two dice	$X = \text{sum of the numbers}$
Toss a coin 25 times	$X = \text{number of heads in 25 tosses}$
Apply different levels of fertilizer to corn plants	$X = \text{yield/acre}$

In defining a random variable, we have also defined a new sample space (the range of the random variable). We must now check formally that our probability function, which is defined on the original sample space, can be used for the random variable.

Suppose we have a sample space

$$S = \{s_1, \dots, s_n\}$$

with a probability function  $P$  and we define a random variable  $X$  with range  $\mathcal{X} = \{x_1, \dots, x_m\}$ . We can define a probability function  $P_X$  on  $\mathcal{X}$  in the following way. We will observe  $X = x_i$  if and only if the outcome of the random experiment is an  $s_j \in S$  such that  $X(s_j) = x_i$ . Thus,

$$(1.4.1) \quad P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\}).$$

Note that the left-hand side of (1.4.1), the function  $P_X$ , is an *induced* probability function on  $\mathcal{X}$ , defined in terms of the original function  $P$ . Equation (1.4.1) formally defines a probability function,  $P_X$ , for the random variable  $X$ . Of course, we have to verify that  $P_X$  satisfies the Kolmogorov Axioms, but that is not a very difficult job (see Exercise 1.52). Because of the equivalence in (1.4.1), we will simply write  $P(X = x_i)$  rather than  $P_X(X = x_i)$ .

*A note on notation:* Random variables will always be denoted with uppercase letters and the realized values of the variable (or its range) will be denoted by the corresponding lowercase letter. Thus, the random variable  $X$  can take the value  $x$ .

**Example 1.4.2:** Consider again the experiment of tossing a fair coin three times from Example 1.3.7. Define the random variable  $X$  to be the number of heads obtained in the three tosses. A complete enumeration of the value of  $X$  for each point in the sample space is given at the top of the next page.

$s$	$X(s)$
HHH	3
HHT	2
HTH	2
THH	2
HTT	1
THT	1
TTH	1
TTT	0

The range for the random variable  $X$  is  $\mathcal{X} = \{0, 1, 2, 3\}$ . Assuming that all eight points in  $S$  have probability  $\frac{1}{8}$ , by simply counting in the above display we see that the induced probability function on  $\mathcal{X}$  is this:

$x$	$P_X(X = x)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

For example,  $P_X(X = 1) = P(\{\text{HTT}, \text{THT}, \text{TTH}\}) = \frac{3}{8}$ . ||

**Example 1.4.3:** It may be possible to determine  $P_X$  even if a complete listing, as in Example 1.4.2, is not possible. Let  $S$  be the  $2^{50}$  strings of 50 zeros and ones,  $X = \text{number of } 1\text{s}$ , and  $\mathcal{X} = \{0, 1, 2, \dots, 50\}$ , as mentioned at the beginning of this section. Suppose that each of the  $2^{50}$  strings is equally likely. The probability that  $X = 27$  can be obtained by counting all of the strings with 27 ones in the original sample space. Since each string is equally likely, it follows that

$$\begin{aligned} P_X(X = 27) &= \frac{\#\text{ strings with 27 ones}}{\#\text{ strings}} \\ &= \frac{\binom{50}{27}}{2^{50}}. \end{aligned}$$

In general, for any  $i \in \mathcal{X}$ ,

$$P_X(X = i) = \frac{\binom{50}{i}}{2^{50}}.$$

The previous illustrations had both a finite  $S$  and finite  $\mathcal{X}$ , and definition of  $P_X$  was straightforward. Such is also the case if  $\mathcal{X}$  is countable. If  $\mathcal{X}$  is uncountable, we define the induced probability function,  $P_X$ , in a manner similar to (1.4.1). For any set  $A \subset \mathcal{X}$ ,

$$(1.4.2) \quad P_X(X \in A) = P(\{s \in S : X(s) \in A\}).$$

This does define a legitimate probability function for which the Kolmogorov Axioms can be verified. (To be precise, we use (1.4.2) to define probabilities only for a certain Borel field of subsets of  $\mathcal{X}$ . But we will not concern ourselves with these technicalities.)

## 1.5 Distribution Functions

With every random variable  $X$ , we associate a function called the cumulative distribution function or *cdf* of a random variable  $X$ , denoted by  $F_X(x)$ , is defined by

$$F_X(x) = P_X(X \leq x), \quad \text{for all } x.$$

**Example 1.5.1:** Consider the experiment of tossing three fair coins, and let  $X$  = number of heads observed. The cdf of  $X$  is

$x$	$F_X(x)$
$-\infty < x < 0$	( $-\infty, 0$ ) 0
$0 \leq x < 1$	[0, 1) $\frac{1}{8}$
$1 \leq x < 2$	[1, 2) $\frac{1}{2}$
$2 \leq x < 3$	[2, 3) $\frac{7}{8}$
$3 \leq x < \infty$	[3, $\infty$ ) 1

The step function  $F_X(x)$  is graphed in Figure 1.5.1.

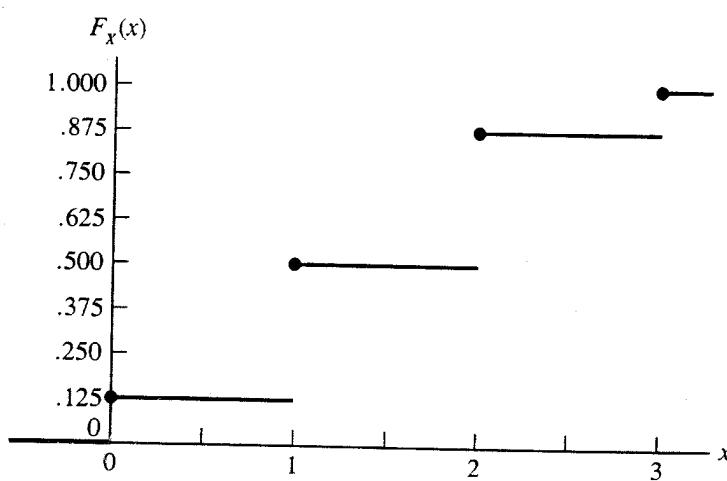


FIGURE 1.5.1 Cdf of Example 1.5.1

There are several points to note from Figure 1.5.1.  $F_X$  is defined for all values of  $x$ , not just those in  $\mathcal{X} = \{0, 1, 2, 3\}$ . Thus, for example,

$$F_X(2.5) = P(X \leq 2.5) = P(X = 0, 1, \text{ or } 2) = \frac{7}{8}.$$

Note that  $F_X$  has jumps at the values of  $x_i \in \mathcal{X}$  and the size of the jump at  $x_i$  is equal to  $P(X = x_i)$ . Also,  $F_X(x) = 0$  for  $x < 0$  since  $X$  cannot be negative and  $F_X(x) = 1$  for  $x \geq 3$  since  $x$  is certain to be less than such a value. ||

As is apparent from Figure 1.5.1,  $F_X$  can be discontinuous, with jumps at certain values of  $x$ . By the way in which  $F_X$  is defined, however, at the jump points  $F_X$  takes the value at the top of the jump. (Note the open and closed endpoints on the intervals in Example 1.5.1.) This is known as *right-continuity*—the function is continuous when a point is approached from the right. The property of right-continuity is a consequence of the definition of the cdf. In contrast, if we had defined  $F_X(x) = P_X(X < x)$  (note strict inequality),  $F_X$  would then be *left-continuous*. The size of the jump at any point  $x$  is equal to  $P(X = x)$ .

Every cdf satisfies certain properties, some of which are obvious when we think of the definition of  $F_X(x)$  in terms of probabilities.

**THEOREM 1.5.1:** The function  $F(x)$  is a cdf if and only if the following three conditions hold:

- a.  $\lim_{x \rightarrow -\infty} F(x) = 0$  and  $\lim_{x \rightarrow \infty} F(x) = 1$ .
- b.  $F(x)$  is a nondecreasing function of  $x$ .
- c.  $F(x)$  is right-continuous. That is, for every number  $x_0$ ,  $\lim_{x \downarrow x_0} F(x) = F(x_0)$ .

*Outline of proof:* To prove necessity, the three properties can be verified by writing  $F$  in terms of the probability function (Exercise 1.55). To prove sufficiency, that if a function  $F$  satisfies the three conditions of the theorem then it is a cdf for some random variable, is much harder. It must be established that there exists a sample space  $S$ , a probability function  $P$  on  $S$ , and a random variable  $X$  defined on  $S$  such that  $F$  is the cdf of  $X$ . □

**Example 1.5.2:** Suppose we do an experiment that consists of tossing a coin until a head appears. Let  $p$  = probability of a head on any given toss and define a random variable  $X$  = number of tosses required to get a head. Then, for any  $x = 1, 2, \dots$ ,

$$(1.5.1) \quad P(X = x) = (1 - p)^{x-1} p,$$

since we must get  $x - 1$  tails followed by a head for the event to occur and all trials are independent. From (1.5.1) we calculate, for any positive integer  $x$ ,

$$(1.5.2) \quad P(X \leq x) = \sum_{i=1}^x P(X = i)$$

$$= \sum_{i=1}^x (1-p)^{i-1} p, \quad x = 1, 2, \dots$$

The partial sum of the geometric series is

$$(1.5.3) \quad \sum_{k=1}^n t^{k-1} = \frac{1-t^n}{1-t}, \quad t \neq 1,$$

a fact that can be established by induction (see Exercise 1.57). Applying (1.5.3) to our probability, we find that the cdf of the random variable  $X$  is

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= \frac{1 - (1-p)^x}{1 - (1-p)} p \\ &= 1 - (1-p)^x, \quad x = 1, 2, \dots \end{aligned}$$

The cdf  $F_X(x)$  is flat between the nonnegative integers, as in Example 1.5.1.

It is easy to show that if  $0 < p < 1$ , then  $F_X(x)$  satisfies the conditions of Theorem 1.5.1. First,

$$\lim_{x \rightarrow -\infty} F_X(x) = 0$$

since  $F_X(x) = 0$  for all  $x < 0$ , and

$$\lim_{x \rightarrow \infty} F_X(x) = \lim_{x \rightarrow \infty} 1 - (1-p)^x = 1,$$

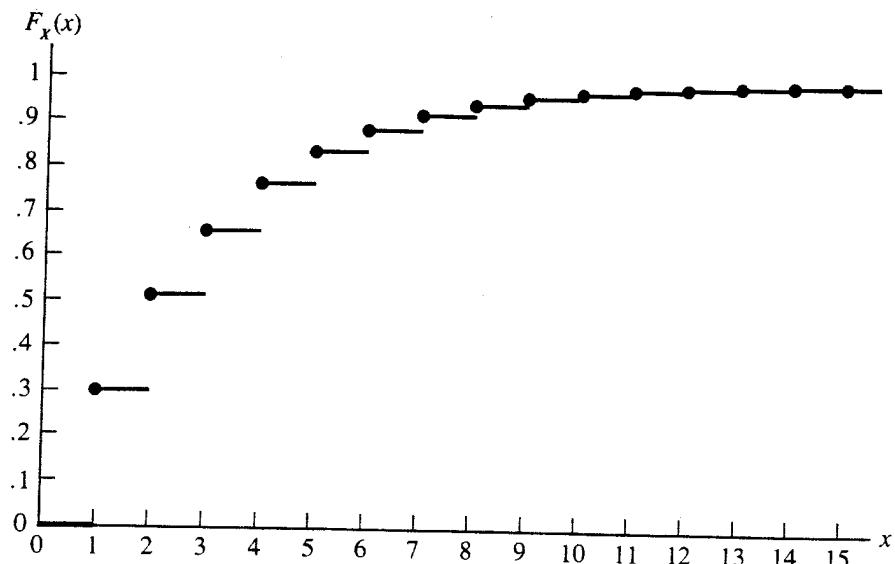
where  $x$  goes only through integer values when this limit is taken. To verify property (b), we simply note that the sum in (1.5.2) contains more *positive* terms as  $x$  increases. Finally, to verify (c), note that, for any  $x$ ,  $F_X(x + \epsilon) = F_X(x)$  if  $\epsilon > 0$  is sufficiently small. Hence,

$$\lim_{\epsilon \downarrow 0} F_X(x + \epsilon) = F_X(x),$$

so  $F_X(x)$  is right-continuous.  $F_X(x)$  is the cdf of a distribution called the *geometric distribution* (after the series) and is pictured in Figure 1.5.2 at the top of page 32. ||

**Example 1.5.3:** An example of a continuous cdf is the function

$$(1.5.4) \quad F_X(x) = \frac{1}{1 + e^{-x}},$$

FIGURE 1.5.2 Geometric cdf,  $p = .3$ 

which satisfies the conditions of Theorem 1.5.1. For example,

$$\lim_{x \rightarrow -\infty} F_X(x) = 0 \quad \text{since} \quad \lim_{x \rightarrow -\infty} e^{-x} = \infty$$

and

$$\lim_{x \rightarrow \infty} F_X(x) = 1 \quad \text{since} \quad \lim_{x \rightarrow \infty} e^{-x} = 0.$$

Differentiating  $F_X(x)$  gives

$$\frac{d}{dx} F_X(x) = \frac{e^{-x}}{(1 + e^{-x})^2} > 0,$$

showing that  $F_X(x)$  is increasing.  $F_X$  is not only right-continuous, but also continuous. This is a special case of the logistic distribution. ||

**Example 1.5.4:** If  $F_X$  is not a continuous function of  $x$ , it is possible for it to be a mixture of continuous pieces and jumps. For example, if we modify  $F_X(x)$  of (1.5.4) to be, for some  $\epsilon$ ,  $1 > \epsilon > 0$ ,

$$(1.5.5) \quad F_Y(y) = \begin{cases} \frac{1-\epsilon}{1+e^{-y}} & \text{if } y < 0 \\ \epsilon + \frac{(1-\epsilon)}{1+e^{-y}} & \text{if } y \geq 0 \end{cases}$$

then  $F_Y(y)$  is the cdf of a random variable  $Y$  (see Exercise 1.54). The function  $F_Y$  has a jump of height  $\epsilon$  at  $y = 0$ , and otherwise is continuous. This model might be appropriate if we were observing the reading from a gauge, a reading that could (theoretically) be anywhere between  $-\infty$  and  $\infty$ . This particular gauge, however, sometimes sticks at 0. We could then model our observations with  $F_Y$  where  $\epsilon$  is the probability that the gauge sticks. ||

Whether a cdf is continuous or has jumps corresponds to the associated random variable being continuous or not. In fact, the association is such that it is convenient to define continuous random variables in this way.

**DEFINITION 1.5.2:** A random variable  $X$  is *continuous* if  $F_X(x)$  is a continuous function of  $x$ . A random variable  $X$  is *discrete* if  $F_X(x)$  is a step function of  $x$ .

We close this section with a theorem formally stating that  $F_X$  completely determines the probability distribution of a random variable  $X$ . We first need the notion of two random variables being identically distributed.

**DEFINITION 1.5.3:** The random variables  $X$  and  $Y$  are *identically distributed* if for every set  $A$ ,  $P(X \in A) = P(Y \in A)$ .

Note that two random variables that are identically distributed are not necessarily equal. That is, Definition 1.5.3 does not say that  $X = Y$ .

**Example 1.5.5:** Consider the experiment of tossing a fair coin three times as in Example 1.4.2. Define the random variables  $X$  and  $Y$  by

$$X = \text{number of heads observed}$$

and

$$Y = \text{number of tails observed}.$$

The distribution of  $X$  is given in Example 1.4.2 and it is easily verified that the distribution of  $Y$  is exactly the same. That is, for each  $k = 0, 1, 2, 3$ , we have  $P(X = k) = P(Y = k)$ . So  $X$  and  $Y$  are identically distributed. However, for no sample points do we have  $X(s) = Y(s)$ . ||

**THEOREM 1.5.2:** The following two statements are equivalent:

- a. The random variables  $X$  and  $Y$  are identically distributed.
- b.  $F_X(x) = F_Y(x)$  for every  $x$ .

*Proof:* To show equivalence we must show that each statement implies the other. We first show that (a)  $\Rightarrow$  (b).

Since  $X$  and  $Y$  are identically distributed, we have for any set  $A$ ,

$$P(X \in A) = P(Y \in A).$$

In particular, for a set  $(-\infty, x]$  we have

$$P(X \in (-\infty, x]) = P(Y \in (-\infty, x]) \quad \text{for all } x.$$

But this last equality is

$$P(X \leq x) = P(Y \leq x), \quad \text{for all } x,$$

or that  $F_X(x) = F_Y(x)$ , for all  $x$ .

The converse implication, that (b)  $\Rightarrow$  (a), is much more difficult to prove. The above argument showed that if the  $X$  and  $Y$  probabilities agreed on all sets, then they agreed on intervals. We now must prove the opposite, that is, if the  $X$  and  $Y$  probabilities agree on all intervals, then they agree on all sets. To show this requires heavy use of Borel fields; we will not go into these details here. Suffice it to say that it is necessary to prove only that the two probability functions agree on all intervals (Chung, 1974).  $\square$

## 1.6 Density and Mass Functions

Associated with a random variable  $X$  and its cdf  $F_X$  is another function, called either the probability density function (pdf) or probability mass function (pmf). The terms pdf and pmf refer, respectively, to the continuous and discrete cases. Both pdfs and pmfs are concerned with “point probabilities” of random variables.

**DEFINITION 1.6.1:** The *probability mass function (pmf)* of a discrete random variable  $X$  is given by

$$f_X(x) = P(X = x) \quad \text{for all } x.$$

**Example 1.6.1:** For the geometric distribution of Example 1.5.2, we have the pmf

$$f_X(x) = P(X = x) = \begin{cases} (1 - p)^{x-1} p & \text{for } x = 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$

Recall that  $P(X = x)$  or, equivalently,  $f_X(x)$  is the size of the jump in the cdf at  $x$ . We can use the pmf to calculate probabilities. Since we can now measure the probability of a single point, we need only sum over all of the points in the appropriate event. Hence, for positive integers  $a$  and  $b$ , with  $a \leq b$ , we have

$$\begin{aligned} P(a \leq X \leq b) &= \sum_{k=a}^b f_X(k) \\ &= \sum_{k=a}^b (1 - p)^{k-1} p. \end{aligned}$$

As a special case of this we get

$$(1.6.1) \quad P(X \leq b) = \sum_{k=1}^b f_X(k) = F_X(b). \quad ||$$

A widely accepted convention, which we will adopt, is to use an uppercase letter for the cdf and the corresponding lowercase letter for the pmf or pdf.

We must be a little more careful in our definition of a pdf in the continuous case. If we naively try to calculate  $P(X = x)$  for a continuous random variable, we get the following. Since  $\{X = x\} \subset \{x - \epsilon < X \leq x\}$  for any  $\epsilon > 0$ , we have from Theorem 1.2.2(c) that

$$P(X = x) \leq P(x - \epsilon < X \leq x) = F_X(x) - F_X(x - \epsilon),$$

for any  $\epsilon > 0$ . Therefore,

$$0 \leq P(X = x) \leq \lim_{\epsilon \downarrow 0} [F_X(x) - F_X(x - \epsilon)] = 0,$$

by the continuity of  $F_X$ . However, if we understand the purpose of the pdf, its definition will become clear.

From Example 1.6.1, we see that a pmf gives us “point probabilities.” In the discrete case, we can sum over values of the pmf to get the cdf (as in (1.6.1)). The analogous procedure in the continuous case is to substitute integrals for sums, and we get

$$P(X \leq x) = F_X(x) = \int_{-\infty}^x f_X(t)dt.$$

Using the Fundamental Theorem of Calculus, if  $f_X(x)$  is continuous, we have the further relationship

$$(1.6.2) \quad \frac{d}{dx} F_X(x) = f_X(x).$$

Note that the analogy with the discrete case is almost exact. We “add up” the “point probabilities”  $f_X(x)$  to obtain interval probabilities.

**DEFINITION 1.6.2:** The *probability density function* or *pdf*,  $f_X(x)$ , of a continuous random variable  $X$  is the function that satisfies

$$(1.6.3) \quad F_X(x) = \int_{-\infty}^x f_X(t)dt \quad \text{for all } x.$$

*A note on notation:* The expression “ $X$  has a distribution given by  $F_X(x)$ ” is abbreviated symbolically by “ $X \sim F_X(x)$ ,” where we read the symbol “ $\sim$ ” as “is distributed as.” We can similarly write  $X \sim f_X(x)$  or, if  $X$  and  $Y$  have the same distribution,  $X \sim Y$ .

In the continuous case we can be somewhat cavalier about the specification of interval probabilities. Since  $P(X = x) = 0$  if  $X$  is a continuous random variable,

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b).$$

It should be clear that the pdf (or pmf) contains the same information as the cdf. This being the case, we can use either one to solve problems and should try to choose the simpler one.

**Example 1.6.2:** For the logistic distribution of Example 1.5.3 we have

$$F_X(x) = \frac{1}{1 + e^{-x}}$$

~~Ex~~

and, hence,

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{e^{-x}}{(1 + e^{-x})^2}.$$

The area under the curve  $f_X(x)$  gives us interval probabilities (see Figure 1.6.1):

$$P(a < X < b) = F_X(b) - F_X(a) = \int_{-\infty}^b f_X(x)dx - \int_{-\infty}^a f_X(x)dx = \int_a^b f_X(x)dx.$$

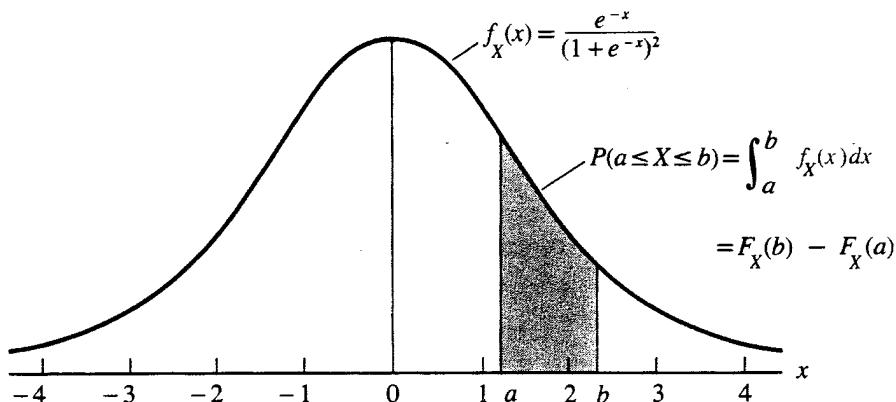


FIGURE 1.6.1 Area under logistic curve

||

There are really only two requirements for a pdf (or pmf), both of which are immediate consequences of the definition.

**THEOREM 1.6.1:** A function  $f_X(x)$  is a pdf (or pmf) of a random variable  $X$  if and only if

- a.  $f_X(x) \geq 0$  for all  $x$ .
- b.  $\sum_x f_X(x) = 1$  (pmf) or  $\int_{-\infty}^{\infty} f_X(x)dx = 1$  (pdf).

*Proof:* If  $f_X(x)$  is a pdf (or pmf), then the two properties are immediate from the definitions. In particular, for a pdf, using (1.6.3) and Theorem 1.5.1, we have that

$$1 = \lim_{x \rightarrow \infty} F_X(x) = \int_{-\infty}^{\infty} f_X(t) dt.$$

The converse implication is equally easy to prove. Once we have  $f_X(x)$  we can define  $F_X(x)$  and appeal to Theorem 1.5.1.  $\square$

From a purely mathematical viewpoint, any nonnegative function with a finite positive integral (or sum) can be turned into a pdf or pmf. For example, if  $h(x)$  is any nonnegative function that is positive on a set  $A$ , 0 elsewhere, and

$$\int_{\{x \in A\}} h(x) dx = K < \infty$$

for some constant  $K > 0$ , then the function  $f_X(x) = h(x)/K$  is a pdf of a random variable  $X$  taking values in  $A$ .

Actually, the relationship (1.6.3) does not always hold because  $F_X(x)$  may be continuous but not differentiable. In fact, there exist continuous random variables for which the integral relationship does not exist for *any*  $f_X(x)$ . These cases are rather pathological and we will ignore them. Thus, in this text, we will assume that (1.6.3) holds for any continuous random variable. In more advanced texts (for example, Chung (1974)) a random variable is called *absolutely continuous* if (1.6.3) holds.

## EXERCISES

---

- 1.1** For each of the following experiments, describe the sample space.
  - a. Toss a coin four times.
  - b. Count the number of insect-damaged leaves on a plant.
  - c. Measure the lifetime (in hours) of a particular brand of light bulb.
  - d. Record the weights of 10-day-old rats.
  - e. Observe the proportion of defectives in a shipment of electronic components.
- 1.2** Verify the following identities.
 

a. $A \setminus B = A \setminus (A \cap B) = A \cap B^c$	b. $B = (B \cap A) \cup (B \cap A^c)$
c. $B \setminus A = B \cap A^c$	d. $A \cup B = A \cup (B \cap A^c)$ .
- 1.3** Finish the proof of Theorem 1.1.1. For any events  $A$ ,  $B$ , and  $C$  defined on a sample space  $S$ , show that
  - a.  $A \cup B = B \cup A$  and  $A \cap B = B \cap A$ . (Commutativity)
  - b.  $A \cup (B \cup C) = (A \cup B) \cup C$  and  $A \cap (B \cap C) = (A \cap B) \cap C$ . (Associativity)
  - c.  $(A \cup B)^c = A^c \cap B^c$  and  $(A \cap B)^c = A^c \cup B^c$ . (DeMorgan's Laws)
- 1.4** Prove the general version of DeMorgan's Laws. Let  $\{A_\alpha : \alpha \in \Gamma\}$  be a (possibly uncountable) collection of sets. Prove that
  - a.  $(\cup_\alpha A_\alpha)^c = \cap_\alpha A_\alpha^c$ .
  - b.  $(\cap_\alpha A_\alpha)^c = \cup_\alpha A_\alpha^c$ .
- 1.5** Formulate and prove a version of DeMorgan's Laws that applies to a finite collection of sets  $A_1, \dots, A_n$ .
- 1.6** Let  $S$  be a sample space. Show that the collection  $\mathcal{B} = \{\emptyset, S\}$  is a Borel field.
- 1.7** Let  $S$  be a sample space, and let  $\mathcal{B} = \{\text{all subsets of } S, \text{ including } S \text{ itself}\}$ . Show that  $\mathcal{B}$  is a Borel field.
- 1.8** Show that the intersection of two Borel fields is a Borel field.

- 1.9** It was noted in Section 1.2.1 that statisticians who follow the De Finetti school do not accept the Axiom of Countable Additivity. However, this axiom can be derived from two others that are somewhat more self-evident. Suppose we have an infinite sequence of nested sets  $A_1 \supset A_2 \supset \dots \supset A_n \supset \dots$  that are decreasing to the empty set, which we denote by  $A_n \downarrow \emptyset$ . Consider the following:

*Axiom of Continuity:* If  $A_n \downarrow \emptyset$  then  $P(A_n) \rightarrow 0$ .

Prove that the Axiom of Continuity together with the Axiom of Finite Additivity imply the Axiom of Countable Additivity.

- 1.10** If  $P(A) = \frac{1}{3}$  and  $P(B^c) = \frac{1}{4}$ , can  $A$  and  $B$  be disjoint? Explain.
- 1.11** Refer to the game of darts explained in Example 1.2.4, and to Figure 1.2.1.
- Derive the general formula for the probability of scoring  $i$  points.
  - Show that  $P(\text{scoring } i \text{ points})$  is a decreasing function of  $i$ , that is, as the points increase, the probability of scoring them decreases.
  - Show that  $P(\text{scoring } i \text{ points})$  satisfies the Kolmogorov Axioms.
- 1.12** For events  $A$  and  $B$ , find formulas for the probabilities of the following events in terms of the quantities  $P(A)$ ,  $P(B)$ , and  $P(A \cap B)$ .
- either  $A$  or  $B$  or both
  - either  $A$  or  $B$  but not both
  - at least one of  $A$  or  $B$
  - at most one of  $A$  or  $B$
- 1.13** Suppose that a sample space  $S$  has  $n$  elements. Prove that the number of subsets that can be formed from the elements of  $S$  is  $2^n$ .
- 1.14** Finish the proof of Theorem 1.2.4. Use the result established for  $k = 2$  as the basis of an induction argument.
- 1.15** How many different sets of initials can be formed if every person has one surname and
- exactly two given names?
  - either one or two given names?
  - either one or two or three given names?
- 1.16** In the game of dominos, each piece is marked with two numbers. The pieces are symmetrical so that the number pair is not ordered (so, for example,  $(2, 6) = (6, 2)$ ). How many different pieces can be formed using the numbers  $1, 2, \dots, n$ ?
- 1.17** If  $n$  balls are placed at random into  $n$  cells, find the probability that exactly one cell remains empty.
- 1.18** There is interest in obtaining information about two species of fish that are known to inhabit a particular pond (other species will be ignored). Samples of size 5 will be drawn from the pond.
- If there are  $M$  fish of species A and  $N$  of species B, describe the sample space of the experiment. In particular, how many elements are in the sample space?
  - Assuming that all samples are equally likely, what is the probability of getting at least three fish of species B in a sample?
  - A second experiment is now planned, in which there is interest in species A, B, C, and D. Samples of size 8 are to be taken. How many different samples are there?
- 1.19** If a multivariate function has continuous partial derivatives, the order in which the derivatives are calculated does not matter. Thus, for example, the function  $f(x, y)$  of two variables has equal third partials

$$\frac{\partial^3}{\partial x^2 \partial y} f(x, y) = \frac{\partial^3}{\partial y \partial x^2} f(x, y).$$

- How many fourth partial derivatives does a function of three variables have?
- Prove that a function of  $n$  variables has  $\binom{n+r-1}{r}$   $r$ th partial derivatives.

- 1.20** A secretary types four letters to four people and addresses the four envelopes. If he inserts the letters at random, one in each envelope, what is the probability that exactly two letters will go into the correct envelopes? exactly three?
- 1.21** In a certain family, the four children take turns washing dishes. Out of a total of four breakages, three were caused by the youngest child, who thereafter was called clumsy. Was it justified to call the child clumsy, or could such an occurrence be reasonably attributed to chance?
- 1.22** A committee is to consist of four academicians and two industrialists, to be chosen from a larger group of eight academicians and five industrialists. How many ways can a committee be formed if
- there are no additional restrictions?
  - two of the chosen academicians must be the two female members of the group of eight?
- 1.23** A way of approximating large factorials is through the use of *Stirling's Formula*:

$$n! \approx \sqrt{2\pi} n^{n+(1/2)} e^{-n},$$

a complete derivation of which is difficult. Instead, prove the easier fact

$$\lim_{n \rightarrow \infty} \frac{n!}{n^{n+(1/2)} e^{-n}} = \text{a constant.}$$

(Hint: Feller (1968) proceeds by using the monotonicity of the logarithm to establish that

$$\int_{k-1}^k \log x dx < \log k < \int_k^{k+1} \log x dx, \quad k = 1, \dots, n,$$

and hence

$$\int_0^n \log x dx < \log n! < \int_1^{n+1} \log x dx.$$

Now compare  $\log n!$  to the average of the two integrals. See Exercise 5.21 for another derivation.)

- 1.24** My telephone rings 12 times each week, the calls being randomly distributed among the 7 days. What is the probability that I get at least one call each day?
- 1.25** Verify the following identities for  $n \geq 2$ .
- $\sum_{k=0}^n (-1)^k \binom{n}{k} = 0$
  - $\sum_{k=1}^n k \binom{n}{k} = n2^{n-1}$
  - $\sum_{k=1}^n (-1)^{k+1} k \binom{n}{k} = 0$
- 1.26** A man is given  $n$  keys, in random order, of which only one will unlock his door. He tries them successively (sampling without replacement). This procedure may require  $1, 2, \dots, n$  trials. Show that each of the  $n$  outcomes has probability  $1/n$ . (In other words,  $P(\text{man succeeds on } k\text{th trial}) = 1/n, k = 1, 2, \dots, n$ .)
- 1.27** A closet contains  $n$  pairs of shoes. If  $2r$  shoes are chosen at random ( $2r < n$ ), what is the probability that there will be no matching pair in the sample?
- 1.28** In a draft lottery containing the 366 days of the year (including February 29), what is the probability that the first 180 days drawn (without replacement) are evenly distributed among the 12 months? What is the probability that the first 30 days drawn contain none from September?

- 1.29 Two people each toss a fair coin  $n$  times. Find the probability that they will score the same number of heads.
- 1.30 An electron spin resonance spectrometer (ESR) detects atoms in a molecule that have a free (unpaired) electron. A *hyperfine coupling constant* measures the magnetic attraction of the electron to such an atom. The output of an ESR gives a set of possible hyperfine coupling constants for the atoms in a molecule. Although each atom corresponds to one coupling constant from the set, different atoms can have the same constant. The exact matching of atoms to constants is immaterial because, for a given set of constants, all possible arrangements of those constants correspond to the same compound. For example, if the set of possible constants is  $\{1, 2, \dots, 10\}$ , then the two assignments of constants  $\{3, 3, 3, 6, 7, 7\}$  diagrammed below are considered equivalent. If there are  $n$  atoms in a molecule, and  $r$  distinct values in the set of possible coupling constants, how many different coupling assignments are there? (See Duling, Motten, and Mason (1988) for a more thorough description of this problem.)

	7 OH·		3 OH·
3 H	H 3	3 H	H 3
7 H	H 3	7 H	H 7
H 6		H 6	

- 1.31 An employer is about to hire one new employee from a group of  $N$  candidates, whose future potential can be rated on a scale from 1 to  $N$ . The employer proceeds according to the following rules:

- i. Each candidate is seen in succession (in random order) and a decision is made whether to hire the candidate.
- ii. Having rejected  $m - 1$  candidates ( $m > 1$ ), the employer can hire the  $m$ th candidate only if the  $m$ th candidate is better than the previous  $m - 1$ .

Suppose a candidate is hired on the  $i$ th trial. What is the probability that the best candidate was hired?

- 1.32 Approximately one-third of all human twins are identical (one-egg) and two-thirds are fraternal (two-egg) twins. Identical twins are necessarily the same sex, with male and female being equally likely. Among fraternal twins, approximately one-fourth are both female, one-fourth are both male, and half are one male and one female. Finally, among all U.S. births, approximately 1 in 90 is a twin birth. Define the following events:

$$A = \{\text{a U.S. birth results in twin females}\}$$

$$B = \{\text{a U.S. birth results in identical twins}\}$$

$$C = \{\text{a U.S. birth results in twins}\}$$

- a. State, in words, the event  $A \cap B \cap C$ .
  - b. Find  $P(A \cap B \cap C)$ .
- 1.33 Two pennies, one with  $P(\text{head}) = u$  and one with  $P(\text{head}) = w$ , are to be tossed together independently. Define

$$p_0 = P(0 \text{ heads occur}),$$

$$p_1 = P(1 \text{ head occurs}),$$

$$p_2 = P(2 \text{ heads occur}).$$

Can  $u$  and  $w$  be chosen such that  $p_0 = p_1 = p_2$ ? Prove your answer.

- 1.34** In a town of  $n + 1$  inhabitants, a person tells a rumor to a second person, who in turn repeats it to a third person, etc. At each step the recipient of the rumor is chosen at random from the  $n$  people. Find the probability that the rumor will be told exactly  $r$  times
- before returning to the originator.
  - without being repeated to any person.
- 1.35** An airport bus deposits 25 passengers at 7 stops. Each passenger is as likely to get off at any stop as at any other, and the passengers act independently of one another. The bus makes a stop only if someone wants to get off. What is the probability that nobody gets off at the third stop?
- 1.36** Two players, A and B, alternately and independently flip a coin and the first player to obtain a head wins. Assume player A flips first.
- If the coin is fair, what is the probability that A wins?
  - Suppose that  $P(\text{head}) = p$ , not necessarily  $\frac{1}{2}$ . What is the probability that A wins?
  - Show that for all  $p$ ,  $0 < p < 1$ ,  $P(\text{A wins}) > \frac{1}{2}$ . (*Hint:* Try to write  $P(\text{A wins})$  in terms of the events  $E_1, E_2, \dots$ , where  $E_i = \{\text{head first appears on } i\text{th toss}\}$ .)
- 1.37** Suppose that 5% of men and .25% of women are color-blind. A person is chosen at random and that person is color-blind. What is the probability that the person is male? (Assume males and females to be in equal numbers.)
- 1.38** An insurance company has three types of customers—high risk, medium risk, and low risk. Twenty percent of its customers are high risk, 30% are medium risk, and 50% are low risk. Also, the probability that a customer has at least one accident in the current year is .25 for high risk, .16 for medium risk, and .10 for low risk.
- Find the probability that a customer chosen at random will have at least one accident in the current year.
  - Find the probability that a customer is high risk, given that the person has had at least one accident during the current year.
- 1.39** In a certain population, 1% of the people are color-blind. A subset is to be randomly chosen. How large must the subset be if the probability of its containing at least one color-blind person is to be .95 or more? (Assume that the population is large enough to be considered infinite, so the selection can be considered to be with replacement.)
- 1.40** Two litters of a particular rodent species have been born, one with two brown-haired and one gray-haired (litter 1), and the other with three brown-haired and two gray-haired (litter 2). We select a litter at random, and then select an offspring at random from the selected litter.
- What is the probability that the animal chosen is brown-haired?
  - Given that a brown-haired offspring was selected, what is the probability that the sampling was from litter 1?
- 1.41** Prove that if  $P(\cdot)$  is a legitimate probability function and  $B$  is a set with  $P(B) > 0$ , then  $P(\cdot|B)$  also satisfies Kolmogorov's Axioms.
- 1.42** Refer to the dart game of Example 1.2.4. Suppose we do not assume that the probability of hitting the dart board is 1, but rather is proportional to the area of the dart board. Assume that the dart board is mounted on a wall that is hit with probability 1, and the wall has area  $A$ .
- Using the fact that the probability of hitting a region is proportional to area, construct

a probability function for  $P(\text{scoring } i \text{ points})$ ,  $i = 0, \dots, 5$ . (No points are scored if the dart board is not hit.)

b. Show that the conditional probability distribution  $P(\text{scoring } i \text{ points} | \text{board is hit})$  is exactly the probability distribution of Example 1.2.4.

- 1.43** Prove each of the following statements. (Assume that any conditioning event has positive probability.)

a. If  $P(B) = 1$  then  $P(A|B) = P(A)$  for any  $A$ .

b. If  $A \subset B$  then  $P(B|A) = 1$  and  $P(A|B) = P(A)/P(B)$ .

c. If  $A$  and  $B$  are mutually exclusive, then

$$P(A|A \cup B) = \frac{P(A)}{P(A) + P(B)}.$$

d.  $P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$ .

- 1.44** A pair of events  $A$  and  $B$  cannot be simultaneously *mutually exclusive* and *independent*. Prove that if  $P(A) > 0$  and  $P(B) > 0$ , then

a. If  $A$  and  $B$  are mutually exclusive they cannot be independent.

b. If  $A$  and  $B$  are independent they cannot be mutually exclusive.

- 1.45** Finish the proof of Theorem 1.3.1. If  $A$  and  $B$  are independent events, show that the following pairs are also independent.

a.  $A^c$  and  $B$

b.  $A^c$  and  $B^c$

- 1.46** If the probability of hitting a target is  $\frac{1}{5}$ , and ten shots are fired independently, what is the probability of the target being hit at least twice? What is the conditional probability that the target is hit at least twice, given that it is hit at least once?

- 1.47** Here we will look at some variations of Example 1.3.2. A similar, but somewhat more complicated, problem is discussed by Selvin (1975).

a. In the warden's calculation of Example 1.3.2 it was assumed that if A were to be pardoned, then with equal probability the warden would tell A that either B or C would die. However, this need not be the case. The warden can assign probabilities  $\gamma$  and  $1 - \gamma$  to these events, as shown here:

Prisoner pardoned	Warden tells A	
A	B dies	with probability $\gamma$
A	C dies	with probability $1 - \gamma$
B	C dies	
C	B dies	

Calculate  $P(A|\mathcal{W})$  as a function of  $\gamma$ . For what values of  $\gamma$  is  $P(A|\mathcal{W})$  less than, equal to, or greater than  $\frac{1}{3}$ ?

b. Suppose again that  $\gamma = \frac{1}{2}$ , as in the example. After the warden tells A that B will die, A thinks for a while and realizes that his original calculation was false. However, A then gets a bright idea. A asks the warden if he can swap fates with C. The warden, thinking that no information has been passed, agrees to this. Prove that A's reasoning is now correct and that his probability of survival has jumped to  $\frac{2}{3}$ !

- 1.48** A fair die is cast until a 6 appears. What is the probability that it must be cast more than five times?

- 1.49** The Smiths have two children. At least one of them is a boy. What is the probability that both children are boys? (See Gardner (1961) for a complete discussion of this problem.)

- 1.50** As in Example 1.3.3, consider telegraph signals “dot” and “dash” sent in the proportion 3:4, where erratic transmissions cause a dot to become a dash with probability  $\frac{1}{4}$ , and a dash to become a dot with probability  $\frac{1}{3}$ .
- If a dash is received, what is the probability that a dash has been sent?
  - Assuming independence between signals, if the message dot-dot was received, what is the probability distribution of the four possible messages that could have been sent?
- 1.51** Standardized tests provide an interesting application of probability theory. Suppose first that a test consists of 20 multiple-choice questions, each with 4 possible answers. If the student guesses on each question, then the taking of the exam can be modeled as a sequence of 20 independent events. Find the probability that the student gets at least 10 questions correct, given that he is guessing.
- 1.52** Show that the induced probability function defined in (1.4.1) defines a legitimate probability function in that it satisfies the Kolmogorov Axioms.
- 1.53** Seven balls are distributed randomly into seven cells. Let  $X_i$  = the number of cells containing exactly  $i$  balls. What is the probability distribution of  $X_3$ ? (That is, find  $P(X_3 = x)$  for every possible  $x$ .)
- 1.54** Prove that the following functions are cdfs.
- $\frac{1}{2} + \frac{1}{\pi} \tan^{-1}(x), x \in (-\infty, \infty)$
  - $(1 + e^{-x})^{-1}, x \in (-\infty, \infty)$
  - $e^{-e^{-x}}, x \in (-\infty, \infty)$
  - $1 - e^{-x}, x \in (0, \infty)$
  - the function defined in (1.5.5)
- 1.55** Prove the necessity part of Theorem 1.5.1.
- 1.56** A cdf  $F_X$  is *stochastically greater* than a cdf  $F_Y$  if  $F_X(t) \leq F_Y(t)$  for all  $t$  and  $F_X(t) < F_Y(t)$  for some  $t$ . Prove that if  $X \sim F_X$  and  $Y \sim F_Y$ , then

$$P(X > t) \geq P(Y > t) \quad \text{for every } t$$

and

$$P(X > t) > P(Y > t) \quad \text{for some } t,$$

that is,  $X$  tends to be bigger than  $Y$ .

- 1.57** Verify formula (1.5.3), the formula for the partial sum of the geometric series.
- 1.58** An appliance store receives a shipment of 30 microwave ovens, 5 of which are (unknown to the manager) defective. The store manager selects 4 ovens at random, without replacement, and tests to see if they are defective. Let  $X$  = number of defectives found. Calculate the pmf and cdf of  $X$  and plot the cdf.
- 1.59** Let  $X$  be a continuous random variable with pdf  $f(x)$  and cdf  $F(x)$ . For a fixed number  $x_0$ , define the function

$$g(x) = \begin{cases} f(x)/[1 - F(x_0)] & x \geq x_0 \\ 0 & x < x_0. \end{cases}$$

Prove that  $g(x)$  is a pdf. (Assume that  $F(x_0) < 1$ .)

- 1.60** A certain river floods every year. Suppose that the low-water mark is set at 1 and the high-water mark  $Y$  has distribution function

$$F_Y(y) = P(Y \leq y) = 1 - \frac{1}{y^2}, \quad 1 \leq y < \infty.$$

- a. Verify that  $F_Y(y)$  is a cdf.  
 b. Find  $f_Y(y)$ , the pdf of  $Y$ .  
 c. If the low-water mark is reset at zero and we use a unit of measurement which is  $\frac{1}{10}$  of that given previously, the high-water mark becomes  $Z = 10(Y - 1)$ . Find  $F_Z(z)$ .
- 1.61** For each of the following, determine the value of  $c$  that makes  $f(x)$  a pdf.
- a.  $f(x) = c \sin x, 0 < x < \pi/2$       b.  $f(x) = ce^{-|x|}, -\infty < x < \infty$
- 1.62** Suppose an electronic device has lifetime denoted by  $T$ . The device has value  $V = 5$  if it fails before time  $t = 3$ ; otherwise, it has value  $V = 2T$ . If  $T$  has pdf

$$f_T(t) = \frac{1}{1.5} e^{-t/(1.5)}, \quad t > 0,$$

find the cdf of  $V$ .

- 1.63** A die is constructed so that the probability of tossing  $i$  is proportional to  $i$  ( $i = 1, \dots, 6$ ). What is  $P(\text{toss an } i)$ ?

# 2 Transformations and Expectations

*"Like all Holmes's reasoning the thing seemed simplicity itself when it was once explained."*

**Dr. Watson**  
*The Stockbroker's Clerk*

Often, if we are able to model a phenomenon in terms of a random variable  $X$  with cdf  $F_X(x)$ , we will also be concerned with the behavior of functions of  $X$ . In this chapter we study techniques that allow us to gain information about functions of  $X$  that may be of interest, information that can range from very complete (the distributions of these functions) to more vague (the average behavior).

## 2.1 Distributions of Functions of a Random Variable

If  $X$  is a random variable with cdf  $F_X(x)$ , then any function of  $X$ , say  $g(X)$ , is also a random variable. Often  $g(X)$  is of interest itself and we write  $Y = g(X)$  to denote the new random variable  $g(X)$ . Since  $Y$  is a function of  $X$ , we can describe the probabilistic behavior of  $Y$  in terms of that of  $X$ . That is, for any set  $A$ ,

$$P(Y \in A) = P(g(X) \in A),$$

showing that the distribution of  $Y$  depends on the functions  $F_X$  and  $g$ . Depending on the choice of  $g$ , it is sometimes possible to obtain a tractable expression for this probability.

Formally, if we write  $y = g(x)$ , the function  $g(x)$  defines a mapping from the original sample space of  $X$ ,  $\mathcal{X}$ , to a new sample space,  $\mathcal{Y}$ , the sample space of the random variable  $Y$ . That is,

$$g(x): \mathcal{X} \rightarrow \mathcal{Y}.$$

We associate with  $g$  an inverse mapping, denoted by  $g^{-1}$ , which is a mapping from subsets of  $\mathcal{Y}$  to subsets of  $\mathcal{X}$ , and is defined by

$$(2.1.1) \quad g^{-1}(A) = \{x \in \mathcal{X}: g(x) \in A\}.$$

Note that the mapping  $g^{-1}$  takes sets into sets, that is,  $g^{-1}(A)$  is the set of points in  $\mathcal{X}$  that  $g(x)$  takes into the set  $A$ . It is possible for  $A$  to be a point set, say  $A = \{y\}$ . Then

$$g^{-1}(\{y\}) = \{x \in \mathcal{X}: g(x) = y\}.$$

In this case we often write  $g^{-1}(y)$  instead of  $g^{-1}(\{y\})$ . The quantity  $g^{-1}(y)$  can still be a set, however, if there is more than one  $x$  for which  $g(x) = y$ . If there is only one  $x$  for which  $g(x) = y$ , then  $g^{-1}(y)$  is the point set  $\{x\}$ , and we will write  $g^{-1}(y) = x$ . If the random variable  $Y$  is now defined by  $Y = g(X)$ , we can write for any set  $A \subset \mathcal{Y}$ ,

$$\begin{aligned} P(Y \in A) &= P(g(X) \in A) \\ (2.1.2) \quad &= P(\{x \in \mathcal{X}: g(x) \in A\}) \\ &= P(X \in g^{-1}(A)). \end{aligned}$$

This defines the probability distribution of  $Y$ . It is straightforward to show that this probability distribution satisfies the Kolmogorov Axioms.

If  $X$  is a discrete random variable then  $\mathcal{X}$  is countable. The sample space for  $Y = g(X)$  is  $\mathcal{Y} = \{y: y = g(x), x \in \mathcal{X}\}$ , which is also a countable set. Thus,  $Y$  is also a discrete random variable. Using (2.1.2), the pmf for  $Y$  is

$$f_Y(y) = P(Y = y) = \sum_{x \in g^{-1}(y)} P(X = x) = \sum_{x \in g^{-1}(y)} f_X(x), \quad \text{for } y \in \mathcal{Y},$$

and  $f_Y(y) = 0$  for  $y \notin \mathcal{Y}$ . In this case, finding the pmf of  $Y$  involves simply identifying  $g^{-1}(y)$ , for each  $y \in \mathcal{Y}$ , and summing the appropriate probabilities.

**Example 2.1.1:** A discrete random variable  $X$  has a *binomial distribution* if its pmf is of the form

$$f_X(x) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \dots, n,$$

where  $n$  is a positive integer and  $0 \leq p \leq 1$ . Values such as  $n$  and  $p$  that can be set to different values, producing different probability distributions, are called *parameters*. Consider the random variable  $Y = g(X)$ , where  $g(x) = n - x$ . That is,  $Y = n - X$ . Here  $\mathcal{X} = \{0, 1, \dots, n\}$  and  $\mathcal{Y} = \{y: y = g(x), x \in \mathcal{X}\} = \{0, 1, \dots, n\}$ . For any  $y \in \mathcal{Y}$ ,  $n - x = g(x) = y$  if and only if  $x = n - y$ . Thus,  $g^{-1}(y)$  is the single point  $x = n - y$ , and

$$\begin{aligned} f_Y(y) &= \sum_{x \in g^{-1}(y)} f_X(x) \\ &= f_X(n - y) \\ &= \binom{n}{n-y} p^{n-y} (1 - p)^{n-(n-y)} \\ &= \binom{n}{y} (1 - p)^y p^{n-y}. \end{aligned} \quad \begin{array}{l} \text{Definition 1.2.4} \\ \left( \text{implies } \binom{n}{y} = \binom{n}{n-y} \right) \end{array}$$

Thus, we see that  $Y$  also has a binomial distribution, but with parameters  $n$  and  $1 - p$ . ||

If  $X$  and  $Y$  are continuous random variables, then in some cases it is possible to find simple formulas for the cdf and pdf of  $Y$  in terms of the cdf and pdf of  $X$  and the function  $g$ . In the remainder of this section, we consider some of these cases.

The cdf of  $Y = g(X)$  is

$$\begin{aligned}
 F_Y(y) &= P(Y \leq y) \\
 &= P(g(X) \leq y) \\
 (2.1.3) \quad &= P(\{x \in \mathcal{X}: g(x) \leq y\}) \\
 &= \int_{\{x \in \mathcal{X}: g(x) \leq y\}} f_X(x) dx.
 \end{aligned}$$

Sometimes there may be difficulty in identifying  $\{x \in \mathcal{X}: g(x) \leq y\}$  and carrying out the integration of  $f_X(x)$  over this region, as the next example shows.

**Example 2.1.2:** Suppose  $X$  has a uniform distribution on the interval  $(0, 2\pi)$ , that is,

$$f_X(x) = \begin{cases} 1/(2\pi) & 0 < x < 2\pi \\ 0 & \text{otherwise} \end{cases}.$$

Consider  $Y = \sin^2(X)$ . Then (see Figure 2.1.1)

$$(2.1.4) \quad P(Y \leq y) = P(X \leq x_1) + P(x_2 \leq X \leq x_3) + P(X \geq x_4).$$

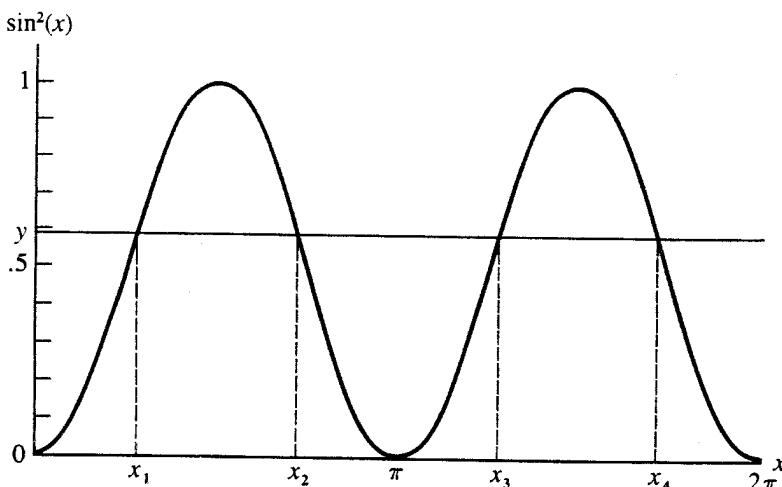


FIGURE 2.1.1 Graph of the transformation  $y = \sin^2(x)$  of Example 2.1.2

From the symmetry of the function  $\sin^2(x)$ , and the fact that  $X$  has a uniform distribution, we have

$$P(X \leq x_1) = P(X \geq x_4) \quad \text{and} \quad P(x_2 \leq X \leq x_3) = 2P(x_2 \leq X \leq \pi),$$

so

$$(2.1.5) \quad P(Y \leq y) = 2P(X \leq x_1) + 2P(x_2 \leq X \leq \pi)$$

where  $x_1$  and  $x_2$  are the two solutions to

$$\sin^2(x) = y, \quad 0 < x < \pi.$$

Thus, even though this example dealt with a seemingly simple situation, the resulting expression for the cdf of  $Y$  was not simple. ||

When making transformations, it is important to keep track of the sample spaces of the random variables; otherwise, much confusion can arise. When making a transformation from  $X$  to  $Y = g(X)$ , it is most convenient to use

$$(2.1.6) \quad \mathcal{X} = \{x: f_X(x) > 0\} \quad \text{and} \quad \mathcal{Y} = \{y: y = g(x) \text{ for some } x \in \mathcal{X}\}.$$

The pdf of the random variable  $X$  is positive only on the set  $\mathcal{X}$  and is zero elsewhere. Such a set is called the *support set* of a distribution or, more informally, the *support* of a distribution. This terminology can also apply to a pmf or, in general, to any nonnegative function.

It is easiest to deal with functions  $g(x)$  that are *monotone*, that is, those that satisfy either

a.  $u > v \Rightarrow g(u) > g(v)$  (increasing)

or

b.  $u < v \Rightarrow g(u) > g(v)$  (decreasing).

If the transformation  $x \rightarrow g(x)$  is monotone, then it is *one-to-one and onto* from  $\mathcal{X} \rightarrow \mathcal{Y}$ . That is, each  $x$  goes to only one  $y$  and each  $y$  comes from at most one  $x$  (one-to-one). Also, for  $\mathcal{Y}$  defined as in (2.1.6), for each  $y \in \mathcal{Y}$  there is an  $x \in \mathcal{X}$  such that  $g(x) = y$  (onto). Thus, the transformation  $g$  uniquely pairs  $xs$  and  $ys$ . If  $g$  is monotone, then  $g^{-1}$  is single-valued, that is,  $g^{-1}(y) = x$  if and only if  $y = g(x)$ . If  $g$  is increasing, this implies that

$$(2.1.7a) \quad \begin{aligned} \{x \in \mathcal{X}: g(x) \leq y\} &= \{x \in \mathcal{X}: g^{-1}(g(x)) \leq g^{-1}(y)\} \\ &= \{x \in \mathcal{X}: x \leq g^{-1}(y)\}. \end{aligned}$$

If  $g$  is decreasing, this implies that

$$(2.1.7b) \quad \begin{aligned} \{x \in \mathcal{X}: g(x) \leq y\} &= \{x \in \mathcal{X}: g^{-1}(g(x)) \geq g^{-1}(y)\} \\ &= \{x \in \mathcal{X}: x \geq g^{-1}(y)\}. \end{aligned}$$

(A graph will illustrate why the inequality reverses in the decreasing case.) If  $g(x)$  is an increasing function, then using (2.1.3), we can write

$$\begin{aligned} F_Y(y) &= \int_{\{x \in \mathcal{X}: x \leq g^{-1}(y)\}} f_X(x) dx \\ &= \int_{-\infty}^{g^{-1}(y)} f_X(x) dx \\ &= F_X(g^{-1}(y)). \end{aligned}$$

If  $g(x)$  is decreasing, we have

$$\begin{aligned} F_Y(y) &= \int_{g^{-1}(y)}^{\infty} f_X(x) dx \\ &= 1 - F_X(g^{-1}(y)). \quad (\text{continuity of } X \text{ is used here}) \end{aligned}$$

We summarize these results in the following theorem.

**THEOREM 2.1.1:** Let  $X$  have cdf  $F_X(x)$ , let  $Y = g(X)$ , and let  $\mathcal{X}$  and  $\mathcal{Y}$  be defined as in (2.1.6).

- a. If  $g$  is an increasing function on  $\mathcal{X}$ ,  $F_Y(y) = F_X(g^{-1}(y))$  for  $y \in \mathcal{Y}$ .
- b. If  $g$  is a decreasing function on  $\mathcal{X}$  and  $X$  is a continuous random variable,  $F_Y(y) = 1 - F_X(g^{-1}(y))$  for  $y \in \mathcal{Y}$ .

**Example 2.1.3:** Suppose  $X \sim f_X(x) = 1$  if  $0 < x < 1$  and 0 otherwise, the uniform(0, 1) distribution. It is straightforward to check that  $F_X(x) = x$ ,  $0 < x < 1$ . We now make the transformation  $Y = g(X) = -\log X$ . Since

$$\frac{d}{dx}g(x) = \frac{d}{dx}(-\log x) = \frac{-1}{x} < 0, \quad \text{for } 0 < x < 1,$$

$g(x)$  is a decreasing function. As  $X$  ranges between 0 and 1,  $-\log x$  ranges between 0 and  $\infty$ , that is,  $\mathcal{Y} = (0, \infty)$ . For  $y > 0$ ,  $y = -\log x$  implies  $x = e^{-y}$ , so  $g^{-1}(y) = e^{-y}$ . Therefore, for  $y > 0$ ,

$$\begin{aligned} F_Y(y) &= 1 - F_X(g^{-1}(y)) \\ &= 1 - F_X(e^{-y}) \\ &= 1 - e^{-y}. \quad (F_X(x) = x) \end{aligned}$$

Of course,  $F_Y(y) = 0$  for  $y \leq 0$ . Note that it was necessary only to verify that  $g(x) = -\log x$  is monotone on  $(0, 1)$ , the support of  $X$ . ||

If the pdf of  $Y$  is continuous, it can be obtained by differentiating the cdf. The resulting expression is given in the following theorem.

**THEOREM 2.1.2:** Let  $X$  have pdf  $f_X(x)$  and let  $Y = g(X)$ , where  $g$  is a monotone function. Let  $\mathcal{X}$  and  $\mathcal{Y}$  be defined by (2.1.6). Suppose that  $f_X(x)$  is continuous on  $\mathcal{X}$  and that  $g^{-1}(y)$  has a continuous derivative on  $\mathcal{Y}$ . Then the pdf of  $Y$  is given by

$$(2.1.8) \quad f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & y \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases}$$

*Proof:* From Theorem 2.1.1 we have, by the chain rule,

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \begin{cases} f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) & \text{if } g \text{ is increasing} \\ -f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) & \text{if } g \text{ is decreasing} \end{cases},$$

which can be expressed concisely as (2.1.8). □

**Example 2.1.4:** Let  $f_X(x)$  be the *gamma pdf*

$$f(x) = \frac{1}{(n-1)! \beta^n} x^{n-1} e^{-x/\beta}, \quad 0 < x < \infty,$$

where  $\beta$  is a positive constant and  $n$  is a positive integer. Suppose we want to find the pdf of  $g(X) = 1/X$ . Note that here the support sets  $\mathcal{X}$  and  $\mathcal{Y}$  are both the interval  $(0, \infty)$ . If we let  $y = g(x)$ , then  $g^{-1}(y) = 1/y$  and  $\frac{d}{dy} g^{-1}(y) = -1/y^2$ . Applying the above theorem, for  $y \in (0, \infty)$ ,

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= \frac{1}{(n-1)! \beta^n} \left( \frac{1}{y} \right)^{n-1} e^{-1/(\beta y)} \frac{1}{y^2} \\ &= \frac{1}{(n-1)! \beta^n} \left( \frac{1}{y} \right)^{n+1} e^{-1/(\beta y)}, \end{aligned}$$

a special case of a pdf known as the *inverted gamma pdf*. ||

In many applications, the function  $g$  may be neither increasing nor decreasing, hence the above results will not apply. However, it is often the case that  $g$  will be monotone over certain intervals and that allows us to get an expression for  $Y = g(X)$ . (If  $g$  is not monotone over certain intervals then we are in *deep trouble*.)

**Example 2.1.5:** Suppose  $X$  is a continuous random variable. For  $y > 0$ , the cdf of  $Y = X^2$  is

$$\begin{aligned}
 F_Y(y) &= P(Y \leq y) \\
 &= P(X^2 \leq y) \\
 &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\
 &= P(-\sqrt{y} < X \leq \sqrt{y}) \quad (\text{continuity of } X) \\
 &= P(X \leq \sqrt{y}) - P(X \leq -\sqrt{y}) \\
 &= F_X(\sqrt{y}) - F_X(-\sqrt{y}).
 \end{aligned}$$

The pdf of  $Y$  can now be obtained from the cdf by differentiation:

$$\begin{aligned}
 f_Y(y) &= \frac{d}{dy} F_Y(y) \\
 &= \frac{d}{dy} [F_X(\sqrt{y}) - F_X(-\sqrt{y})], \\
 &= \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}),
 \end{aligned}$$

where we use the chain rule to differentiate  $F_X(\sqrt{y})$  and  $F_X(-\sqrt{y})$ . Therefore, the pdf is

$$(2.1.9) \quad f_Y(y) = \frac{1}{2\sqrt{y}} (f_X(\sqrt{y}) + f_X(-\sqrt{y})). \quad ||$$

Notice that the pdf of  $Y$  in (2.1.9) is expressed as the sum of two pieces, pieces that represent the intervals where  $g(x) = x^2$  is monotone. In general, this will be the case.

**THEOREM 2.1.3:** Let  $X$  have pdf  $f_X(x)$ , let  $Y = g(X)$ , and define the sample space  $\mathcal{X}$  as in (2.1.6). Suppose there exists a partition,  $A_0, A_1, \dots, A_k$ , of  $\mathcal{X}$  such that  $P(X \in A_0) = 0$  and  $f_X(x)$  is continuous on each  $A_i$ . Further, suppose there exist functions  $g_1(x), \dots, g_k(x)$ , defined on  $A_1, \dots, A_k$ , respectively, satisfying

- a.  $g(x) = g_i(x)$ , for  $x \in A_i$ ,
- b.  $g_i(x)$  is monotone on  $A_i$ ,
- c. the set  $\mathcal{Y} = \{y: y = g_i(x) \text{ for some } x \in A_i\}$  is the same for each  $i = 1, \dots, k$ , and
- d.  $g_i^{-1}(y)$  has a continuous derivative on  $\mathcal{Y}$ , for each  $i = 1, \dots, k$ .

Then

$$f_Y(y) = \begin{cases} \sum_{i=1}^k f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right| & y \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases}. \quad \square$$

The important point in Theorem 2.1.3 is that  $\mathcal{X}$  can be divided into sets  $A_1, \dots, A_k$  such that  $g(x)$  is monotone on each  $A_i$ . We can ignore the “exceptional set”  $A_0$  since  $P(X \in A_0) = 0$ . It is a technical device that is used, for example, to handle endpoints of intervals. It is important to note that each  $g_i(x)$  is a one-to-one transformation from  $A_i$  onto  $\mathcal{Y}$ . Furthermore,  $g_i^{-1}(y)$  is a one-to-one function from  $\mathcal{Y}$  onto  $A_i$  such that, for  $y \in \mathcal{Y}$ ,  $g_i^{-1}(y)$  gives the unique  $x = g_i^{-1}(y) \in A_i$  for which  $g_i(x) = y$ .

**Example 2.1.6:** Let  $X$  have the *standard normal distribution*,

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty.$$

Consider  $Y = X^2$ . The function  $g(x) = x^2$  is monotone on  $(-\infty, 0)$  and on  $(0, \infty)$ . The set  $\mathcal{Y} = (0, \infty)$ . Applying Theorem 2.1.3, we take

$$A_0 = \{0\};$$

$$A_1 = (-\infty, 0), \quad g_1(x) = x^2, \quad g_1^{-1}(y) = -\sqrt{y};$$

$$A_2 = (0, \infty), \quad g_2(x) = x^2, \quad g_2^{-1}(y) = \sqrt{y}.$$

The pdf of  $Y$  is

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{y})^2/2} \left| -\frac{1}{2\sqrt{y}} \right| + \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2/2} \left| \frac{1}{2\sqrt{y}} \right| \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2}, \quad 0 < y < \infty. \end{aligned}$$

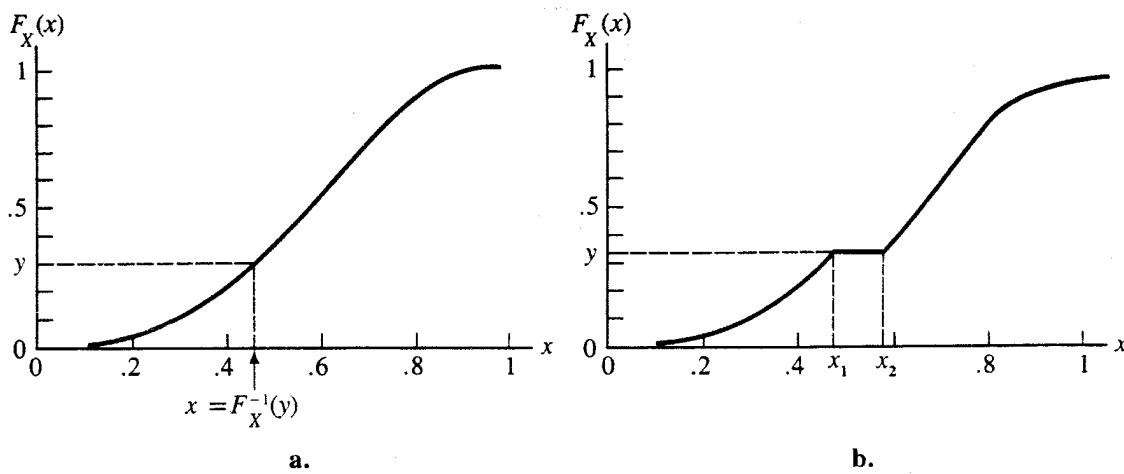
The pdf of  $Y$  is one that we will often encounter, that of a *chi squared random variable* with 1 degree of freedom. ||

We close this section with a special and very useful transformation, the *probability integral transformation*.

**THEOREM 2.1.4:** Let  $X$  have continuous cdf  $F_X(x)$  and define the random variable  $Y$  as  $Y = F_X(X)$ . Then  $Y$  is uniformly distributed on  $(0, 1)$ , that is,  $P(Y \leq y) = y$ ,  $0 < y < 1$ .

Before we prove this theorem, we will digress for a moment and look at  $F_X^{-1}$ , the inverse of the cdf  $F_X$ , in some detail. If  $F_X$  is strictly increasing, then  $F_X^{-1}$  is well defined by

$$(2.1.10) \quad F_X^{-1}(y) = x \Leftrightarrow F_X(x) = y.$$

FIGURE 2.1.2 (a)  $F(x)$  strictly increasing; (b)  $F(x)$  nondecreasing

However, if  $F_X$  is constant on some interval, then  $F_X^{-1}$  is not well defined by (2.1.10), as Figure 2.1.2 illustrates. Any  $x$  satisfying  $x_1 \leq x \leq x_2$  satisfies  $F_X(x) = y$ .

This problem is avoided by defining  $F_X^{-1}(y)$  for  $0 < y < 1$ , by

$$(2.1.11) \quad F_X^{-1}(y) = \inf\{x: F_X(x) \geq y\},$$

a definition that agrees with (2.1.10) when  $F_X$  is nonconstant and provides an  $F_X^{-1}$  which is single-valued even when  $F_X$  is not strictly increasing. Using this definition, in Figure 2.1.2b, we have  $F_X^{-1}(y) = x_1$ . At the endpoints of the range of  $y$ ,  $F_X^{-1}(y)$  can also be defined.  $F_X^{-1}(1) = \infty$  if  $F_X(x) < 1$  for all  $x$  and, for any  $F_X$ ,  $F_X^{-1}(0) = -\infty$ .

*Proof of Theorem 2.1.4:* For  $Y = F_X(X)$  we have, for  $0 < y < 1$ ,

$$\begin{aligned} P(Y \leq y) &= P(F_X(X) \leq y) \\ &= P(F_X^{-1}[F_X(X)] \leq F_X^{-1}(y)) && (F_X^{-1} \text{ is increasing}) \\ &= P(X \leq F_X^{-1}(y)) && (\text{see paragraph below}) \\ &= F_X(F_X^{-1}(y)) && (\text{definition of } F_X) \\ &= y. && (\text{continuity of } F_X) \end{aligned}$$

At the endpoints we have  $P(Y \leq y) = 1$  for  $y \geq 1$  and  $P(Y \leq y) = 0$  for  $y \leq 0$ , showing that  $Y$  has a uniform distribution.

The reasoning behind the equality

$$P(F_X^{-1}(F_X(X)) \leq F_X^{-1}(y)) = P(X \leq F_X^{-1}(y))$$

is somewhat subtle and deserves additional attention. If  $F_X$  is strictly increasing, then it is true that  $F_X^{-1}(F_X(x)) = x$ . (Refer to Figure 2.1.2a.) However, if  $F_X$  is flat, it may be that  $F_X^{-1}(F_X(x)) \neq x$ . Suppose  $F_X$  is as in Figure 2.1.2b and let  $x \in [x_1, x_2]$ . Then  $F_X^{-1}(F_X(x)) = x_1$  for any  $x$  in this interval. Even in this case, though,

probability equality holds, since  $P(X \leq x) = P(X \leq x_1)$  for any  $x \in [x_1, x_2]$ . The flat cdf denotes a region of 0 probability ( $P(x_1 < X \leq x) = F_X(x) - F_X(x_1) = 0$ ).  $\square$

One application of Theorem 2.1.4 is in the generation of random samples from a particular distribution. If it is required to generate an observation  $x$  from a population with cdf  $F_X$ , we need only generate a uniform random number  $u$ , between 0 and 1, and solve for  $x$  in the equation  $F_X(x) = u$ . (For many distributions there are other methods of generating observations that take less computer time, but this method is still useful because of its general applicability.)

## 2.2 Expected Values

The expected value, or expectation, of a random variable is merely its average value, where we speak of “average” value as one that is weighted according to the probability distribution. The expected value of a distribution can be thought of as a measure of center, as we think of averages as being middle values. By weighting the values of the random variable according to the probability distribution, we hope to obtain a number that summarizes a typical or expected value of an observation of the random variable.

**DEFINITION 2.2.1:** The *expected value* or *mean* of a random variable  $g(X)$ , denoted by  $Eg(X)$ , is

(2.2.1)

$$Eg(X) = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x) dx & \text{if } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} g(x) f_X(x) = \sum_{x \in \mathcal{X}} g(x) P(X = x) & \text{if } X \text{ is discrete} \end{cases}$$

provided that the integral or sum exists. If  $E|g(X)| = \infty$ , we say that  $Eg(X)$  does not exist.

**Example 2.2.1:** Suppose  $X$  has an *exponential* ( $\lambda$ ) *distribution*, that is, it has pdf given by

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad 0 \leq x < \infty, \quad \lambda > 0.$$

Then  $EX$  is given by

$$\begin{aligned} EX &= \int_0^{\infty} \frac{1}{\lambda} x e^{-x/\lambda} dx \\ &= -xe^{-x/\lambda} \Big|_0^{\infty} + \int_0^{\infty} e^{-x/\lambda} dx \quad (\text{integration by parts}) \\ &= \int_0^{\infty} e^{-x/\lambda} dx = \lambda \end{aligned}$$

||

**Example 2.2.2:** If  $X$  has a *binomial distribution*, its pmf is given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n$$

where  $n$  is a positive integer,  $0 \leq p \leq 1$ , and for every fixed pair  $n$  and  $p$  the pmf sums to 1. The expected value of a binomial random variable is given by

$$EX = \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^n x \binom{n}{x} p^x (1-p)^{n-x}$$

( $x = 0$  term is 0). Using the identity  $x \binom{n}{x} = n \binom{n-1}{x-1}$ , we have

$$\begin{aligned} EX &= \sum_{x=1}^n n \binom{n-1}{x-1} p^x (1-p)^{n-x} \\ &= \sum_{y=0}^{n-1} n \binom{n-1}{y} p^{y+1} (1-p)^{n-(y+1)} \quad (\text{substitute } y = x-1) \\ &= np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{n-1-y} \\ &= np, \end{aligned}$$

since the last summation must be 1, being the sum over all possible values of a binomial( $n - 1$ ,  $p$ ) pmf. ||

**Example 2.2.3:** A classic example of a random variable whose expected value does not exist is a *Cauchy random variable*, that is, one with pdf

$$f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \quad -\infty < x < \infty.$$

It is straightforward to check that  $\int_{-\infty}^{\infty} f_X(x) dx = 1$ , but  $E|X| = \infty$ . Write

$$E|X| = \int_{-\infty}^{\infty} \frac{|x|}{\pi} \frac{1}{1+x^2} dx = \frac{2}{\pi} \int_0^{\infty} \frac{x}{1+x^2} dx.$$

For any positive number  $M$ ,

$$\int_0^M \frac{x}{1+x^2} dx = \frac{\log(1+x^2)}{2} \Big|_0^M = \frac{\log(1+M^2)}{2}.$$

Thus,

$$E|X| = \lim_{M \rightarrow \infty} \frac{2}{\pi} \int_0^M \frac{x}{1+x^2} dx = \frac{1}{\pi} \lim_{M \rightarrow \infty} \log(1+M^2) = \infty$$

and  $EX$  does not exist. ||

The process of taking expectations is a linear operation, which means that the expectation of a linear function of  $X$  can be easily evaluated by noting that for any constants  $a$  and  $b$ ,

$$(2.2.2) \quad E(aX + b) = aEX + b.$$

For example, if  $X$  is binomial( $n, p$ ), so  $EX = np$ , then

$$E(X - np) = EX - np = np - np = 0.$$

The expectation operator, in fact, has many properties that can help ease calculational effort. Most of these properties follow from the properties of the integral or sum, and are summarized in the following theorem.

**THEOREM 2.2.1:** Let  $X$  be a random variable and let  $a, b$ , and  $c$  be constants. Then for any functions  $g_1(x)$  and  $g_2(x)$  whose expectations exist,

- a.  $E(ag_1(X) + bg_2(X) + c) = aEg_1(X) + bEg_2(X) + c.$
- b. If  $g_1(x) \geq 0$  for all  $x$ , then  $Eg_1(X) \geq 0.$
- c. If  $g_1(x) \geq g_2(x)$  for all  $x$ , then  $Eg_1(X) \geq Eg_2(X).$
- d. If  $a \leq g_1(x) \leq b$  for all  $x$ , then  $a \leq Eg_1(X) \leq b.$

*Proof:* We will give details only for the continuous case, the discrete case being similar. By definition,

$$\begin{aligned} E(ag_1(X) + bg_2(X) + c) &= \int_{-\infty}^{\infty} (ag_1(x) + bg_2(x) + c)f_X(x) dx \\ &= \int_{-\infty}^{\infty} ag_1(x)f_X(x) dx + \int_{-\infty}^{\infty} bg_2(x)f_X(x) dx + \int_{-\infty}^{\infty} cf_X(x) dx \end{aligned}$$

by the additivity of the integral. Since  $a, b$ , and  $c$  are constants, they factor out of their respective integrals and we have

$$\begin{aligned} E(ag_1(X) + bg_2(X) + c) &= a \int_{-\infty}^{\infty} g_1(x)f_X(x) dx + b \int_{-\infty}^{\infty} g_2(x)f_X(x) dx + c \int_{-\infty}^{\infty} f_X(x) dx \\ &= aEg_1(X) + bEg_2(X) + c, \end{aligned}$$

establishing (a). The other three properties are proved in a similar manner. □

**Example 2.2.4:** The expected value of a random variable has another property, one that we can think of as relating to the interpretation of  $EX$  as a good guess at a value of  $X$ .

Suppose we measure the distance between a random variable  $X$  and a constant  $b$  by  $(X - b)^2$ . The closer  $b$  is to  $X$ , the smaller this quantity is. We can now determine the value of  $b$  that minimizes  $E(X - b)^2$  and, hence, will provide us with a good predictor of  $X$ . (Note that it does no good to look for a value of  $b$  that minimizes  $(X - b)^2$ , since the answer would depend on  $X$ , making it a useless predictor of  $X$ .)

We could proceed with the minimization of  $E(X - b)^2$  by using calculus, but there is a simpler method. (See Exercise 2.20 for a calculus-based proof.) Using the belief that there is something special about  $EX$ , write

$$\begin{aligned} E(X - b)^2 &= E(X - EX + EX - b)^2 && \left( \begin{array}{l} \text{add } \pm EX, \text{ which} \\ \text{changes nothing} \end{array} \right) \\ &= E((X - EX) + (EX - b))^2 && (\text{group terms}) \\ &= E(X - EX)^2 + (EX - b)^2 + 2E((X - EX)(EX - b)), \end{aligned}$$

where we have expanded the square. Now, note that

$$E((X - EX)(EX - b)) = (EX - b)E(X - EX) = 0,$$

since  $(EX - b)$  is constant and comes out of the expectation, and  $E(X - EX) = EX - EX = 0$ . This means that

$$(2.2.3) \quad E(X - b)^2 = E(X - EX)^2 + (EX - b)^2.$$

We have no control over the first term on the right-hand side of (2.2.3) and the second term, which is always greater than or equal to 0, can be made equal to 0 by choosing  $b = EX$ . Hence,

$$(2.2.4) \quad \min_b E(X - b)^2 = E(X - EX)^2.$$

See Exercise 2.19 for a similar result about the median. ||

When evaluating expectations of nonlinear functions of  $X$ , we can proceed in one of two ways. From the definition of  $Eg(X)$ , we could directly calculate

$$(2.2.5) \quad Eg(X) = \int_{-\infty}^{\infty} g(x)f_X(x)dx.$$

But we could also find the pdf  $f_Y(y)$  of  $Y = g(X)$  and we would have

$$(2.2.6) \quad Eg(X) = EY = \int_{-\infty}^{\infty} yf_Y(y)dy.$$

**Example 2.2.5:** Let  $X$  have a uniform(0, 1) distribution, that is, the pdf of  $X$  is given by

$$f_X(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases},$$

and define a new random variable  $g(X) = -\log X$ . Then

$$\begin{aligned} \mathbb{E}g(X) &= \mathbb{E}(-\log X) = \int_0^1 -\log x \, dx \\ &= x - x \log x \Big|_0^1 \\ &= 1. \end{aligned}$$

But we also saw in Example 2.1.3 that  $Y = -\log X$  has cdf  $1 - e^{-y}$  and, hence, pdf  $f_Y(y) = \frac{d}{dy}(1 - e^{-y}) = e^{-y}$ ,  $0 < y < \infty$ , which is a special case of the exponential pdf with  $\lambda = 1$ . Thus, by Example 2.2.1,  $\mathbb{E}Y = 1$ . ||

## 2.3 Moments and Moment Generating Functions

The various moments of a distribution are an important class of expectations.

**DEFINITION 2.3.1:** For each integer  $n$ , the *n*th moment of  $X$  (or  $F_X(x)$ ),  $\mu'_n$ , is

$$\mu'_n = \mathbb{E}X^n.$$

The *n*th central moment of  $X$ ,  $\mu_n$ , is

$$\mu_n = \mathbb{E}(X - \mu)^n,$$

where  $\mu = \mu'_1 = \mathbb{E}X$ .

Aside from the mean,  $\mathbb{E}X$ , of a random variable, perhaps the most important moment is the second central moment, more commonly known as the variance.

**DEFINITION 2.3.2:** The *variance* of a random variable  $X$  is its second central moment,  $\text{Var } X = \mathbb{E}(X - \mathbb{E}X)^2$ . The positive square root of  $\text{Var } X$  is the *standard deviation* of  $X$ .

The variance gives a measure of the degree of spread of a distribution around its mean. We saw earlier in Example 2.2.4 that the quantity  $\mathbb{E}(X - b)^2$  is minimized by choosing  $b = \mathbb{E}X$ . Now we consider the absolute size of this minimum. The interpretation attached to the variance is that larger values mean  $X$  is more variable. At the extreme, if  $\text{Var } X = \mathbb{E}(X - \mathbb{E}X)^2 = 0$ , then  $X$  is equal to  $\mathbb{E}X$ , with probability 1, and there is no variation in  $X$ . The standard deviation has the same qualitative interpretation: Small values mean  $X$  is very likely to be close to  $\mathbb{E}X$ , and large

values mean  $X$  is very variable. The standard deviation is easier to interpret in that the measurement unit on the standard deviation is the same as that for the original variable  $X$ . The measurement unit on the variance is the square of the original unit.

**Example 2.3.1:** Let  $X$  have the exponential( $\lambda$ ) distribution, defined in Example 2.2.1. There we calculated  $EX = \lambda$ , and we can now calculate the variance by

$$\begin{aligned}\text{Var } X &= E(X - \lambda)^2 = \int_0^\infty (x - \lambda)^2 \frac{1}{\lambda} e^{-x/\lambda} dx \\ &= \int_0^\infty (x^2 - 2x\lambda + \lambda^2) \frac{1}{\lambda} e^{-x/\lambda} dx.\end{aligned}$$

To complete the integration, we can integrate each of the terms separately, using integration by parts on the terms involving  $x$  and  $x^2$ . Upon doing this, we find that  $\text{Var } X = \lambda^2$ . ||

We see that the variance of an exponential distribution is directly related to the parameter  $\lambda$ . Figure 2.3.1 shows several exponential distributions corresponding to different values of  $\lambda$ . Notice how the distribution is more concentrated about its mean for smaller values of  $\lambda$ . The behavior of the variance of an exponential, as a function of  $\lambda$ , is a special case of the variance behavior summarized in the following theorem.

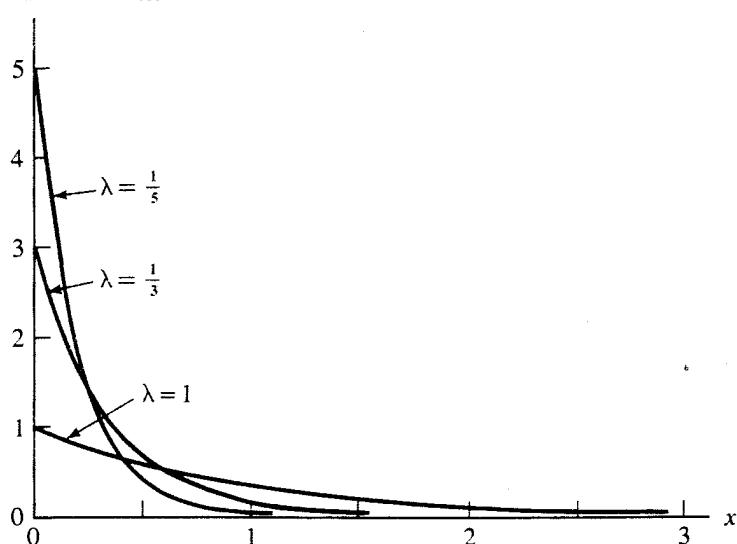


FIGURE 2.3.1 Exponential densities for  $\lambda = 1, \frac{1}{3}, \frac{1}{5}$

**THEOREM 2.3.1:** If  $X$  is a random variable with finite variance, then for any constants  $a$  and  $b$ ,

$$\text{Var}(aX + b) = a^2 \text{Var } X.$$

*Proof:* From the definition, we have

$$\begin{aligned}
 \text{Var}(aX + b) &= E((aX + b) - E(aX + b))^2 \\
 &= E(aX - aEX)^2 \quad (E(aX + b) = a(EX) + b) \\
 &= a^2 E(X - EX)^2 \\
 &= a^2 \text{Var } X. \quad \square
 \end{aligned}$$

It is sometimes easier to use an alternative formula for the variance, given by

$$(2.3.1) \quad \text{Var } X = EX^2 - (EX)^2,$$

which is easily established by noting

$$\begin{aligned}
 \text{Var } X &= E(X - EX)^2 = E[X^2 - 2XEX + (EX)^2] \\
 &= EX^2 - 2(EX)^2 + (EX)^2 \\
 &= EX^2 - (EX)^2,
 \end{aligned}$$

where we use the fact that  $E(XEX) = (EX)(EX) = (EX)^2$ , since  $EX$  is a constant. We now illustrate some moment calculations with a discrete distribution.

**Example 2.3.2:** Let  $X \sim \text{binomial}(n, p)$ , that is,

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n.$$

We have previously seen that  $EX = np$ . To calculate  $\text{Var } X$  we first calculate  $EX^2$ . We have

$$(2.3.2) \quad EX^2 = \sum_{x=0}^n x^2 \binom{n}{x} p^x (1-p)^{n-x}.$$

In order to sum this series, we must first manipulate the binomial coefficient in a manner similar to that used for  $EX$  (Example 2.2.2). We write

$$(2.3.3) \quad x^2 \binom{n}{x} = x \frac{n!}{(x-1)!(n-x)!} = xn \binom{n-1}{x-1}.$$

The summand in (2.3.2) corresponding to  $x = 0$  is zero, and using (2.3.3) we have

$$\begin{aligned}
 EX^2 &= n \sum_{x=1}^n x \binom{n-1}{x-1} p^x (1-p)^{n-x} \\
 &= n \sum_{y=0}^{n-1} (y+1) \binom{n-1}{y} p^{y+1} (1-p)^{n-1-y} \quad (\text{setting } y = x-1)
 \end{aligned}$$

$$= np \sum_{y=0}^{n-1} y \binom{n-1}{y} p^y (1-p)^{n-1-y} + np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{n-1-y}.$$

Now it is easy to see that the first sum is equal to  $(n-1)p$  (since it is the mean of a binomial( $n-1, p$ )), while the second sum is equal to 1. Hence,

$$(2.3.4) \quad EX^2 = n(n-1)p^2 + np.$$

Using (2.3.1), we have

$$\begin{aligned} \text{Var } X &= n(n-1)p^2 + np - (np)^2 \\ &= -np^2 + np \\ &= np(1-p). \end{aligned} \quad ||$$

Calculation of higher moments proceeds in an analogous manner, but usually the mathematical manipulations become quite involved. In applications, moments of order 3 or 4 are sometimes of interest, but there is usually little statistical reason for examining higher moments than these.

We now introduce a new function that is associated with a probability distribution, the *moment generating function* (mgf). As its name suggests, the mgf can be used to generate moments. In practice, it is easier in many cases to calculate moments directly than to use the mgf. However, the main use of the mgf is not to generate moments, but to help in characterizing a distribution. This property can lead to some extremely powerful results when used properly.

**DEFINITION 2.3.3:** Let  $X$  be a random variable with cdf  $F_X$ . The *moment generating function (mgf) of  $X$*  (or  $F_X$ ), denoted by  $M_X(t)$ , is

$$M_X(t) = Ee^{tX},$$

provided that the expectation exists for  $t$  in some neighborhood of 0. That is, there is an  $h > 0$  such that, for all  $t$  in  $-h < t < h$ ,  $Ee^{tX}$  exists. If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

More explicitly, we can write the mgf of  $X$  as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \quad \text{if } X \text{ is continuous.}$$

or

$$M_X(t) = \sum_x e^{tx} P(X = x) \quad \text{if } X \text{ is discrete.}$$

It is very easy to see how the mgf generates moments. We summarize the result in the following theorem.

**THEOREM 2.3.2:** If  $X$  has mgf  $M_X(t)$ , then

$$EX^n = M_X^{(n)}(0),$$

where we define

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)|_{t=0}.$$

That is, the  $n$ th moment is equal to the  $n$ th derivative of  $M_X(t)$  evaluated at  $t = 0$ .

*Proof:* Assuming that we can differentiate under the integral sign (see the next section), we have

$$\begin{aligned} \frac{d}{dt} M_X(t) &= \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \\ &= \int_{-\infty}^{\infty} \left( \frac{d}{dt} e^{tx} \right) f_X(x) dx \\ &= \int_{-\infty}^{\infty} (xe^{tx}) f_X(x) dx \\ &= EXe^{tX}. \end{aligned}$$

Thus,

$$\frac{d}{dt} M_X(t)|_{t=0} = EXe^{tX}|_{t=0} = EX.$$

Proceeding in an analogous manner, we can establish that

$$\frac{d^n}{dt^n} M_X(t)|_{t=0} = EX^n e^{tX}|_{t=0} = EX^n.$$
□

**Example 2.3.3:** In Example 2.1.4 we encountered a special case of the gamma pdf

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad 0 < x < \infty, \quad \alpha > 0, \quad \beta > 0,$$

where  $\Gamma(\alpha)$  denotes the gamma function, some of whose properties are given in Section 3.2. The mgf is given by

$$M_X(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^{\infty} e^{tx} x^{\alpha-1} e^{-x/\beta} dx$$

$$(2.3.5) \quad \begin{aligned} &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1} e^{-(\frac{1}{\beta}-t)x} dx \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1} e^{-x/(\frac{1}{\beta}-t)} dx. \end{aligned}$$

We now recognize the integrand in (2.3.5) as the *kernel* of another gamma pdf. (The *kernel* of a function is the main part of the function, the part that remains when constants are disregarded.) Using the fact that, for any positive constants  $a$  and  $b$ ,

$$f(x) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-x/b}$$

is a pdf, we have that

$$\int_0^\infty \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-x/b} dx = 1$$

and, hence,

$$(2.3.6) \quad \int_0^\infty x^{a-1} e^{-x/b} dx = \Gamma(a)b^a.$$

Applying (2.3.6) to (2.3.5), we have

$$M_X(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \Gamma(\alpha) \left( \frac{\beta}{1-\beta t} \right)^\alpha = \left( \frac{1}{1-\beta t} \right)^\alpha \quad \text{if } t < \frac{1}{\beta}.$$

If  $t \geq 1/\beta$ , then the quantity  $(1/\beta) - t$ , in the integrand of (2.3.5), is nonpositive and the integral in (2.3.6) is infinite. Thus, the mgf of the gamma distribution exists only if  $t < 1/\beta$ . (In Section 3.2 we will explore the gamma function in more detail.)

The mean of the gamma distribution is given by

$$EX = \frac{d}{dt} M_X(t)|_{t=0} = \frac{\alpha\beta}{(1-\beta t)^{\alpha+1}} \Big|_{t=0} = \alpha\beta.$$

Other moments can be calculated in a similar manner. ||

**Example 2.3.4:** For a second illustration of calculating a moment generating function, we consider a discrete distribution, the binomial distribution. The binomial( $n, p$ ) pmf is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

so

$$\begin{aligned} M_X(t) &= \sum_{x=0}^n e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^n \binom{n}{x} (pe^t)^x (1-p)^{n-x}. \end{aligned}$$

The binomial formula (see Theorem 3.1.1) gives

$$(2.3.7) \quad \sum_{x=0}^n \binom{n}{x} u^x v^{n-x} = (u+v)^n.$$

Hence, letting  $u = pe^t$  and  $v = 1 - p$ , we have

$$M_X(t) = [pe^t + (1-p)]^n. \quad ||$$

As previously mentioned, the major usefulness of the moment generating function is not in its ability to generate moments. Rather, its usefulness stems from the fact that, in many cases, the moment generating function can characterize a distribution. There are, however, some technical difficulties associated with using moments to characterize a distribution, which we will now investigate.

If the mgf exists, it characterizes an infinite set of moments. The natural question is whether characterizing the infinite set of moments uniquely determines a distribution function. The answer to this question, unfortunately, is no. Characterizing the set of moments is not enough to determine a distribution uniquely because there may be two distinct random variables having the same moments.

**Example 2.3.5:** Consider the two pdfs given by

$$\begin{aligned} f_1(x) &= \frac{1}{\sqrt{2\pi x}} e^{-(\log x)^2/2}, \quad 0 \leq x < \infty, \\ f_2(x) &= f_1(x)[1 + \sin(2\pi \log x)], \quad 0 \leq x < \infty. \end{aligned}$$

(The pdf  $f_1$  is a special case of a *lognormal pdf*.)

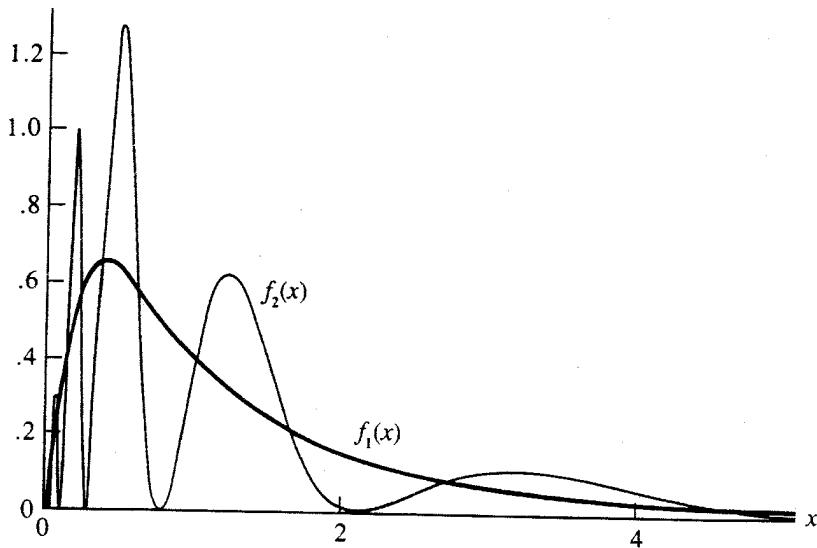
It can be shown that if  $X_1 \sim f_1(x)$ , then

$$EX_1^r = e^{r^2/2}, \quad r = 0, 1, \dots,$$

so  $X_1$  has all of its moments. Now suppose that  $X_2 \sim f_2(x)$ . We have

$$\begin{aligned} EX_2^r &= \int_0^\infty x^r f_2(x) dx \\ &= \int_0^\infty x^r f_1(x)[1 + \sin(2\pi \log x)] dx \\ &= EX_1^r + \int_0^\infty x^r f_1(x) \sin(2\pi \log x) dx. \end{aligned}$$

However, the transformation  $y = \log x - r$  shows that this last integral is that of an odd function over  $(-\infty, \infty)$  and hence is equal to 0 for  $r = 0, 1, \dots$ . Thus, even though  $X_1$  and  $X_2$  have distinct pdfs, they have the same moments for all  $r$ . The two pdfs are pictured in Figure 2.3.2.



**FIGURE 2.3.2** Two pdfs with the same moments:

$$f_1(x) = \frac{1}{\sqrt{2\pi x}} e^{-(\log x)^2/2} \quad \text{and} \quad f_2(x) = f_1(x)[1 + \sin(2\pi \log x)]$$

See Exercise 2.34 for details, and also Exercises 2.35 and 2.36 for more about mgfs and distributions. Also, the *Miscellanea* section gives conditions that guarantee uniqueness of moments. ||

The problem of uniqueness of moments does not occur if the cdfs have bounded support. If that is the case, then the infinite sequence of moments does uniquely determine the distribution (see, for example, Chung (1974) or Feller (1971)). Furthermore, if the mgf exists in a neighborhood of zero, then the distribution is uniquely determined, no matter what its support. Thus, existence of all moments is not equivalent to existence of the moment generating function. The following theorem shows how a distribution can be characterized.

**THEOREM 2.3.3:** Let  $F_X(x)$  and  $F_Y(y)$  be two cdfs all of whose moments exist.

- a. If  $F_X$  and  $F_Y$  have bounded support, then  $F_X(u) = F_Y(u)$  for all  $u$  if and only if  $EX^r = EY^r$  for all integers  $r = 0, 1, 2, \dots$
- b. If the moment generating functions exist and  $M_X(t) = M_Y(t)$  for all  $t$  in some neighborhood of 0, then  $F_X(u) = F_Y(u)$  for all  $u$ . □

In the next theorem, which deals with a sequence of mgfs that converges, we do not treat the bounded support case separately. Note that the uniqueness assumption is automatically satisfied if the limiting mgf exists in a neighborhood of 0 (see the *Miscellanea*).

**THEOREM 2.3.4 (Convergence of mgfs):** Suppose  $\{X_i, i = 1, 2, \dots\}$  is a sequence of random variables, each with mgf  $M_{X_i}(t)$ . Furthermore, suppose that

$$\lim_{i \rightarrow \infty} M_{X_i}(t) = M_X(t), \quad \text{for all } t \text{ in a neighborhood of 0,}$$

and  $M_X(t)$  is an mgf. Then there is a unique cdf  $F_X$  whose moments are determined by  $M_X(t)$  and, for all  $x$  where  $F_X(x)$  is continuous, we have

$$\lim_{i \rightarrow \infty} F_{X_i}(x) = F_X(x).$$

That is, convergence, for  $|t| < h$ , of mgfs to an mgf implies convergence of cdfs.  $\square$

The proofs of Theorems 2.3.3 and 2.3.4 rely on the theory of *Laplace transforms*. (The classic reference is Widder (1946), but Laplace transforms also get a comprehensive treatment by Feller (1971).) The defining equation for  $M_X(t)$ , that is,

$$(2.3.8) \quad M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx,$$

defines a Laplace transform ( $M_X(t)$  is the Laplace transform of  $f_X(x)$ ). A key fact about Laplace transforms is their uniqueness. If (2.3.8) is valid for all  $t$  such that  $|t| < h$ , where  $h$  is some positive number, then given  $M_X(t)$  there is only one function  $f_X(x)$  that satisfies (2.3.8). Given this fact, the two previous theorems are quite reasonable. While rigorous proofs of these theorems are not beyond the scope of this book, the proofs are technical in nature, and shed no real understanding. We omit them.

The possible nonuniqueness of the moment sequence is an annoyance. If we show that a sequence of moments converges, we will not be able to conclude formally that the random variables converge. To do so, we would have to verify the uniqueness of the moment sequence, a generally horrible job (see the *Miscellanea*). However, if the sequence of mgfs converges in a neighborhood of 0, then the random variables converge. Thus, we can consider the convergence of mgfs as a sufficient, but not necessary, condition for the sequence of random variables to converge.

**Example 2.3.6:** One approximation that is usually taught in elementary statistics courses is that binomial probabilities (see Example 2.3.2) can be approximated by *Poisson* probabilities, which are generally easier to calculate. The binomial distribution is characterized by two quantities, denoted by  $n$  and  $p$ . It is taught that the Poisson approximation is valid “when  $n$  is large and  $np$  is small” and rules of thumb are sometimes given.

The  $\text{Poisson}(\lambda)$  pmf is given by

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots,$$

where  $\lambda$  is a positive constant. The approximation states that if  $X \sim \text{binomial}(n, p)$  and  $Y \sim \text{Poisson}(\lambda)$ , with  $\lambda = np$ , then

$$(2.3.9) \quad P(X = x) \approx P(Y = x)$$

for large  $n$  and small  $np$ . We now show that the mgfs converge, lending credence to this approximation. Recall that

$$(2.3.10) \quad M_X(t) = [pe^t + (1 - p)]^n.$$

For the Poisson( $\lambda$ ) distribution, we can calculate (see Exercise 2.37)

$$M_Y(t) = e^{\lambda(e^t - 1)},$$

and if we define  $p = \lambda/n$ , then  $M_X(t) \rightarrow M_Y(t)$  as  $n \rightarrow \infty$ . The validity of the approximation in (2.3.9) will then follow from Theorem 2.3.4.

We first must digress a bit and mention an important limit result, one that has wide applicability in statistics.

**LEMMA 2.3.1:** Let  $a_1, a_2, \dots$  be a sequence of numbers converging to  $a$ , that is,  $\lim_{n \rightarrow \infty} a_n = a$ . Then

$$\lim_{n \rightarrow \infty} \left(1 + \frac{a_n}{n}\right)^n = e^a.$$

*Proof:* The proof of this lemma may be found in many standard calculus texts.  $\square$

Returning to the example, we have

$$\begin{aligned} M_X(t) &= [pe^t + (1 - p)]^n \\ &= \left[1 + \frac{1}{n}(e^t - 1)(np)\right]^n \\ &= \left[1 + \frac{1}{n}(e^t - 1)\lambda\right]^n. \quad (\text{since } \lambda = np) \end{aligned}$$

Now set  $a_n = a = (e^t - 1)\lambda$ , and apply Lemma 2.3.1 to get

$$\lim_{n \rightarrow \infty} M_X(t) = e^{\lambda(e^t - 1)} = M_Y(t),$$

the moment generating function of the Poisson.  $\parallel$

The Poisson approximation can be quite good even for moderate  $p$  and  $n$ . In Figure 2.3.3 (page 68) we show a binomial mass function along with its Poisson approximation, with  $\lambda = np$ . The approximation appears to be satisfactory.

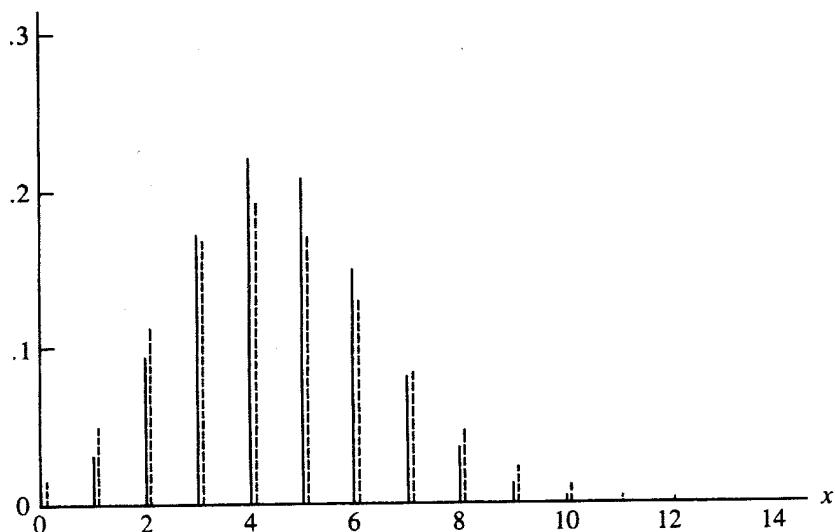


FIGURE 2.3.3 Poisson (dotted line) approximation to the binomial (solid line),  $n = 15, p = .3$

We close this section with a useful result concerning mgfs.

**THEOREM 2.3.5:** For any constants  $a$  and  $b$ , the mgf of the random variable  $aX + b$  is given by

$$M_{aX+b}(t) = e^{bt} M_X(at).$$

*Proof:* By definition,

$$\begin{aligned} M_{aX+b}(t) &= \mathbb{E}(e^{(aX+b)t}) \\ &= \mathbb{E}(e^{(aX)t} e^{bt}) \quad (\text{properties of exponentials}) \\ &= e^{bt} \mathbb{E}(e^{X(at)}) \quad (e^{bt} \text{ is constant}) \\ &= e^{bt} M_X(at), \quad (\text{definition of mgf}) \end{aligned}$$

proving the theorem. □

## 2.4 Differentiating Under an Integral Sign

In the previous section we encountered an instance in which we desired to interchange the order of integration and differentiation. This situation is encountered frequently in theoretical statistics. The purpose of this section is to characterize conditions under which this operation is legitimate. We will also discuss interchanging the order of differentiation and summation.

Many of these conditions can be established using standard theorems from calculus and detailed proofs can be found in most calculus textbooks. Thus, detailed proofs will not be presented here.

We first want to establish the method of calculating

$$(2.4.1) \quad \frac{d}{d\theta} \int_{a(\theta)}^{b(\theta)} f(x, \theta) dx$$

where  $-\infty < a(\theta), b(\theta) < \infty$  for all  $\theta$ . The rule for differentiating (2.4.1) is called Leibnitz's Rule and is an application of the Fundamental Theorem of Calculus and the chain rule.

**LEIBNITZ'S RULE:** If  $f(x, \theta)$ ,  $a(\theta)$ , and  $b(\theta)$  are differentiable with respect to  $\theta$ , then

$$\frac{d}{d\theta} \int_{a(\theta)}^{b(\theta)} f(x, \theta) dx = f(b(\theta), \theta) \frac{d}{d\theta} b(\theta) - f(a(\theta), \theta) \frac{d}{d\theta} a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial \theta} f(x, \theta) dx.$$

Notice that if  $a(\theta)$  and  $b(\theta)$  are constant, we have a special case of Leibnitz's Rule:

$$\frac{d}{d\theta} \int_a^b f(x, \theta) dx = \int_a^b \frac{\partial}{\partial \theta} f(x, \theta) dx.$$

Thus, in general, if we have the integral of a differentiable function over a finite range, differentiation of the integral poses no problem. If the range of integration is infinite, however, problems can arise.

Note that the interchange of derivative and integral in the above equation equates a partial derivative with an ordinary derivative. Formally, this must be the case since the left-hand side is a function only of  $\theta$ , while the integrand on the right-hand side is a function of both  $\theta$  and  $x$ .

The question of whether interchanging the order of differentiation and integration is justified is really a question of whether limits and integration can be interchanged, since a derivative is a special kind of limit. Recall that if  $f(x, \theta)$  is differentiable, then

$$\frac{\partial}{\partial \theta} f(x, \theta) = \lim_{\delta \rightarrow 0} \frac{f(x, \theta + \delta) - f(x, \theta)}{\delta},$$

so we have

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta) dx = \int_{-\infty}^{\infty} \lim_{\delta \rightarrow 0} \left[ \frac{f(x, \theta + \delta) - f(x, \theta)}{\delta} \right] dx,$$

while

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta) dx = \lim_{\delta \rightarrow 0} \int_{-\infty}^{\infty} \left[ \frac{f(x, \theta + \delta) - f(x, \theta)}{\delta} \right] dx.$$

Therefore, if we can justify the interchanging of the order of limits and integration, differentiation under the integral sign will be justified. Treatment of this problem in full generality will, unfortunately, necessitate the use of measure theory, a topic that will not be covered in this book. However, the statements and conclusions of

some important results can be given. The following theorems are all corollaries of Lebesgue's Dominated Convergence Theorem (see, for example, Rudin (1976)).

**THEOREM 2.4.1:** Suppose the function  $h(x, y)$  is continuous at  $y_0$  for each  $x$ , and there exists a function  $g(x)$  satisfying

- a.  $|h(x, y)| \leq g(x)$  for all  $x$  and  $y$
- b.  $\int_{-\infty}^{\infty} g(x) dx < \infty$ .

Then

$$\lim_{y \rightarrow y_0} \int_{-\infty}^{\infty} h(x, y) dx = \int_{-\infty}^{\infty} \lim_{y \rightarrow y_0} h(x, y) dx.$$
□

The key condition in this theorem is the existence of a dominating function  $g(x)$ , with a finite integral, which ensures that the integrals cannot be too badly behaved. We can now apply this theorem to the case we are considering by identifying  $h(x, y)$  with the difference  $(f(x, \theta + \delta) - f(x, \theta))/\delta$ .

**THEOREM 2.4.2:** Suppose  $f(x, \theta)$  is differentiable at  $\theta = \theta_0$ , that is,

$$\lim_{\delta \rightarrow 0} \frac{f(x, \theta_0 + \delta) - f(x, \theta_0)}{\delta} = \frac{\partial}{\partial \theta} f(x, \theta)|_{\theta=\theta_0}$$

exists for every  $x$ , and there exists a function  $g(x, \theta_0)$  and a constant  $\delta_0 > 0$  such that

- a.  $\left| \frac{f(x, \theta_0 + \delta) - f(x, \theta_0)}{\delta} \right| \leq g(x, \theta_0)$  for all  $x$  and  $|\delta| \leq \delta_0$
- b.  $\int_{-\infty}^{\infty} g(x, \theta_0) dx < \infty$ .

Then

$$(2.4.2) \quad \frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta) dx|_{\theta=\theta_0} = \int_{-\infty}^{\infty} \left[ \frac{\partial}{\partial \theta} f(x, \theta)|_{\theta=\theta_0} \right] dx.$$
□

The conclusion of Theorem 2.4.2 is a little cumbersome, but it is important to realize that although we seem to be treating  $\theta$  as a variable, the statement of the theorem is for one value of  $\theta$ . That is, for each value  $\theta_0$  for which  $f(x, \theta)$  is differentiable at  $\theta_0$  and satisfies conditions (a) and (b), the order of integration and differentiation can be interchanged. Often the distinction between  $\theta$  and  $\theta_0$  is not stressed and (2.4.2) is written

$$(2.4.3) \quad \frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta) dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta) dx.$$

Typically,  $f(x, \theta)$  is differentiable at all  $\theta$ , not just at one value  $\theta_0$ . In this case, condition (a) of Theorem 2.4.2 can be replaced by another condition that often proves easier to verify. By an application of the mean value theorem, it follows that, for fixed  $x$  and  $\theta_0$ , and  $|\delta| \leq \delta_0$ ,

$$\frac{f(x, \theta_0 + \delta) - f(x, \theta_0)}{\delta} = \frac{\partial}{\partial \theta} f(x, \theta)|_{\theta=\theta_0+\delta^*(x)}$$

for some number  $\delta^*(x)$ ,  $|\delta^*(x)| \leq \delta_0$ . Therefore, condition (a) will be satisfied if we find a  $g(x, \theta)$  that satisfies condition (b) and

$$(2.4.4) \quad \left| \frac{\partial}{\partial \theta} f(x, \theta)|_{\theta=\theta'} \right| \leq g(x, \theta) \quad \text{for all } \theta' \text{ such that } |\theta' - \theta| \leq \delta_0.$$

Note that in (2.4.4)  $\delta_0$  is implicitly a function of  $\theta$ , as is the case in Theorem 2.4.2. This is permitted since the theorem is applied to each value of  $\theta$  individually. From (2.4.4) we get the following corollary.

**COROLLARY 2.4.1:** Suppose  $f(x, \theta)$  is differentiable in  $\theta$  and there exists a function  $g(x, \theta)$  such that (2.4.4) is satisfied and  $\int_{-\infty}^{\infty} g(x, \theta) dx < \infty$ . Then (2.4.3) holds.  $\square$

Notice that both condition (a) of Theorem 2.4.2 and (2.4.4) impose a uniformity requirement on the functions to be bounded; some type of uniformity is generally needed before derivatives and integrals can be interchanged.

**Example 2.4.1:** Let  $X$  have the exponential( $\lambda$ ) pdf given by  $f(x) = (1/\lambda)e^{-x/\lambda}$ ,  $0 < x < \infty$ , and suppose we want to calculate

$$(2.4.5) \quad \frac{d}{d\lambda} EX^n = \frac{d}{d\lambda} \int_0^\infty x^n \left( \frac{1}{\lambda} \right) e^{-x/\lambda} dx,$$

for integer  $n > 0$ . If we could move the differentiation inside the integral, we would have

$$(2.4.6) \quad \begin{aligned} \frac{d}{d\lambda} EX^n &= \int_0^\infty \frac{\partial}{\partial \lambda} x^n \left( \frac{1}{\lambda} \right) e^{-x/\lambda} dx \\ &= \int_0^\infty \frac{x^n}{\lambda^2} \left( \frac{x}{\lambda} - 1 \right) e^{-x/\lambda} dx \\ &= \frac{1}{\lambda^2} EX^{n+1} - \frac{1}{\lambda} EX^n. \end{aligned}$$

To justify the interchange of integration and differentiation, we bound the derivative of  $x^n(1/\lambda)e^{-x/\lambda}$ . Now

$$(2.4.7) \quad \begin{aligned} \left| \frac{\partial}{\partial \lambda} \left( \frac{x^n e^{-x/\lambda}}{\lambda} \right) \right| &= \frac{x^n e^{-x/\lambda}}{\lambda^2} \left| \frac{x}{\lambda} - 1 \right| \\ &\leq \frac{x^n e^{-x/\lambda}}{\lambda^2} \left( \frac{x}{\lambda} + 1 \right). \quad \left( \text{since } \frac{x}{\lambda} > 0 \right) \end{aligned}$$

For some constant  $\delta_0$  satisfying  $0 < \delta_0 < \lambda$ , take

$$g(x, \lambda) = \frac{x^n e^{-x/(\lambda+\delta_0)}}{(\lambda - \delta_0)^2} \left( \frac{x}{\lambda - \delta_0} + 1 \right).$$

We then have

$$\left| \frac{\partial}{\partial \lambda} \left( \frac{x^n e^{-x/\lambda}}{\lambda} \right) \Big|_{\lambda=\lambda'} \right| \leq g(x, \lambda) \quad \text{for all } \lambda' \text{ such that } |\lambda' - \lambda| \leq \delta_0.$$

Since the exponential distribution has all of its moments, it follows that  $\int_{-\infty}^{\infty} g(x, \lambda) dx < \infty$  as long as  $\lambda - \delta_0 > 0$ , so the interchange of integration and differentiation is justified. ||

The property illustrated for the exponential distribution holds for a large class of densities, which will be dealt with in Section 3.3.

Notice that (2.4.6) gives us a recursion relation for the moments of the exponential distribution,

$$(2.4.8) \quad EX^{n+1} = \lambda EX^n + \lambda^2 \frac{d}{d\lambda} EX^n,$$

making the calculation of the  $(n+1)$ st moment relatively easy. This type of relationship exists for other distributions. In particular, if  $X$  has a normal distribution with mean  $\mu$  and variance 1, so it has pdf  $f(x) = (1/\sqrt{2\pi})e^{-(x-\mu)^2/2}$ , then

$$EX^{n+1} = \mu EX^n - \frac{d}{d\mu} EX^n.$$

We illustrate one more interchange of differentiation and integration, one involving the moment generating function.

**Example 2.4.2:** Again let  $X$  have a normal distribution with mean  $\mu$  and variance 1, and consider the mgf of  $X$ ,

$$M_X(t) = Ee^{tX} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-(x-\mu)^2/2} dx.$$

In Section 2.3 it was stated that we can calculate moments by differentiation of  $M_X(t)$  and differentiation under the integral sign was justified:

$$(2.4.9) \quad \frac{d}{dt} M_X(t) = \frac{d}{dt} Ee^{tX} = E \frac{\partial}{\partial t} e^{tX} = EX e^{tX}.$$

We can apply the results of this section to justify the operations in (2.4.9). Notice that when applying either Theorem 2.4.2 or Corollary 2.4.1 here, we identify  $t$  with the variable  $\theta$  in Theorem 2.4.2. The parameter  $\mu$  is treated as a constant.

From Corollary 2.4.1, we must find a function  $g(x, t)$ , with finite integral, that satisfies

$$(2.4.10) \quad \frac{\partial}{\partial t} e^{tx} e^{-(x-\mu)^2/2} \Big|_{t=t'} \leq g(x, t) \quad \text{for all } t' \text{ such that } |t' - t| \leq \delta_0.$$

Doing the obvious, we have

$$\begin{aligned} \left| \frac{\partial}{\partial t} e^{tx} e^{-(x-\mu)^2/2} \right| &= \left| x e^{tx} e^{-(x-\mu)^2/2} \right| \\ &\leq |x| e^{tx} e^{-(x-\mu)^2/2}. \end{aligned}$$

It is easiest to define our function  $g(x, t)$  separately for  $x \geq 0$  and  $x < 0$ . We take

$$g(x, t) = \begin{cases} |x| e^{(t-\delta_0)x} e^{-(x-\mu)^2/2} & \text{if } x < 0 \\ |x| e^{(t+\delta_0)x} e^{-(x-\mu)^2/2} & \text{if } x \geq 0 \end{cases}$$

It is clear that this function satisfies (2.4.10); it remains to check that its integral is finite.

For  $x \geq 0$  we have

$$g(x, t) = x e^{-(x^2 - 2x(\mu + t + \delta_0) + \mu^2)/2}.$$

We now complete the square in the exponent, that is, we write

$$\begin{aligned} x^2 - 2x(\mu + t + \delta_0) + \mu^2 &= x^2 - 2x(\mu + t + \delta_0) + (\mu + t + \delta_0)^2 - (\mu + t + \delta_0)^2 + \mu^2 \\ &= (x - (\mu + t + \delta_0))^2 + \mu^2 - (\mu + t + \delta_0)^2, \end{aligned}$$

and so, for  $x \geq 0$ ,

$$g(x, t) = x e^{-[x - (\mu + t + \delta_0)]^2/2} e^{-[\mu^2 - (\mu + t + \delta_0)^2]/2}.$$

Since the last exponential factor in the above expression does not depend on  $x$ ,  $\int_0^\infty g(x, t) dx$  is essentially calculating the mean of a normal distribution with mean  $\mu + t + \delta_0$ , except that the integration is only over  $[0, \infty)$ . However, it follows that the integral is finite because the normal distribution has a finite mean (to be shown in Chapter 3). A similar development for  $x < 0$  shows that

$$g(x, t) = |x| e^{-[x - (\mu + t - \delta_0)]^2/2} e^{-[\mu^2 - (\mu + t - \delta_0)^2]/2}$$

and so  $\int_{-\infty}^0 g(x, t) dx < \infty$ . Therefore, we have found an integrable function satisfying (2.4.10) and the operation in (2.4.9) is justified. ||

We now turn to the question of when it is possible to interchange differentiation and summation, an operation that plays an important role in discrete distributions. Of course, we are concerned only with infinite sums, since a derivative can always be taken inside a finite sum.

**Example 2.4.3:** Let  $X$  be a discrete random variable with the *geometric distribution*

$$P(X = x) = \theta(1 - \theta)^x, \quad x = 0, 1, \dots, \quad 0 < \theta < 1.$$

We have that  $\sum_{x=0}^{\infty} \theta(1 - \theta)^x = 1$  and, provided that the operations are justified,

$$\begin{aligned} \frac{d}{d\theta} \sum_{x=0}^{\infty} \theta(1 - \theta)^x &= \sum_{x=0}^{\infty} \frac{d}{d\theta} \theta(1 - \theta)^x \\ &= \sum_{x=0}^{\infty} [(1 - \theta)^x - \theta x(1 - \theta)^{x-1}] \\ &= \frac{1}{\theta} \sum_{x=0}^{\infty} \theta(1 - \theta)^x - \frac{1}{1 - \theta} \sum_{x=0}^{\infty} x\theta(1 - \theta)^x. \end{aligned}$$

Since  $\sum_{x=0}^{\infty} \theta(1 - \theta)^x = 1$  for all  $0 < \theta < 1$ , its derivative is zero. So we have

$$(2.4.11) \quad \frac{1}{\theta} \sum_{x=0}^{\infty} \theta(1 - \theta)^x - \frac{1}{1 - \theta} \sum_{x=0}^{\infty} x\theta(1 - \theta)^x = 0.$$

Now the first sum in (2.4.11) is equal to 1 and the second sum is  $EX$ , hence (2.4.11) becomes

$$\frac{1}{\theta} - \frac{1}{1 - \theta} EX = 0,$$

or

$$EX = \frac{1 - \theta}{\theta}.$$

We have, in essence, summed the series  $\sum_{x=0}^{\infty} x\theta(1 - \theta)^x$  by differentiating. ||

Justification of taking the derivative inside the summation is more straightforward than the integration case. The following theorem provides the details.

**THEOREM 2.4.3:** Suppose that the series  $\sum_{x=0}^{\infty} h(\theta, x)$  converges for all  $\theta$  in an interval  $(a, b)$  of real numbers and

- a.  $\frac{\partial}{\partial \theta} h(\theta, x)$  is continuous in  $\theta$  for each  $x$ ,
- b.  $\sum_{x=0}^{\infty} \frac{\partial}{\partial \theta} h(\theta, x)$  converges uniformly on every closed bounded subinterval of  $(a, b)$ .

Then

$$(2.4.12) \quad \frac{d}{d\theta} \sum_{x=0}^{\infty} h(\theta, x) = \sum_{x=0}^{\infty} \frac{\partial}{\partial \theta} h(\theta, x). \quad \square$$

The condition of uniform convergence is the key one to verify in order to establish that the differentiation can be taken inside the summation. Recall that a series converges uniformly if its sequence of partial sums converges uniformly, a fact that we use in the following example.

**Example (2.4.3) (Continued):** To apply Theorem 2.4.3 we identify

$$h(\theta, x) = \theta(1 - \theta)^x,$$

and

$$\frac{\partial}{\partial \theta} h(\theta, x) = (1 - \theta)^x - \theta x(1 - \theta)^{x-1},$$

and verify that  $\sum_{x=0}^{\infty} \frac{\partial}{\partial \theta} h(\theta, x)$  converges uniformly. Define  $S_n(\theta)$  by

$$S_n(\theta) = \sum_{x=0}^n [(1 - \theta)^x - \theta x(1 - \theta)^{x-1}].$$

The convergence will be uniform on  $[c, d] \subset (0, 1)$  if, given  $\epsilon > 0$ , we can find an  $N$  such that

$$n > N \Rightarrow |S_n(\theta) - S_{\infty}(\theta)| < \epsilon \quad \text{for all } \theta \in [c, d].$$

Recall the partial sum of the geometric series (1.5.3). If  $y \neq 1$ , then we can write

$$\sum_{k=0}^n y^k = \frac{1 - y^{n+1}}{1 - y}.$$

Applying this, we have

$$\begin{aligned}
 \sum_{x=0}^n (1-\theta)^x &= \frac{1-(1-\theta)^{n+1}}{\theta} \\
 \sum_{x=0}^n \theta x (1-\theta)^{x-1} &= \theta \sum_{x=0}^n -\frac{\partial}{\partial \theta} (1-\theta)^x \\
 &= -\theta \frac{d}{d\theta} \sum_{x=0}^n (1-\theta)^x \\
 &= -\theta \frac{d}{d\theta} \left[ \frac{1-(1-\theta)^{n+1}}{\theta} \right].
 \end{aligned}$$

Here we (justifiably) pull the derivative through the finite sum. Calculating this derivative gives

$$\sum_{x=0}^n \theta x (1-\theta)^{x-1} = \frac{(1-(1-\theta)^{n+1}) - (n+1)\theta(1-\theta)^n}{\theta}$$

and, hence,

$$\begin{aligned}
 S_n(\theta) &= \frac{1-(1-\theta)^{n+1}}{\theta} - \frac{(1-(1-\theta)^{n+1}) - (n+1)\theta(1-\theta)^n}{\theta} \\
 &= (n+1)(1-\theta)^n.
 \end{aligned}$$

It is clear that, for  $0 < \theta < 1$ ,  $S_\infty = \lim_{n \rightarrow \infty} S_n(\theta) = 0$ . Since  $S_n(\theta)$  is continuous, the convergence is uniform on any closed bounded interval. Therefore, the series of derivatives converges uniformly and the interchange of differentiation and summation is justified. ||

We close this section with a theorem that is similar to Theorem 2.4.3, but treats the case of interchanging the order of summation and integration.

**THEOREM 2.4.4:** Suppose the series  $\sum_{x=0}^{\infty} h(\theta, x)$  converges uniformly on  $[a, b]$  and that, for each  $x$ ,  $h(\theta, x)$  is a continuous function of  $\theta$ . Then

$$\int_a^b \sum_{x=0}^{\infty} h(\theta, x) d\theta = \sum_{x=0}^{\infty} \int_a^b h(\theta, x) d\theta. \quad \square$$

## EXERCISES

---

**2.1** In each of the following find the pdf of  $Y$ . Show that the pdf integrates to 1.

- a.  $f_X(x) = 42x^5(1-x)$ ,  $0 < x < 1$ ;  $Y = X^3$
- b.  $f_X(x) = 7e^{-7x}$ ,  $0 < x < \infty$ ;  $Y = 4X + 3$

**2.2** In each of the following find the pdf of  $Y$ .

- a.  $f_X(x) = 1, 0 < x < 1; Y = X^2$
- b.  $X$  has pdf

$$f_X(x) = \frac{(n+m+1)!}{n!m!} x^n (1-x)^m, \quad 0 < x < 1, \quad m, n \text{ positive integers};$$

$$Y = -\log X$$

- c.  $X$  has pdf

$$f_X(x) = \frac{1}{\sigma^2} x e^{-(x/\sigma)^2/2}, \quad 0 < x < \infty, \quad \sigma^2 \text{ a positive constant};$$

$$Y = e^X$$

**2.3** Suppose  $X$  has the geometric pmf,  $f_X(x) = \frac{1}{3} \left(\frac{2}{3}\right)^x, x = 0, 1, 2, \dots$ . Determine the probability distribution of  $Y = X/(X+1)$ . Note that here both  $X$  and  $Y$  are discrete random variables. To specify the probability distribution of  $Y$ , specify its pmf.

**2.4** Let  $\lambda$  be a fixed positive constant, and define the function  $f(x)$  by  $f(x) = \frac{1}{2} \lambda e^{-\lambda x}$  if  $x \geq 0$  and  $f(x) = \frac{1}{2} \lambda e^{\lambda x}$  if  $x < 0$ .

- a. Verify that  $f(x)$  is a pdf.
- b. If  $X$  is a random variable with pdf given by  $f(x)$ , find  $P(X < t)$  for all  $t$ . Evaluate all integrals.
- c. Find  $P(|X| < t)$  for all  $t$ . Evaluate all integrals.

**2.5** Let  $X$  have pdf

$$f(x) = \begin{cases} 30x^2(1-x)^2 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the pdf of  $X^2$ .

**2.6** Use Theorem 2.1.3 to find the pdf of  $Y$  in Example 2.1.2. Show that the same answer is obtained by differentiating the cdf given in (2.1.5).

**2.7** In each of the following find the pdf of  $Y$  and show that the pdf integrates to 1.

a.  $f_X(x) = \frac{1}{2} e^{-|x|}, -\infty < x < \infty; Y = |X|^3$

b.  $f_X(x) = \frac{3}{8}(x+1)^2, -1 < x < 1; Y = 1 - X^2$

c.  $f_X(x) = \frac{3}{8}(x+1)^2, -1 < x < 1; Y = 1 - X^2 \text{ if } X \leq 0 \text{ and } Y = 1 - X \text{ if } X > 0$

**2.8** Let  $X$  have pdf

$$f_X(x) = \frac{2}{9}(x+1), \quad -1 \leq x \leq 2.$$

Find the pdf of  $Y = X^2$ . Note that Theorem 2.1.3 is not directly applicable in this problem.

**2.9** In each of the following show that the given function is a cdf and find  $F_X^{-1}(y)$ .

a.  $F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-x} & \text{if } x \geq 0 \end{cases}$

b.  $F_X(x) = \begin{cases} e^x/2 & \text{if } x < 0 \\ 1/2 & \text{if } 0 \leq x < 1 \\ 1 - (e^{1-x}/2) & \text{if } 1 \leq x \end{cases}$

$$\text{c. } F_X(x) = \begin{cases} e^x/4 & \text{if } x < 0 \\ 1 - (e^{-x}/4) & \text{if } x \geq 0 \end{cases}$$

Note that, in part (c),  $F_X(x)$  is discontinuous but (2.1.11) is still the appropriate definition of  $F_X^{-1}(y)$ .

- 2.10** If the random variable  $X$  has pdf

$$f(x) = \begin{cases} \frac{x-1}{2} & 1 < x < 3 \\ 0 & \text{otherwise} \end{cases},$$

find a monotone function  $u(x)$  such that the random variable  $Y = u(X)$  has a uniform(0, 1) distribution.

- 2.11** In Theorem 2.1.4 the probability integral transform was proved, relating the uniform cdf to any continuous cdf. In this exercise we investigate the relationship between discrete random variables and uniform random variables. Let  $X$  be a discrete random variable with cdf  $F_X(x)$  and define the random variable  $Y$  as  $Y = F_X(X)$ .
- Prove that  $Y$  is stochastically greater than a uniform(0, 1); that is, if  $U \sim \text{uniform}(0, 1)$ , then

$$P(Y > y) \geq P(U > y) = 1 - y, \quad \text{for all } y, \quad 0 < y < 1,$$

$$P(Y > y) > P(U > y) = 1 - y, \quad \text{for some } y, \quad 0 < y < 1.$$

(Recall that *stochastically greater* was defined in Exercise 1.56.)

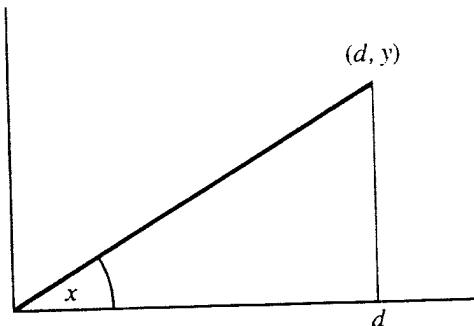
- Equivalently, show that the cdf of  $Y$  satisfies  $F_Y(y) \leq y$  for all  $0 < y < 1$  and  $F_Y(y) < y$  for some  $0 < y < 1$ . (*Hint:* Let  $x_0$  be a jump point of  $F_X$ , and define  $y_0 = F_X(x_0)$ . Show that  $P(Y \leq y_0) = y_0$ . Now establish the inequality by considering  $y = y_0 + \epsilon$ . Pictures of the cdfs will help.)

- 2.12** Let  $X$  have the standard normal pdf,  $f_X(x) = (1/\sqrt{2\pi})e^{-x^2/2}$ .

- Find  $EX^2$  directly, and then by using the pdf of  $Y = X^2$  from Example 2.1.5 and calculating  $EY$ .

- Find the pdf of  $Y = |X|$ , and find its mean and variance.

- 2.13** A random right triangle can be constructed in the following manner. Let  $X$  be a random angle whose distribution is uniform on  $(0, \pi/2)$ . For each  $X$ , construct a triangle as pictured:



Here,  $Y = \text{height of the random triangle}$ . For a fixed constant  $d$ , find the distribution of  $Y$ . Also, find  $EY$ .

- 2.14** Consider a sequence of independent coin flips, each of which has probability  $p$  of being Heads. Define a random variable  $X$  as the length of the run (of either Heads or Tails) started by the first trial. (For example,  $X = 3$  if either TTTH or HHHT is observed.) Find the distribution of  $X$ , and find  $EX$ .

- 2.15** Let  $X$  be a continuous, nonnegative random variable [ $f(x) = 0$  for  $x < 0$ ]. Show that

$$EX = \int_0^\infty [1 - F_X(x)] dx,$$

where  $F_X(x)$  is the cdf of  $X$ .

- 2.16** Let  $X$  be a discrete random variable whose range is the nonnegative integers. Show that

$$EX = \sum_{k=0}^{\infty} (1 - F_X(k)),$$

where  $F_X(k) = P(X \leq k)$ . Compare this with Exercise 2.15.

- 2.17** Use the result of Exercise 2.15 to find the mean duration of certain telephone calls, where we assume that the duration,  $T$ , of a particular call can be described probabilistically by  $P(T > t) = ae^{-\lambda t} + (1 - a)e^{-\mu t}$ , where  $a$ ,  $\lambda$ , and  $\mu$  are constants,  $0 < a < 1$ ,  $\lambda > 0$ ,  $\mu > 0$ .

- 2.18** A *median* of a distribution is a value  $m$  such that  $P(X \leq m) \geq \frac{1}{2}$  and  $P(X \geq m) \geq \frac{1}{2}$ . (If  $X$  is continuous,  $m$  satisfies  $\int_{-\infty}^m f(x) dx = \int_m^\infty f(x) dx = \frac{1}{2}$ .) Find the median of the following distributions.

a.  $f(x) = 3x^2$ ,  $0 < x < 1$       b.  $f(x) = \frac{1}{\pi(1+x^2)}$ ,  $-\infty < x < \infty$

- 2.19** Show that

$$\min_a E|X - a| = E|X - m|,$$

where  $m$  is the median of  $X$  (see Exercise 2.18).

- 2.20** Prove that

$$\frac{d}{da} E(X - a)^2 = 0 \Leftrightarrow EX = a,$$

by differentiating the integral. Verify, using calculus, that  $a = EX$  is indeed a minimum. List the assumptions about  $F_X$  and  $f_X$  that are needed.

- 2.21** A couple decides to continue to have children until a daughter is born. What is the expected number of children of this couple? (*Hint:* See Example 1.5.2.)

- 2.22** Prove the “two-way” rule for expectations, equation (2.2.6), which says  $Eg(X) = EY$ , where  $Y = g(X)$ . Assume that  $g(x)$  is a monotone function.

- 2.23** Let  $X$  have the pdf

$$f(x) = \frac{4}{\beta^3 \sqrt{\pi}} x^2 e^{-x^2/\beta^2}, \quad 0 < x < \infty, \quad \beta > 0.$$

- a. Verify that  $f(x)$  is a pdf.      b. Find  $EX$  and  $\text{Var } X$ .

- 2.24** Let  $X$  have the pdf

$$f(x) = \frac{1}{2}(1+x), \quad -1 < x < 1.$$

- a. Find the pdf of  $Y = X^2$ .      b. Find  $EY$  and  $\text{Var } Y$ .

**2.25** Compute  $EX$  and  $\text{Var } X$  for each of the following probability distributions.

- $f_X(x) = ax^{a-1}, 0 < x < 1, a > 0$
- $f_X(x) = \frac{1}{n}, x = 1, 2, \dots, n; n > 0$  an integer
- $f_X(x) = \frac{3}{2}(x-1)^2, 0 < x < 2$

**2.26** Suppose the pdf  $f_X(x)$  of a random variable  $X$  is an *even function*. ( $f_X(x)$  is an *even function* if  $f_X(x) = f_X(-x)$  for every  $x$ .) Show that

- $X$  and  $-X$  are identically distributed.
- $M_X(t)$  is symmetric about zero.

**2.27** Let  $f(x)$  be a pdf and let  $a$  be a number such that, for all  $\epsilon > 0$ ,  $f(a + \epsilon) = f(a - \epsilon)$ . Such a pdf is said to be *symmetric* about the point  $a$ .

- Give three examples of symmetric pdfs.
- Show that if  $X \sim f(x)$ , symmetric, then the median of  $X$  (see Exercise 2.18) is the number  $a$ .
- Show that if  $X \sim f(x)$ , symmetric, and  $EX$  exists, then  $EX = a$ .
- Show that  $f(x) = e^{-x}, x \geq 0$ , is not a symmetric pdf.
- Show that for the pdf in part (d), the median is less than the mean.

**2.28** Let  $f(x)$  be a pdf and let  $a$  be a number such that, if  $a \geq x \geq y$  then  $f(a) \geq f(x) \geq f(y)$  and, if  $a \leq x \leq y$  then  $f(a) \geq f(x) \geq f(y)$ . Such a pdf is called *unimodal* with a *mode* equal to  $a$ .

- Give an example of a unimodal pdf for which the mode is unique.
- Give an example of a unimodal pdf for which the mode is not unique.
- Show that if  $f(x)$  is both symmetric (see Exercise 2.27) and unimodal, then the point of symmetry is a mode.
- Consider the pdf  $f(x) = e^{-x}, x \geq 0$ . Show that this pdf is unimodal. What is its mode?

**2.29** Let  $\mu_n$  denote the  $n$ th central moment of a random variable  $X$ . Two quantities of interest, in addition to the mean and variance, are

$$\alpha_3 = \frac{\mu_3}{(\mu_2)^{3/2}} \quad \text{and} \quad \alpha_4 = \frac{\mu_4}{\mu_2^2}.$$

The value  $\alpha_3$  is called the *skewness* and  $\alpha_4$  is called the *kurtosis*. The skewness measures the lack of symmetry in the pdf (see Exercise 2.27). The kurtosis, although harder to interpret, measures the peakedness or flatness of the pdf. See Ruppert (1987) for a discussion of the meaning of  $\alpha_4$ .

- Show that if a pdf is symmetric about a point  $a$ , then  $\alpha_3 = 0$ .
- Calculate  $\alpha_3$  for  $f(x) = e^{-x}, x \geq 0$ , a pdf that is *skewed to the right*.
- Calculate  $\alpha_4$  for each of the following pdfs and comment on the peakedness of each.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty$$

$$f(x) = \frac{1}{2}, \quad -1 < x < 1$$

$$f(x) = \frac{1}{2}e^{-|x|}, \quad -\infty < x < \infty$$

**2.30** Does a distribution exist for which  $M_X(t) = t/(1-t), |t| < 1$ ? If yes, find it. If no, prove it.

**2.31** Let  $M_X(t)$  be the moment generating function of  $X$ , and define  $S(t) = \log(M_X(t))$ .

Show that

$$\frac{d}{dt} S(t)|_{t=0} = EX \quad \text{and} \quad \frac{d^2}{dt^2} S(t)|_{t=0} = \text{Var } X.$$

- 2.32 Find the moment generating function corresponding to

a.  $f(x) = \frac{1}{c}, \quad 0 < x < c$

b.  $f(x) = \frac{2x}{c^2}, \quad 0 < x < c$

c.  $f(x) = \frac{1}{2\beta} e^{-|x-\alpha|/\beta}, \quad -\infty < x < \infty, \quad -\infty < \alpha < \infty, \quad \beta > 0$

d.  $P(X = x) = \binom{r+x-1}{x} p^r (1-p)^x, \quad x = 0, 1, \dots, \quad 0 < p < 1,$   
 $r > 0$  an integer

- 2.33 Let  $X$  be a random variable with moment generating function  $M_X(t)$ ,  $-h < t < h$ .

Prove that

a.  $P(X \geq a) \leq e^{-at} M_X(t), \quad 0 < t < h$

b.  $P(X \leq a) \leq e^{-at} M_X(t), \quad -h < t < 0$ .

- 2.34 Fill in the gaps in Example 2.3.5.

- a. Show that if  $X_1 \sim f_1(x)$ , then

$$EX_1^r = e^{r^2/2}, \quad r = 0, 1, \dots$$

So  $f_1(x)$  has all of its moments, and all of the moments are finite.

b. Now show that

$$\int_0^\infty x^r f_1(x) \sin(2\pi \log x) dx = 0,$$

for all positive integers  $r$ , so  $EX_1^r = EX_2^r$  for all  $r$ . (Romano and Siegel (1986) discuss an extreme version of this example, where an entire class of distinct pdfs have the same moments. Also, Berg (1988) has shown that this moment behavior can arise with simpler transforms of the normal distribution such as  $X^3$ .)

- 2.35 The *lognormal distribution*, on which Example 2.3.5 is based, has an interesting property. If we have the pdf

$$f(x) = \frac{1}{\sqrt{2\pi}x} e^{-(\log x)^2/2}, \quad 0 \leq x < \infty,$$

then Exercise 2.34 shows that all moments exist and are finite. However, this distribution does not have a moment generating function, that is,

$$M_X(t) = \int_0^\infty \frac{e^{tx}}{\sqrt{2\pi}x} e^{-(\log x)^2/2} dx$$

does not exist. Prove this.

- 2.36 A distribution cannot be uniquely determined by a finite collection of moments, as this example from Romano and Siegel (1986) shows. Let  $X$  have the normal distribution, that is,  $X$  has pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty.$$

Define a discrete random variable  $Y$  by

$$P(Y = \sqrt{3}) = P(Y = -\sqrt{3}) = \frac{1}{6}, \quad P(Y = 0) = \frac{2}{3}.$$

Show that

$$EX^r = EY^r \quad \text{for } r = 1, 2, 3, 4, 5.$$

(Romano and Siegel point out that for any finite  $n$  there exists a discrete, and hence nonnormal, random variable whose first  $n$  moments are equal to those of  $X$ .)

- 2.37 In each of the following cases verify the expression given for the moment generating function and in each case use the mgf to calculate  $EX$  and  $\text{Var } X$ .

a.  $P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad M_X(t) = e^{\lambda(e^t - 1)}, \quad x = 0, 1, \dots; \quad \lambda > 0$

b.  $P(X = x) = p(1-p)^x, \quad M_X(t) = \frac{p}{1-(1-p)e^t}, \quad x = 0, 1, \dots; \quad 0 < p < 1$

c.  $f_X(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma}, \quad M_X(t) = e^{\mu t + \sigma^2 t^2/2}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$

- 2.38 Let  $X$  have the negative binomial distribution with pmf

$$f(x) = \binom{r+x-1}{x} p^r (1-p)^x, \quad x = 0, 1, 2, \dots$$

where  $0 < p < 1$  and  $r > 0$  is an integer.

- a. Calculate the mgf of  $X$ .

- b. Define a new random variable by  $Y = 2pX$ . Show that, as  $p \downarrow 0$ , the mgf of  $Y$  converges to that of a chi squared random variable with  $2r$  degrees of freedom by showing that

$$\lim_{p \rightarrow 0} M_Y(t) = \left( \frac{1}{1-2t} \right)^r, \quad |t| < \frac{1}{2}.$$

- 2.39 In each of the following cases calculate the indicated derivatives, justifying all operations.

a.  $\frac{d}{dx} \int_0^x e^{-\lambda t} dt \qquad \qquad \qquad$  b.  $\frac{d}{d\lambda} \int_0^\infty e^{-\lambda t} dt$

c.  $\frac{d}{dt} \int_t^1 \frac{1}{x^2} dx \qquad \qquad \qquad$  d.  $\frac{d}{dt} \int_1^\infty \frac{1}{(x-t)^2} dx$

- 2.40 Prove

$$\sum_{k=0}^x \binom{n}{k} p^k (1-p)^{n-k} = (n-x) \binom{n}{x} \int_0^{1-p} t^{n-x-1} (1-t)^x dt.$$

(Hint: Integrate by parts or differentiate both sides with respect to  $p$ .)

## Miscellanea

### Uniqueness of Moment Sequences

A distribution is not necessarily determined by its moments, but if the moment sequence is unique (in that there is only one probability distribution with this sequence of moments) then the moment generating function clearly determines the distribution. A sufficient condition for

the moment sequence to be unique is *Carleman's Condition* (Chung, 1974). If  $X \sim F_X$  and we denote  $\mathbb{E}X^r = \mu'_r$ , then the moment sequence is unique if

$$\sum_{r=1}^{\infty} \frac{1}{(\mu'_{2r})^{1/(2r)}} = +\infty.$$

As mentioned earlier, this condition is, in general, not easy to verify.

Feller (1971) has a very complete development of Laplace transforms, of which mgfs are a special case. In particular, Feller shows that whenever

$$M_X(t) = \sum_{r=0}^{\infty} \frac{(-1)^r \mu'_r t^r}{r!}$$

converges on an interval  $-t_0 \leq t < t_0$ ,  $t_0 > 0$ , the distribution  $F_X$  is uniquely determined. Thus, when the mgf exists, the moment sequence determines the distribution  $F_X$  uniquely.

It should be clear that using the mgf to determine the distribution is a difficult task. A better method is through the use of *characteristic functions*, which are explained below. Although characteristic functions simplify the characterization of a distribution, they necessitate understanding complex analysis. You win some and you lose some.

### **Other Generating Functions**

In addition to the moment generating function, there are a number of other generating functions available. In most cases, the characteristic function is the most useful of these. Except for rare circumstances, the other generating functions are less useful, but there are situations where they can ease calculations.

**Cumulant generating function** For a random variable  $X$ , the cumulant generating function is the function  $\log[M_X(t)]$ . This function can be used to generate the *cumulants* of  $X$ , which are defined (rather circuitously) as the coefficients in the Taylor series of the cumulant generating function (see Exercise 2.31).

**Factorial moment generating function** The factorial moment generating function of  $X$  is defined as  $\mathbb{E}t^X$ , if the expectation exists. The name comes from the fact that this function satisfies

$$\frac{d^r}{dt^r} \mathbb{E}t^X \Big|_{t=1} = \mathbb{E}\{X(X-1)\cdots(X-r+1)\}.$$

If  $X$  is a discrete random variable, then we can write

$$\mathbb{E}t^X = \sum_x t^x P(X=x),$$

and the factorial moment generating function is called the *probability generating function*, since the coefficients of the power series give the probabilities. That is, to obtain the probability that  $X = k$ , calculate

$$\frac{1}{k!} \frac{d^k}{dt^k} \mathbb{E}t^X \Big|_{t=1} = P(X=k).$$

**Characteristic function** Perhaps the most useful of all of these types of functions is the characteristic function. The characteristic function of  $X$  is defined by

$$\checkmark \quad \phi_X(t) = Ee^{itX},$$

where  $i$  is the complex number  $\sqrt{-1}$ , so the above expectation requires complex integration. The characteristic function does much more than the mgf does. When the moments of  $F_X$  exist,  $\phi_X$  can be used to generate them, much like an mgf. The characteristic function always exists and it completely determines the distribution. That is, every cdf has a unique characteristic function. So we can state a theorem like Theorem 2.3.3, for example, but without qualification.

**Theorem (Convergence of Characteristic Functions):** Suppose  $X_k, k = 1, 2, \dots$ , is a sequence of random variables, each with characteristic function  $\phi_{X_k}(t)$ . Furthermore, suppose that

$$\lim_{k \rightarrow \infty} \phi_{X_k}(t) = \phi_X(t), \quad \text{for all } t \text{ in a neighborhood of 0}$$

and  $\phi_X(t)$  is a characteristic function. Then, for all  $x$  where  $F_X(x)$  is continuous,

$$\lim_{k \rightarrow \infty} F_{X_k}(x) = F_X(x). \quad \square$$

A full treatment of generating functions is given by Feller (1968). Characteristic functions can be found in almost any advanced probability text; a particularly readable treatment is in Chung (1974).

### Expected Values

Betteley (1977) provides an interesting addition law for expectations. Recall the addition law for probabilities: For any sets  $A$  and  $B$ ,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Let  $X$  and  $Y$  be any two random variables and define

$$X \wedge Y = \min(X, Y) \quad \text{and} \quad X \vee Y = \max(X, Y).$$

Since  $X + Y = (X \vee Y) + (X \wedge Y)$ , it follows that

$$E(X \vee Y) = EX + EY - E(X \wedge Y),$$

which can often be useful in calculating  $E(X \vee Y)$ . Betteley gives a number of examples.

# 3 Common Families of Distributions

*"I don't admit that a fresh illustration is an explanation," said I with some asperity. "Bravo, Watson! A very dignified and logical remonstrance."*

**Dr. Watson and Sherlock Holmes**  
*The Disappearance of Lady Frances Carfax*

Statistical distributions are used to model populations; as such, we usually deal with a *family* of distributions rather than a single distribution. This family is indexed by one or more parameters, which allow us to vary certain characteristics of the distribution while staying with one functional form. For example, we may specify that the normal distribution is a reasonable choice to model a particular population, but we cannot precisely specify the mean. Then, we deal with a parametric family, normal distributions with mean  $\mu$ , where  $\mu$  is an unspecified parameter,  $-\infty < \mu < \infty$ .

In this chapter we catalog many of the more common statistical distributions, some of which we have previously encountered. For each distribution we will give its mean and variance, and many other useful or descriptive measures that may aid understanding. We will also indicate some typical applications of these distributions and some interesting and useful interrelationships. Some of these facts are summarized in tables at the end of the book. This chapter is, by no means, comprehensive in its coverage of statistical distributions. That task has been accomplished by N. L. Johnson and S. Kotz in their multiple-volume work *Distributions in Statistics*.

## 3.1 Discrete Distributions

A random variable  $X$  is said to have a discrete distribution if the range of  $X$ , the sample space, is countable. In most situations, the random variable has integer-valued outcomes.

### *Discrete Uniform Distribution*

A random variable  $X$  has a *discrete uniform*(1,  $N$ ) *distribution* if

$$(3.1.1) \quad P(X = x|N) = \frac{1}{N}, \quad x = 1, 2, \dots, N,$$

where  $N$  is a specified integer. This distribution puts equal mass on each of the outcomes  $1, 2, \dots, N$ .

*A note on notation:* When we are dealing with parametric distributions, as will almost always be the case, the distribution is dependent on values of the parameters. In order to emphasize this fact and to keep track of the parameters, we write them in the pmf preceded by a “|” (given). This convention will also be used with cdfs, pdfs, expectations, and other places where it might be necessary to keep track of the parameters. When there is no possibility of confusion, the parameters may be omitted in order not to clutter up notation too much.

To calculate the mean and variance of  $X$ , recall the identities (provable by induction)

$$\sum_{i=1}^k i = \frac{k(k+1)}{2} \quad \text{and} \quad \sum_{i=1}^k i^2 = \frac{k(k+1)(2k+1)}{6}.$$

We then have

$$EX = \sum_{x=1}^N xP(X=x|N) = \sum_{x=1}^N x \frac{1}{N} = \frac{N+1}{2}$$

and

$$EX^2 = \sum_{x=1}^N x^2 \frac{1}{N} = \frac{(N+1)(2N+1)}{6}$$

and so

$$\begin{aligned} \text{Var } X &= EX^2 - (EX)^2 \\ &= \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 \\ &= \frac{(N+1)(N-1)}{12}. \end{aligned}$$

This distribution can be generalized so that the sample space is any range of integers,  $N_0, N_0 + 1, \dots, N_1$ , with pmf  $P(X = x|N_0, N_1) = 1/(N_1 - N_0 + 1)$ .

### Hypergeometric Distribution

The hypergeometric distribution has many applications in finite population sampling and is best understood through the classic example of the urn model.

Suppose we have a large urn filled with  $N$  balls that are identical in every way except that  $M$  are red and  $N - M$  are green. We reach in, blindfolded, and select  $K$  balls at random (the  $K$  balls are taken all at once, a case of sampling without replacement). What is the probability that exactly  $x$  of the balls are red?

The total number of samples of size  $K$  that can be drawn from the  $N$  balls is  $\binom{N}{K}$ , as was discussed in Section 1.2.3. It is required that  $x$  of the balls be red,

and this can be accomplished in  $\binom{M}{x}$  ways, leaving  $\binom{N-M}{K-x}$  ways of filling out the sample with  $K - x$  green balls. Thus, if we let  $X$  denote the number of red balls in a sample of size  $K$ , then  $X$  has a *hypergeometric distribution* given by

$$(3.1.2) \quad P(X = x|N, M, K) = \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}}, \quad x = 0, 1, \dots, K.$$

Note that there is, implicit in (3.1.2), an additional assumption on the range of  $X$ . Binomial coefficients, of the form  $\binom{n}{r}$ , have been defined only if  $n \geq r$ , and so the range of  $X$  is additionally restricted by the pair of inequalities

$$M \geq x \quad \text{and} \quad N - M \geq K - x,$$

which can be combined as

$$M - (N - K) \leq x \leq M.$$

In many cases  $K$  is small compared to  $M$  and  $N$ , so the range  $0 \leq x \leq K$  will be contained in the above range and, hence, will be appropriate. The formula for the hypergeometric probability function is usually quite difficult to deal with. In fact, it is not even trivial to verify that

$$\sum_{x=0}^K P(X = x) = \sum_{x=0}^K \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}} = 1.$$

The hypergeometric distribution illustrates the fact that, statistically, dealing with finite populations (finite  $N$ ) is a difficult task.

The mean of the hypergeometric distribution is given by

$$EX = \sum_{x=0}^K x \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}} = \sum_{x=1}^K x \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}}. \quad (\text{summand is zero at } x = 0)$$

To evaluate this expression, we use the identities (already encountered in Section 2.3),

$$x \binom{M}{x} = M \binom{M-1}{x-1},$$

$$\binom{N}{K} = \frac{N}{K} \binom{N-1}{K-1},$$

and obtain

$$EX = \sum_{x=1}^K \frac{M \binom{M-1}{x-1} \binom{N-M}{K-x}}{\frac{N}{K} \binom{N-1}{K-1}} = \frac{KM}{N} \sum_{x=1}^K \frac{\binom{M-1}{x-1} \binom{N-M}{K-x}}{\binom{N-1}{K-1}}.$$

We now can recognize the second sum above as the sum of the probabilities for another hypergeometric distribution based on parameter values  $N - 1$ ,  $M - 1$ , and  $K - 1$ . This can be seen clearly by defining  $y = x - 1$  and writing

$$\begin{aligned} \sum_{x=1}^K \frac{\binom{M-1}{x-1} \binom{N-M}{K-x}}{\binom{N-1}{K-1}} &= \sum_{y=0}^{K-1} \frac{\binom{M-1}{y} \binom{(N-1)-(M-1)}{K-1-y}}{\binom{N-1}{K-1}} \\ &= \sum_{y=0}^{K-1} P(Y = y | N - 1, M - 1, K - 1) = 1, \end{aligned}$$

where  $Y$  is a hypergeometric random variable with parameters  $N - 1$ ,  $M - 1$ , and  $K - 1$ . Therefore, for the hypergeometric distribution,

$$EX = \frac{KM}{N}.$$

A similar, but more lengthy, calculation will establish that

$$\text{Var } X = \frac{KM}{N} \left( \frac{(N - M)(N - K)}{N(N - 1)} \right).$$

Note the manipulations used here to calculate  $EX$ . The sum was transformed to another hypergeometric distribution with different parameter values and, by recognizing this fact, we were able to sum the series.

**Example 3.1.1:** The hypergeometric distribution has application in acceptance sampling, as this example will illustrate. Suppose a retailer buys goods in lots and each item can be either acceptable or defective. Let

$$N = \# \text{ items in a lot},$$

$$M = \# \text{ defectives in a lot}.$$

Then we can calculate the probability that a sample of size  $K$  contains  $x$  defectives. To be specific, suppose that a lot of 25 machine parts is delivered, where a part is considered acceptable only if it passes tolerance. We sample 10 parts and find that none are defective (all are within tolerance). What is the probability of this event if there are 6 defectives in the lot of 25? Applying the hypergeometric distribution with  $N = 25$ ,  $M = 6$ ,  $K = 10$ , we have

$$P(X = 0) = \frac{\binom{6}{0} \binom{19}{10}}{\binom{25}{10}} = .028,$$

showing that our observed event is quite unlikely if there are six (or more!) defectives in the lot.

### Binomial Distribution

The binomial distribution, one of the more useful discrete distributions, is based on the idea of a *Bernoulli trial*. A Bernoulli trial (named for James Bernoulli, one of the founding fathers of probability theory) is an experiment with two, and only two, possible outcomes. A random variable  $X$  has a *Bernoulli( $p$ ) distribution* if

$$(3.1.3) \quad X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}, \quad 0 \leq p \leq 1.$$

The value  $X = 1$  is often termed a “success” and  $p$  is referred to as the success probability. The value  $X = 0$  is termed a “failure.” The mean and variance of a Bernoulli( $p$ ) random variable are easily seen to be

$$\begin{aligned} EX &= 1p + 0(1 - p) = p, \\ \text{Var } X &= (1 - p)^2 p + (0 - p)^2 (1 - p) = p(1 - p). \end{aligned}$$

Many experiments can be modeled as a sequence of Bernoulli trials, the simplest being the repeated tossing of a coin;  $p$  = probability of a head,  $X = 1$  if the coin shows heads. Other examples include gambling games (for example, in roulette let  $X = 1$  if red occurs, so  $p$  = probability of red), election polls ( $X = 1$  if candidate A gets a vote), and incidence of a disease ( $p$  = probability that a random person gets infected).

If  $n$  identical Bernoulli trials are performed, define the events

$$A_i = \{X = 1 \text{ on the } i\text{th trial}\}, \quad i = 1, 2, \dots, n.$$

If we assume that the events  $A_1, \dots, A_n$  are a collection of independent events (as is the case in coin tossing), it is then easy to derive the distribution of the total number of successes in  $n$  trials. Define a random variable  $Y$  by

$$Y = \text{total number of successes in } n \text{ trials.}$$

The event  $\{Y = y\}$  will occur only if, out of the events  $A_1, \dots, A_n$ , exactly  $y$  of them occur, and necessarily  $n - y$  of them do not occur. One particular outcome (one particular ordering of occurrences and nonoccurrences) of the  $n$  Bernoulli trials might be  $A_1 \cap A_2 \cap A_3^c \cap \dots \cap A_{n-1} \cap A_n^c$ . This has probability of occurrence

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3^c \cap \cdots \cap A_{n-1} \cap A_n^c) &= pp(1-p) \cdots p(1-p) \\ &= p^y(1-p)^{n-y}, \end{aligned}$$

where we have used the independence of the  $A_i$ 's in this calculation. Notice that the calculation is not dependent on *which* set of  $y A_i$ 's occurs, only that *some* set of  $y$  occurs. Furthermore, the event  $\{Y = y\}$  will occur no matter which set of  $y A_i$ 's occurs. Putting this all together, we see that a particular sequence of  $n$  trials with exactly  $y$  successes has probability  $p^y(1-p)^{n-y}$  of occurring. Since there are  $\binom{n}{y}$  such sequences (the number of orderings of  $y$  ones and  $n - y$  zeros), we have

$$P(Y = y|n, p) = \binom{n}{y} p^y(1-p)^{n-y}, \quad y = 0, 1, 2, \dots, n,$$

and  $Y$  is called a *binomial( $n, p$ ) random variable*.

The random variable  $Y$  can be alternatively, and equivalently, defined in the following way: In a sequence of  $n$  identical, independent Bernoulli trials, each with success probability  $p$ , define the random variables  $X_1, \dots, X_n$  by

$$X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}.$$

The random variable

$$Y = \sum_{i=1}^n X_i$$

has the binomial( $n, p$ ) distribution.

The fact that  $\sum_{y=0}^n P(Y = y) = 1$  follows from the following general theorem.

**THEOREM 3.1.1 (Binomial Theorem):** For any real numbers  $x$  and  $y$  and integer  $n \geq 0$ ,

$$(3.1.4) \quad (x + y)^n = \sum_{i=0}^n \binom{n}{i} x^i y^{n-i}.$$

*Proof:* Write

$$(x + y)^n = (x + y)(x + y) \cdots (x + y),$$

and consider how the right-hand side would be calculated. From each factor  $(x + y)$  we choose either an  $x$  or  $y$ , and multiply together the  $n$  choices. For each  $i = 0, 1, \dots, n$ , the number of such terms in which  $x$  appears exactly  $i$  times is  $\binom{n}{i}$ . Therefore, this term is of the form  $\binom{n}{i} x^i y^{n-i}$  and the result follows.  $\square$

If we take  $x = p$  and  $y = 1 - p$  in (3.1.4), we get

$$1 = (p + (1 - p))^n = \sum_{i=0}^n \binom{n}{i} p^i (1 - p)^{n-i},$$

and we see that each term in the sum is a binomial probability. As another special case, take  $x = y = 1$  in Theorem 3.1.1 and get the identity

$$2^n = \sum_{i=0}^n \binom{n}{i}.$$

The mean and variance of the binomial distribution have already been derived in Examples 2.2.2 and 2.3.2, so we will not repeat the derivations here. For completeness, we state them. If  $X \sim \text{binomial}(n, p)$  then

$$EX = np, \quad \text{Var } X = np(1 - p).$$

The mgf of the binomial distribution was calculated in Example 2.3.4. It is

$$M_X(t) = [pe^t + (1 - p)]^n.$$

**Example 3.1.2:** Suppose we are interested in finding the probability of obtaining at least one 6 in four rolls of a fair die. This experiment can be modeled as a sequence of four Bernoulli trials with success probability  $p = \frac{1}{6} = P(\text{die shows 6})$ . Define the random variable  $X$  by

$$X = \text{total number of 6s in four rolls.}$$

Then  $X \sim \text{binomial}\left(4, \frac{1}{6}\right)$  and

$$\begin{aligned} P(\text{at least one 6}) &= P(X > 0) = 1 - P(X = 0) \\ &= 1 - \binom{4}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^4 \\ &= 1 - \left(\frac{5}{6}\right)^4 \\ &= .518. \end{aligned}$$

Now we consider another game; throw a pair of dice 24 times and ask for the probability of at least one double 6. This, again, can be modeled by the binomial distribution with success probability  $p$ , where

$$p = P(\text{roll a double 6}) = \frac{1}{36}.$$

So, if  $Y = \text{number of double 6s in 24 rolls}$ ,  $Y \sim \text{binomial}(24, \frac{1}{36})$  and

$$\begin{aligned} P(\text{at least one double 6}) &= P(Y > 0) \\ &= 1 - P(Y = 0) \\ &= 1 - \binom{24}{0} \left(\frac{1}{36}\right)^0 \left(\frac{35}{36}\right)^{24} \\ &= 1 - \left(\frac{35}{36}\right)^{24} \\ &= .491. \end{aligned}$$

This is the calculation originally done in the eighteenth century by Pascal at the request of the gambler de Meré, who thought both events had the same probability. (He began to believe he was wrong when he started losing money on the second bet.) ||

### Poisson Distribution

The Poisson distribution is a widely applied discrete distribution, and can serve as a model for a number of different types of experiments. For example, if we are modeling a phenomenon in which we are waiting for an occurrence (such as waiting for a bus, waiting for customers to arrive in a bank), the number of occurrences in a given time interval can sometimes be modeled by the Poisson distribution. One of the basic assumptions on which the Poisson distribution is built is that, for small time intervals, the probability of an arrival is proportional to the length of waiting time. This makes it a reasonable model for situations like those indicated above. For example, it makes sense to assume that the longer we wait, the more likely it is that a customer will enter the bank. See the *Miscellanea* section for a more formal treatment of this.

Another area of application is in spatial distributions, where, for example, the Poisson may be used to model the distribution of bomb hits in an area or the distribution of fish in a lake.

The Poisson distribution has a single parameter  $\lambda$ , sometimes called the intensity parameter. A random variable  $X$ , taking values in the nonnegative integers, has a *Poisson( $\lambda$ ) distribution* if

$$(3.1.5) \quad P(X = x|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, \dots$$

To see that  $\sum_{x=0}^{\infty} P(X = x|\lambda) = 1$ , recall the Taylor series expansion of  $e^y$ ,

$$e^y = \sum_{i=0}^{\infty} \frac{y^i}{i!}.$$

Thus,

$$\sum_{x=0}^{\infty} P(X = x|\lambda) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1.$$

The mean of  $X$  is easily seen to be

$$\begin{aligned} EX &= \sum_{x=0}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!} \\ &= \sum_{x=1}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!} \\ &= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \quad (\text{substitute } y = x - 1) \\ &= \lambda. \end{aligned}$$

A similar calculation will show that

$$\text{Var } X = \lambda,$$

and so the parameter  $\lambda$  is both the mean and the variance of the Poisson distribution.

The mgf can also be obtained by a straightforward calculation, again following from the Taylor series of  $e^y$ . We have

$$M_X(t) = e^{\lambda(e^t - 1)}.$$

(See Exercise 2.37 and Example 2.3.6.)

**Example 3.1.3:** As an example of a waiting-for-occurrence application, consider a telephone operator who, on the average, handles five calls every 3 minutes. What is the probability that there will be no calls in the next minute? At least two calls?

If we let  $X$  = number of calls in a minute, then  $X$  has a Poisson distribution with  $EX = \lambda = \frac{5}{3}$ . So

$$P(\text{no calls in the next minute}) = P(X = 0)$$

$$\begin{aligned} &= \frac{e^{-5/3} \left(\frac{5}{3}\right)^0}{0!} \\ &= e^{-5/3} = .189; \end{aligned}$$

$$\begin{aligned}
 P(\text{at least two calls in the next minute}) &= P(X \geq 2) \\
 &= 1 - P(X = 0) - P(X = 1) \\
 &= 1 - .189 - \frac{e^{-5/3} \left(\frac{5}{3}\right)^1}{1!} \\
 &= .496. \quad \parallel
 \end{aligned}$$

Calculation of Poisson probabilities can be done rapidly by noting the following recursion relation:

$$(3.1.6) \quad P(X = x) = \frac{\lambda}{x} P(X = x - 1), \quad x = 1, 2, \dots$$

This relation is easily proved by writing out the pmf of the Poisson. Similar relations hold for other discrete distributions. For example, if  $Y \sim \text{binomial}(n, p)$ , then

$$(3.1.7) \quad P(Y = y) = \frac{(n - y + 1)}{y} \frac{p}{1 - p} P(Y = y - 1).$$

The recursion relations (3.1.6) and (3.1.7) can be used to establish the Poisson approximation to the binomial, which we have already seen in Section 2.3, where the approximation was justified using mgfs. Set  $\lambda = np$  and, if  $p$  is small, we can write

$$\frac{n - y + 1}{y} \frac{p}{1 - p} = \frac{np - p(y - 1)}{y - py} \approx \frac{\lambda}{y}$$

since, for small  $p$ , the terms  $p(y - 1)$  and  $py$  can be ignored. Therefore, to this level of approximation, (3.1.7) becomes

$$(3.1.8) \quad P(Y = y) = \frac{\lambda}{y} P(Y = y - 1),$$

which is the Poisson recursion relation. Therefore, to complete the approximation, we only need establish that  $P(X = 0) \approx P(Y = 0)$ , since all other probabilities will follow from (3.1.8). Now

$$P(Y = 0) = (1 - p)^n = \left(1 - \frac{np}{n}\right)^n = \left(1 - \frac{\lambda}{n}\right)^n$$

upon setting  $np = \lambda$ . Recall from Section 2.3 that for fixed  $\lambda$ ,  $\lim_{n \rightarrow \infty} (1 - (\lambda/n))^n = e^{-\lambda}$ , so for large  $n$  we have the approximation

$$P(Y = 0) = \left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda} = P(X = 0),$$

completing the Poisson approximation to the binomial.

The approximation is valid when  $n$  is large and  $p$  is small, which is exactly when it is most useful, freeing us from calculation of binomial coefficients and powers for large  $n$ .

**Example 3.1.4:** A typesetter, on the average, makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?

If we assume that setting a word is a Bernoulli trial with success probability  $p = \frac{1}{500}$  (notice that we are labelling an error as a “success”) and that the trials are independent, then  $X =$  number of errors in five pages (1500 words) is binomial  $(1500, \frac{1}{500})$ . Thus

$$P(\text{no more than two errors}) = P(X \leq 2)$$

$$\begin{aligned} &= \sum_{x=0}^2 \binom{1500}{x} \left(\frac{1}{500}\right)^x \left(\frac{499}{500}\right)^{1500-x} \\ &= .4230, \end{aligned}$$

which is a fairly cumbersome calculation. If we use the Poisson approximation with  $\lambda = 1500(\frac{1}{500}) = 3$ , we have

$$P(X \leq 2) \approx e^{-3} \left(1 + 3 + \frac{3^2}{2}\right) = .4232. \quad ||$$

### Negative Binomial Distribution

The binomial distribution counts the number of successes in a fixed number of Bernoulli trials. Suppose that, instead, we count the number of Bernoulli trials required to get a fixed number of successes. This latter formulation leads to the negative binomial distribution.

In a sequence of independent Bernoulli( $p$ ) trials, let the random variable  $X$  denote the trial at which the  $r$ th success occurs, where  $r$  is a fixed integer. Then

$$(3.1.9) \quad P(X = x|r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \dots$$

and we say that  $X$  has a *negative binomial( $r, p$ ) distribution*.

The derivation of (3.1.9) follows quickly from the binomial distribution. The event  $\{X = x\}$  can occur only if there are exactly  $r-1$  successes in the first  $x-1$  trials, and a success on the  $x$ th trial. The probability of  $r-1$  successes in  $x-1$  trials is the binomial probability  $\binom{x-1}{r-1} p^{r-1} (1-p)^{x-r}$  and with probability  $p$  there is a success on the  $x$ th trial. Multiplying these probabilities gives (3.1.9).

The negative binomial distribution is sometimes defined in terms of the random variable  $Y = \text{number of failures before the } r\text{th success}$ . This formulation is statistically equivalent to the one given above in terms of  $X = \text{trial at which the } r\text{th success occurs}$ , since  $Y = X - r$ . Using the relationship between  $Y$  and  $X$ , the alternative form of the negative binomial distribution is

$$(3.1.10) \quad P(Y = y) = \binom{r+y-1}{y} p^r (1-p)^y, \quad y = 0, 1, \dots$$

Unless otherwise noted, when we refer to the negative binomial( $r, p$ ) distribution we will use this pmf.

The negative binomial distribution gets its name from the relationship

$$\binom{r+y-1}{y} = (-1)^y \binom{-r}{y} = (-1)^y \frac{(-r)(-r-1)(-r-2)\dots(-r-y+1)}{(y)(y-1)(y-2)\dots(2)(1)},$$

which is, in fact, the defining equation for binomial coefficients with negative integers (see Feller (1968) for a complete treatment). Substituting into (3.1.10) yields

$$P(Y = y) = (-1)^y \binom{-r}{y} p^r (1-p)^y,$$

which bears a striking resemblance to the binomial distribution.

The fact that  $\sum_{y=0}^{\infty} P(Y = y) = 1$  is not easy to verify, but follows from an extension of the Binomial Theorem, an extension that includes negative exponents. We will not pursue this further here. An excellent exposition on binomial coefficients can be found in Feller (1968).

The mean and variance of  $Y$  can be calculated using techniques similar to those used for the binomial distribution:

$$\begin{aligned} EY &= \sum_{y=0}^{\infty} y \binom{r+y-1}{y} p^r (1-p)^y \\ &= \sum_{y=1}^{\infty} \frac{(r+y-1)!}{(y-1)!(r-1)!} p^r (1-p)^y \\ &= \sum_{y=1}^{\infty} r \binom{r+y-1}{y-1} p^r (1-p)^y. \end{aligned}$$

Now write  $z = y - 1$ , and the sum becomes

$$\begin{aligned} EY &= \sum_{z=0}^{\infty} r \binom{r+z}{z} p^r (1-p)^{z+1} \\ &= r \frac{(1-p)}{p} \sum_{z=0}^{\infty} \binom{(r+1)+z-1}{z} p^{r+1} (1-p)^z \quad \left( \begin{array}{l} \text{summand is negative} \\ \text{binomial pmf} \end{array} \right) \end{aligned}$$

$$= r \frac{(1-p)}{p}.$$

Since the sum is over all values of a negative binomial( $r + 1, p$ ) distribution, it equals 1. A similar calculation will show

$$\text{Var } Y = \frac{r(1-p)}{p^2}.$$

There is an interesting, and sometimes useful, reparameterization of the negative binomial distribution in terms of its mean. If we define the parameter  $\mu = r(1-p)/p$ , then  $EY = \mu$  and a little algebra will show

$$\text{Var } Y = \mu + \frac{1}{r}\mu^2.$$

The variance is a quadratic function of the mean. This relationship can be useful in both data analysis and theoretical considerations (Morris, 1982).

The negative binomial family of distributions includes the Poisson distribution as a limiting case. If  $r \rightarrow \infty$  and  $p \rightarrow 1$  such that  $r(1-p) \rightarrow \lambda$ ,  $0 < \lambda < \infty$ , then

$$\begin{aligned} EY &= \frac{r(1-p)}{p} \rightarrow \lambda, \\ \text{Var } Y &= \frac{r(1-p)}{p^2} \rightarrow \lambda, \end{aligned}$$

which agree with the Poisson mean and variance. To demonstrate that the negative binomial( $r, p$ )  $\rightarrow$  Poisson( $\lambda$ ), we can show that all of the probabilities converge. The fact that the mgfs converge leads us to expect this (see Exercise 3.13).

**Example 3.1.5:** A technique known as inverse binomial sampling is useful in sampling biological populations. If the proportion of individuals possessing a certain characteristic is  $p$  and we sample until we see  $r$  such individuals, then the number of individuals sampled is a negative binomial random variable.

For example, suppose that in a population of fruit flies we are interested in the proportion having vestigial wings and decide to sample until we have found 100 such flies. The probability that we will have to examine at least  $N$  flies is (using (3.1.9))

$$\begin{aligned} P(X \geq N) &= \sum_{x=N}^{\infty} \binom{x-1}{99} p^{100} (1-p)^{x-100} \\ &= 1 - \sum_{x=100}^{N-1} \binom{x-1}{99} p^{100} (1-p)^{x-100}. \end{aligned}$$

For given  $p$  and  $N$ , we can evaluate this expression to determine how many fruit flies we are likely to look at. (Although the evaluation is cumbersome, the use of a recursion relation will speed things up.) ||

Example 3.1.5 shows that the negative binomial distribution can, like the Poisson, be used to model phenomena in which we are waiting for an occurrence. In the negative binomial case we are waiting for a specified number of successes.

### Geometric Distribution

The geometric distribution is the simplest of the waiting time distributions, and is a special case of the negative binomial distribution. If we set  $r = 1$  in (3.1.9) we have

$$P(X = x|p) = p(1 - p)^{x-1}, \quad x = 1, 2, \dots$$

which defines the pmf of a *geometric random variable*  $X$  with success probability  $p$ .  $X$  can be interpreted as the trial at which the first success occurs, so we are “waiting for a success.” The fact that  $\sum_{x=1}^{\infty} P(X = x) = 1$  follows from properties of the geometric series. For any number  $a$  with  $|a| < 1$ ,

$$\sum_{x=1}^{\infty} a^{x-1} = \frac{1}{1-a},$$

which we have already encountered in Example 1.5.2.

The mean and variance of  $X$  can be calculated by using the negative binomial formulas and by writing  $X = Y + 1$  to obtain

$$EX = EY + 1 = \frac{1}{p} \quad \text{and} \quad \text{Var } X = \frac{1-p}{p^2}.$$

The geometric distribution has an interesting property, known as the “memoryless” property. For integers  $s > t$ , it is the case that

$$(3.1.11) \quad P(X > s|X > t) = P(X > s - t);$$

that is, the geometric distribution “forgets” what has occurred. The probability of getting an additional  $s - t$  failures, having already observed  $t$  failures, is the same as the probability of observing  $s - t$  failures at the start of the sequence. In other words, the probability of getting a run of failures depends only on the length of the run, not on its position.

To establish (3.1.11), we first note that for any integer  $n$ ,

$$(3.1.12) \quad \begin{aligned} P(X > n) &= P(\text{no successes in } n \text{ trials}) \\ &= (1 - p)^n, \end{aligned}$$

and hence

$$\begin{aligned} P(X > s | X > t) &= \frac{P(X > s \text{ and } X > t)}{P(X > t)} \\ &= \frac{P(X > s)}{P(X > t)} \\ &= (1 - p)^{s-t} \\ &= P(X > s - t). \end{aligned}$$

**Example 3.1.6:** The geometric distribution is sometimes used to model “lifetimes” or “time until failure” of components. For example, if the probability is .001 that a light bulb will fail on any given day, then the probability that it will last at least 30 days is

$$P(X > 30) = \sum_{x=31}^{\infty} .001(1 - .001)^{x-1} = (.999)^{30} = .970.$$

The memoryless property of the geometric distribution describes a very special “lack of aging” property. It indicates that the geometric distribution is not applicable to modeling lifetimes for which the probability of failure is expected to increase with time. There are other distributions used to model various types of aging; see, for example, Barlow and Proschan (1975).

## 3.2 Continuous Distributions

In this section we will discuss some of the more common families of continuous distributions, those with well-known names. The distributions mentioned here by no means constitute all of the distributions used in statistics. Indeed, as was seen in Section 1.6, any nonnegative, integrable function can be transformed into a pdf.

### Uniform Distribution

The continuous *uniform distribution* is defined by spreading mass uniformly over an interval  $[a, b]$ . Its pdf is given by

$$(3.2.1) \quad f(x|a, b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

It is easy to check that  $\int_a^b f(x)dx = 1$ . We also have

$$\begin{aligned} EX &= \int_a^b \frac{x}{b-a} dx = \frac{b+a}{2} \\ \text{Var } X &= \int_a^b \frac{(x - \frac{b+a}{2})^2}{b-a} dx = \frac{(b-a)^2}{12}. \end{aligned}$$

## 2 Gamma Distribution

The gamma family of distributions is a flexible family of distributions on  $[0, \infty)$ , and can be derived by the construction discussed in Section 1.6. If  $\alpha$  is a positive constant, the integral

$$\int_0^\infty t^{\alpha-1} e^{-t} dt$$

is finite. If  $\alpha$  is a positive integer the integral can be expressed in closed form, otherwise it cannot. In either case its value defines the *gamma function*,

$$(3.2.2) \quad \Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt.$$

The gamma function satisfies many useful relationships, in particular

$$(3.2.3) \quad \boxed{\Gamma(\alpha + 1) = \alpha\Gamma(\alpha), \quad \alpha > 0,} \quad \checkmark$$

which can be verified through integration by parts. Combining (3.2.3) with the easily verified fact that  $\Gamma(1) = 1$ , we have for any integer  $n > 0$ ,

$$(3.2.4) \quad \boxed{\Gamma(n) = (n - 1)!} \quad \checkmark$$

(Another useful special case, which will be seen in (3.2.15), is that  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ .)

Expressions (3.2.3) and (3.2.4) give recursion relations that ease the problems of calculating values of the gamma function. The recursion relation allows us to calculate any value of the gamma function from knowing only the values of  $\Gamma(c)$ ,  $0 < c \leq 1$ .

Since the integrand in (3.2.2) is positive, it immediately follows that

$$(3.2.5) \quad f(t) = \frac{t^{\alpha-1} e^{-t}}{\Gamma(\alpha)}, \quad 0 < t < \infty$$

is a pdf. The full gamma family, however, has two parameters, and can be derived by changing variables to get the pdf of the random variable  $X = \beta T$  in (3.2.5), where  $\beta$  is a positive constant. Upon doing this, we get the *gamma( $\alpha, \beta$ ) family*,

$$(3.2.6) \quad f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad 0 < x < \infty, \quad \alpha > 0, \quad \beta > 0.$$

The parameter  $\alpha$  is known as the shape parameter, since it most influences the peakedness of the distribution, while the parameter  $\beta$  is called the scale parameter, since most of its influence is on the spread of the distribution.

The mean of the  $\text{gamma}(\alpha, \beta)$  distribution is

$$(3.2.7) \quad EX = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x x^{\alpha-1} e^{-x/\beta} dx.$$

To evaluate (3.2.7), notice that the integrand is the kernel of a gamma( $\alpha + 1, \beta$ ) pdf. From (3.2.6) we know that, for any  $\alpha, \beta > 0$ ,

$$(3.2.8) \quad \int_0^\infty x^{\alpha-1} e^{-x/\beta} dx = \Gamma(\alpha)\beta^\alpha,$$

so we have

$$\begin{aligned} EX &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^\alpha e^{-x/\beta} dx = \frac{1}{\Gamma(\alpha)\beta^\alpha} \Gamma(\alpha + 1)\beta^{\alpha+1} \\ &= \frac{\alpha\Gamma(\alpha)\beta}{\Gamma(\alpha)} \\ &= \alpha\beta. \end{aligned} \quad (\text{from (3.2.3)})$$

Note that to evaluate  $EX$  we have again used the technique of recognizing the integral as the kernel of another pdf. (We have already used this technique to calculate the gamma mgf in Example 2.3.3 and, in a discrete case, to do binomial calculations in Examples 2.2.2 and 2.3.2.)

The variance of the gamma( $\alpha, \beta$ ) distribution is calculated in a manner analogous to that used for the mean. In particular, in calculating  $EX^2$  we deal with the kernel of a gamma( $\alpha + 2, \beta$ ) distribution. The result is

$$\text{Var } X = \alpha\beta^2.$$

In Example 2.3.3 we calculated the mgf of a gamma( $\alpha, \beta$ ) distribution. It is given by

$$M_X(t) = \left( \frac{1}{1 - \beta t} \right)^\alpha, \quad t < \frac{1}{\beta}.$$

**Example 3.2.1:** There is an interesting relationship between the gamma and Poisson distributions. If  $X$  is a gamma( $\alpha, \beta$ ) random variable, where  $\alpha$  is an integer, then for any  $x$ ,

$$(3.2.9) \quad P(X \leq x) = P(Y \geq \alpha),$$

where  $Y \sim \text{Poisson}(x/\beta)$ . Equation (3.2.9) can be established by successive integrations by parts, as follows. Since  $\alpha$  is an integer, we write  $\Gamma(\alpha) = (\alpha - 1)!$  to get

$$\begin{aligned} P(X \leq x) &= \frac{1}{(\alpha - 1)!\beta^\alpha} \int_0^x t^{\alpha-1} e^{-t/\beta} dt \\ &= \frac{1}{(\alpha - 1)!\beta^\alpha} \left[ -t^{\alpha-1} \beta e^{-t/\beta} \Big|_0^x + \int_0^x (\alpha - 1)t^{\alpha-2} \beta e^{-t/\beta} dt \right], \end{aligned}$$

where we use the integration by parts substitution  $u = t^{\alpha-1}$ ,  $dv = e^{-t/\beta} dt$ . Continuing our evaluation, we have

$$\begin{aligned} P(X \leq x) &= \frac{-1}{(\alpha-1)!\beta^{\alpha-1}} x^{\alpha-1} e^{-x/\beta} + \frac{1}{(\alpha-2)!\beta^{\alpha-1}} \int_0^x t^{\alpha-2} e^{-t/\beta} dt \\ &= \frac{1}{(\alpha-2)!\beta^{\alpha-1}} \int_0^x t^{\alpha-2} e^{-t/\beta} dt - P(Y = \alpha-1), \end{aligned}$$

where  $Y \sim \text{Poisson}(x/\beta)$ . Continuing in this manner, we can establish (3.2.9). (See Exercise 3.17.) ||

There are two important special cases of the gamma distribution. If we set  $\alpha = p/2$ , where  $p$  is an integer, and  $\beta = 2$ , then the gamma pdf becomes

$$(3.2.10) \quad f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}, \quad 0 < x < \infty,$$

which is the *chi squared pdf with  $p$  degrees of freedom*. The mean, variance, and mgf of the chi squared distribution can all be calculated by using the previously derived gamma formulas.

The chi squared distribution plays an important role in statistical inference, especially when sampling from a normal distribution. This topic will be dealt with in detail in Chapter 5.

Another important special case of the gamma distribution is obtained when we set  $\alpha = 1$ . We then have

$$(3.2.11) \quad f(x|\beta) = \frac{1}{\beta} e^{-x/\beta}, \quad 0 < x < \infty,$$

the *exponential pdf* with scale parameter  $\beta$ . Its mean and variance were calculated in Examples 2.2.1 and 2.3.1.

The exponential distribution can be used to model lifetimes, analogous to the use of the geometric distribution in the discrete case. In fact, the exponential distribution shares the “memoryless” property of the geometric. If  $X \sim \text{exponential}(\beta)$ , that is, with pdf given by (3.2.11), then for  $s > t \geq 0$ ,

$$P(X > s | X > t) = P(X > s - t),$$

since

$$\begin{aligned} P(X > s | X > t) &= \frac{P(X > s, X > t)}{P(X > t)} \\ &= \frac{P(X > s)}{P(X > t)} \quad (\text{since } s > t) \end{aligned}$$

$$\begin{aligned}
&= \frac{\int_s^\infty \frac{1}{\beta} e^{-x/\beta} dx}{\int_t^\infty \frac{1}{\beta} e^{-x/\beta} dx} \\
&= \frac{e^{-s/\beta}}{e^{-t/\beta}} \\
&= e^{-(s-t)/\beta} \\
&= P(X > s - t).
\end{aligned}$$

Another distribution related to both the exponential and the gamma families is the *Weibull distribution*. If  $X \sim \text{exponential}(\beta)$  then  $Y = X^{1/\gamma}$  has a Weibull( $\gamma, \beta$ ) distribution

$$(3.2.12) \quad f_Y(y|\gamma, \beta) = \frac{\gamma}{\beta} y^{\gamma-1} e^{-y^\gamma/\beta}, \quad 0 < y < \infty, \quad \gamma > 0, \quad \beta > 0.$$

Clearly, we could have started with the Weibull and then derived the exponential as a special case ( $\gamma = 1$ ). This is a matter of taste. The Weibull distribution plays an extremely important role in the analysis of failure time data (see Kalbfleisch and Prentice (1980) for a comprehensive treatment of this topic). The Weibull, in particular, is very useful for modeling *hazard functions* (see Exercises 3.25 and 3.26).

### Normal Distribution

The normal distribution (sometimes called the *Gaussian distribution*) plays a central role in a large body of statistics. There are three main reasons for this. First, the normal distribution, and distributions associated with it, are very tractable analytically (although this may not seem so at first glance). Second, the normal distribution has the familiar bell shape, whose symmetry makes it an appealing choice for many population models. Although there are many other distributions that are also bell-shaped, most do not possess the analytic tractability of the normal. Third, there is the Central Limit Theorem (see Chapter 5 for details) which shows that, under mild conditions, the normal distribution can be used to approximate a large variety of distributions in large samples.

The normal distribution has two parameters, usually denoted by  $\mu$  and  $\sigma^2$ , which are its mean and variance. The pdf of the *normal distribution* with mean  $\mu$  and variance  $\sigma^2$  (usually denoted as  $n(\mu, \sigma^2)$ ) is given by

$$(3.2.13) \quad f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty.$$

If  $X \sim n(\mu, \sigma^2)$ , then the random variable  $Z = (X - \mu)/\sigma$  has a  $n(0, 1)$  distribution, also known as the *standard normal*. This is easily established by writing

$$\begin{aligned}
P(Z \leq z) &= P\left(\frac{X - \mu}{\sigma} \leq z\right) \\
&= P(X \leq z\sigma + \mu)
\end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{z\sigma+\mu} e^{-(x-\mu)^2/(2\sigma^2)} dx \\
 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt, \quad (\text{substitute } t = \frac{x-\mu}{\sigma})
 \end{aligned}$$

showing that  $P(Z \leq z)$  is the standard normal cdf.

It therefore follows that all normal probabilities can be calculated in terms of the standard normal. Furthermore, calculations of expected values can be simplified by carrying out the details in the  $n(0, 1)$  case, then transforming the result to the  $n(\mu, \sigma^2)$  case. For example, if  $Z \sim n(0, 1)$ ,

$$EZ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ze^{-z^2/2} dz = -\frac{1}{\sqrt{2\pi}} e^{-z^2/2} \Big|_{-\infty}^{\infty} = 0$$

and so, if  $X \sim n(\mu, \sigma^2)$ , it follows from Theorem 2.2.1 that

$$EX = E(\mu + \sigma Z) = \mu + \sigma EZ = \mu.$$

Similarly, we have that  $\text{Var } Z = 1$  and using Theorem 2.3.1,  $\text{Var } X = \sigma^2$ .

We have not yet established that (3.2.13) integrates to 1 over the whole real line. By applying the standardizing transformation, we need only to show that

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz = 1.$$

Notice that the integrand above is symmetric around 0, implying that the integral over  $(-\infty, 0)$  is equal to the integral over  $(0, \infty)$ . Thus, we reduce the problem to showing

$$(3.2.14) \quad \int_0^{\infty} e^{-z^2/2} dz = \frac{\sqrt{2\pi}}{2} = \sqrt{\frac{\pi}{2}}.$$

The function  $e^{-z^2/2}$  does not have an antiderivative that can be written explicitly in terms of elementary functions (that is, in closed form), so we cannot perform the integration directly. In fact, this is an example of an integration that either you know how to do, or else you can spend a very long time going nowhere. Since both sides of (3.2.14) are positive, the equality will hold if we establish that the squares are equal. Square the integral in (3.2.14) to obtain

$$\begin{aligned}
 \left( \int_0^{\infty} e^{-z^2/2} dz \right)^2 &= \left( \int_0^{\infty} e^{-t^2/2} dt \right) \left( \int_0^{\infty} e^{-u^2/2} du \right) \\
 &= \int_0^{\infty} \int_0^{\infty} e^{-(t^2+u^2)/2} dt du.
 \end{aligned}$$

The integration variables are just dummy variables, so changing their names is allowed. Now, we convert to polar coordinates. Define

$$t = r \cos \theta \quad \text{and} \quad u = r \sin \theta.$$

Then  $t^2 + u^2 = r^2$  and  $dt du = r d\theta dr$  and the limits of integration become  $0 < r < \infty$ ,  $0 < \theta < \pi/2$  (the upper limit on  $\theta$  is  $\pi/2$  because  $t$  and  $u$  are restricted to be positive). We now have

$$\begin{aligned} \int_0^\infty \int_0^\infty e^{-(t^2+u^2)/2} dt du &= \int_0^\infty \int_0^{\pi/2} r e^{-r^2/2} d\theta dr \\ &= \frac{\pi}{2} \int_0^\infty r e^{-r^2/2} dr \\ &= \frac{\pi}{2} \left[ -e^{-r^2/2} \right]_0^\infty \\ &= \frac{\pi}{2}, \end{aligned}$$

which establishes (3.2.14).

This integral is closely related to the gamma function; in fact, by making the substitution  $w = \frac{1}{2}z^2$  in (3.2.14), we see that this integral is essentially  $\Gamma(\frac{1}{2})$ . If we are careful to get the constants correct, we will see that (3.2.14) implies

$$(3.2.15) \quad \Gamma\left(\frac{1}{2}\right) = \int_0^\infty w^{-1/2} e^{-w} dw = \sqrt{\pi}.$$

The normal distribution is somewhat special in the sense that its two parameters,  $\mu$  (the mean) and  $\sigma^2$  (the variance), provide us with complete information about the exact shape and location of the distribution. This property, that the distribution is determined by  $\mu$  and  $\sigma^2$ , is not unique to the normal pdf, but is shared by a family of pdfs called location-scale families, to be discussed in Section 3.4.

Straightforward calculus shows that the normal pdf (3.2.13) has its maximum at  $x = \mu$  and inflection points (where the curve changes from concave to convex) at  $\mu \pm \sigma$ . Furthermore, the probability content within 1, 2, or 3 standard deviations of the mean is

$$\begin{aligned} P(|X - \mu| \leq \sigma) &= P(|Z| \leq 1) = .6826 \\ P(|X - \mu| \leq 2\sigma) &= P(|Z| \leq 2) = .9544 \\ P(|X - \mu| \leq 3\sigma) &= P(|Z| \leq 3) = .9974 \end{aligned}$$

where  $X \sim n(\mu, \sigma^2)$ ,  $Z \sim n(0, 1)$ , and the numerical values are from Table 1. Often, two-digit values reported are .68, .95, and .99, respectively. Although these do not represent the rounded values, they are the values commonly used. Figure 3.2.1 (page 106) shows the normal pdf along with these key features.

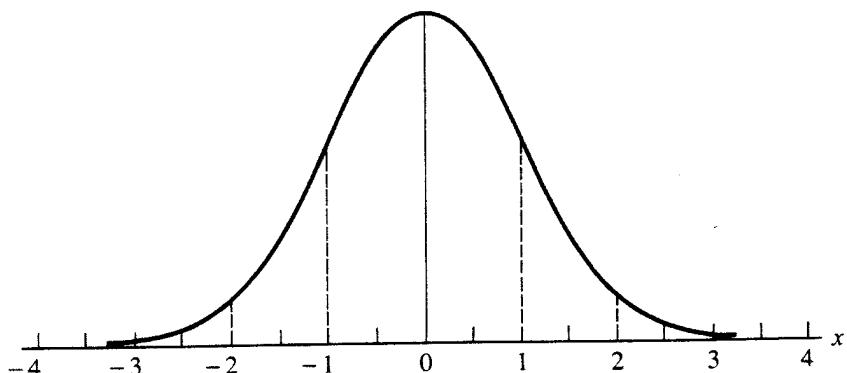


FIGURE 3.2.1 Standard normal density

Among the many uses of the normal distribution, an important one is its use as an approximation to other distributions (which is partially justified by the Central Limit Theorem). For example, if  $X \sim \text{binomial}(n, p)$ , then  $EX = np$  and  $\text{Var } X = np(1 - p)$ , and under suitable conditions, the distribution of  $X$  can be approximated with that of a normal random variable with mean  $\mu = np$  and variance  $\sigma^2 = np(1 - p)$ . The “suitable conditions” are that  $n$  should be large and  $p$  should not be extreme (near 0 or 1). We want  $n$  large so that there are enough (discrete) values of  $X$  to make an approximation by a continuous distribution reasonable and  $p$  should be “in the middle” so the binomial is nearly symmetric, as is the normal. As with most approximations there are no absolute rules, and each application should be checked to decide whether the approximation is good enough for its intended use. A conservative rule to follow is that the approximation will be good if  $\min(np, n(1 - p)) \geq 5$ .

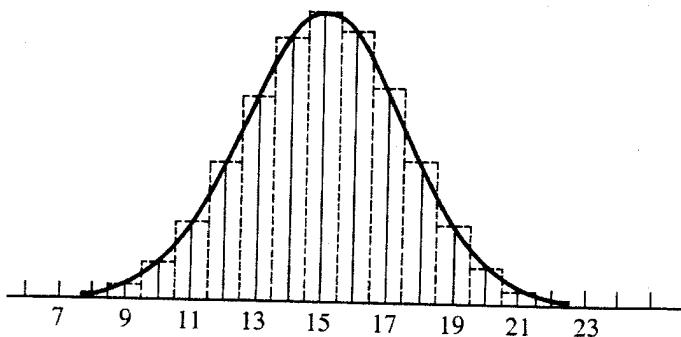
**Example 3.2.2:** Let  $X \sim \text{binomial}(25, .6)$ . We can approximate  $X$  with a normal random variable,  $Y$ , with mean  $\mu = 25(.6) = 15$  and standard deviation  $\sigma = ((25)(.6)(.4))^{1/2} = 2.45$ . Thus

$$P(X \leq 13) \approx P(Y \leq 13) = P\left(Z \leq \frac{13 - 15}{2.45}\right) = P(Z \leq -0.82) = .206,$$

while the exact binomial calculation gives

$$P(X \leq 13) = \sum_{x=0}^{13} \binom{25}{x} (.6)^x (.4)^{25-x} = .267,$$

showing that the normal approximation is good, but not terrific. The approximation can be greatly improved, however, by a “continuity correction.” To see how this works, look at Figure 3.2.2, which shows the  $\text{binomial}(25, .6)$  pmf and the  $n(15, (2.45)^2)$  pdf. We have drawn the binomial pmf using bars of width 1, with height equal to the probability. Thus, the areas of the bars give the binomial probabilities. In the above approximation, notice how the area of the approximating normal is smaller than the binomial area (the normal area is everything to the left of the line at 13, whereas the binomial area includes the entire bar at 13 up to 13.5). The continuity correction adds this area back by adding  $\frac{1}{2}$  to the cutoff point. So instead of approximating



**FIGURE 3.2.2** Normal approximation to the binomial

$P(X \leq 13)$ , we approximate the equivalent expression (because of the discreteness),  $P(X \leq 13.5)$ , and obtain

$$P(X \leq 13) = P(X \leq 13.5) \approx P(Y \leq 13.5) = P(Z \leq -.61) = .271,$$

a much better approximation. In general, the normal approximation with the continuity correction is far superior to the approximation without the continuity correction.

We also make the correction on the lower end. If  $X \sim \text{binomial}(n, p)$ ,  $Y \sim n(np, np(1 - p))$ , then we approximate

$$P(X \leq x) \approx P(Y \leq x + 1/2),$$

$$P(X \geq x) \approx P(Y \geq x - 1/2). \quad ||$$

### Beta Distribution

The beta family of distributions is a continuous family on  $(0, 1)$  indexed by two parameters. The  $\text{beta}(\alpha, \beta)$  pdf is

$$(3.2.16) \quad f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 < x < 1, \quad \alpha > 0, \quad \beta > 0,$$

where  $B(\alpha, \beta)$  denotes the beta function,

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx.$$

The beta function is related to the gamma function through the following identity:

$$(3.2.17) \quad B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

Equation (3.2.17) is very useful in dealing with the beta function, allowing us to take advantage of the properties of the gamma function. In fact, we will never deal directly with the beta function, but rather will use (3.2.17) for all of our evaluations.

The beta distribution is one of the few common “named” distributions that give probability 1 to a finite interval, here taken to be  $(0, 1)$ . As such, the beta is often used

to model proportions, which naturally lie between 0 and 1. We will see illustrations of this in Chapter 4.

Calculation of moments of the beta distribution is quite easy, due to the particular form of the pdf. For  $n > -\alpha$  we have

$$\begin{aligned} EX^n &= \frac{1}{B(\alpha, \beta)} \int_0^1 x^n x^{\alpha-1} (1-x)^{\beta-1} dx \\ &= \frac{1}{B(\alpha, \beta)} \int_0^1 x^{(\alpha+n)-1} (1-x)^{\beta-1} dx. \end{aligned}$$

We now recognize the integrand as the kernel of a  $\text{beta}(\alpha + n, \beta)$  pdf, hence

$$(3.2.18) \quad EX^n = \frac{B(\alpha + n, \beta)}{B(\alpha, \beta)} = \frac{\Gamma(\alpha + n)\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + n)\Gamma(\alpha)}.$$

Using (3.2.3) and (3.2.18) with  $n = 1$  and  $n = 2$ , we calculate the mean and variance of the  $\text{beta}(\alpha, \beta)$  distribution as

$$EX = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \text{Var } X = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

As the parameters  $\alpha$  and  $\beta$  vary, the beta distribution takes on many shapes, as shown in Figure 3.2.3. The pdf can be strictly increasing ( $\alpha > 1, \beta = 1$ ), strictly decreasing ( $\alpha = 1, \beta > 1$ ), U-shaped ( $\alpha < 1, \beta < 1$ ) or unimodal ( $\alpha > 1, \beta > 1$ ). The case  $\alpha = \beta$  yields a pdf symmetric about  $\frac{1}{2}$  with mean  $\frac{1}{2}$  (necessarily) and variance  $(4(2\alpha + 1))^{-1}$ . The pdf becomes more concentrated as  $\alpha$  increases, but stays symmetric, as shown in Figure 3.2.4. Finally, if  $\alpha = \beta = 1$ , the beta distribution reduces to the uniform( $0, 1$ ), showing that the uniform can be considered to be a member of the beta family. The beta distribution is also related, through a

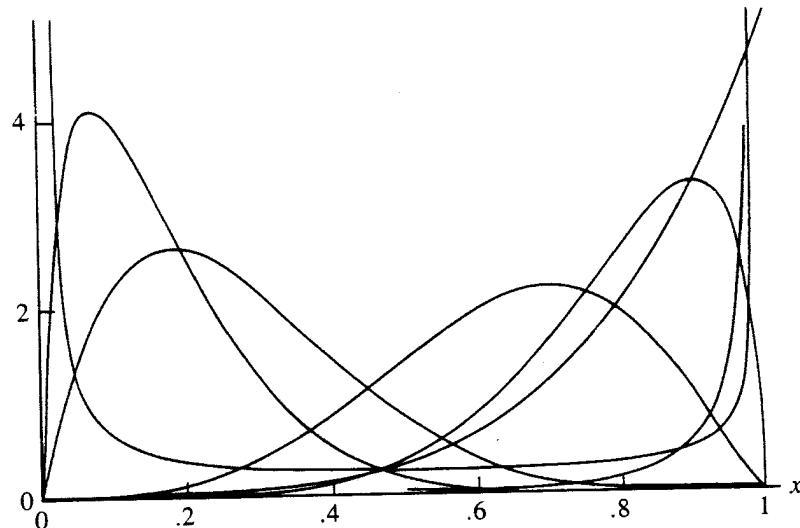


FIGURE 3.2.3 Beta densities

transformation, to the  $F$  distribution, a distribution that plays an extremely important role in statistical analysis (see Section 5.4).

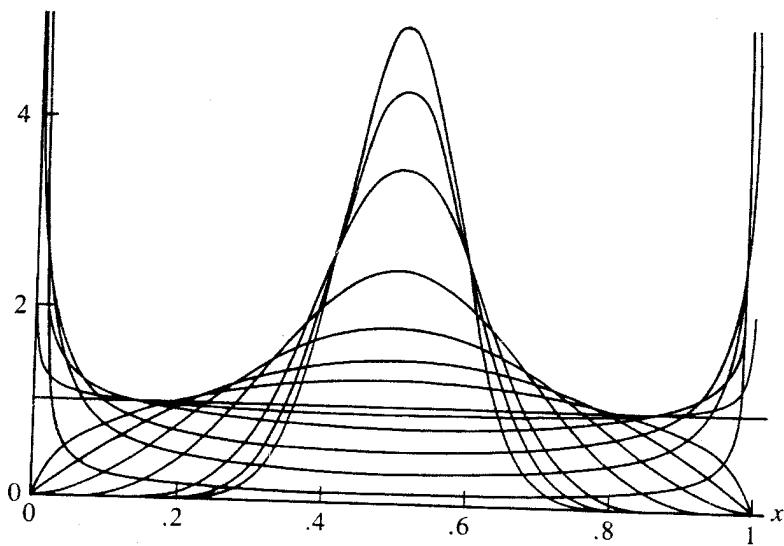


FIGURE 3.2.4 Symmetric beta densities

### *Cauchy Distribution*

The *Cauchy distribution* is a symmetric, bell-shaped distribution on  $(-\infty, \infty)$  with pdf

$$(3.2.19) \quad f(x|\theta) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

(See Exercise 3.34 for a more general version of the Cauchy pdf.) To the eye, the Cauchy does not appear very different from the normal distribution. However, there is a very great difference, indeed. As we have already seen in Chapter 2, the mean of the Cauchy distribution does not exist, that is,

$$(3.2.20) \quad E|X| = \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{|x|}{1 + (x - \theta)^2} dx = \infty.$$

It is easy to see that (3.2.19) defines a proper pdf for all  $\theta$ . Recall that  $\frac{d}{dt} \arctan(t) = (1 + t^2)^{-1}$ , hence

$$\int_{-\infty}^{\infty} \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2} dx = \frac{1}{\pi} \arctan(x - \theta) \Big|_{-\infty}^{\infty} = 1,$$

since  $\arctan(\pm\infty) = \pm\pi/2$ .

Since  $E|X| = \infty$ , it follows that no moments of the Cauchy distribution exist or, in other words, all absolute moments equal  $\infty$ . In particular, the mgf does not exist.

The parameter  $\theta$  in (3.2.19) does measure the center of the distribution; it is the median. If  $X$  has a Cauchy distribution with parameter  $\theta$ , then from Exercise 3.32 it follows that  $P(X \geq \theta) = \frac{1}{2}$ , showing that  $\theta$  is the median of the distribution. Figure 3.2.5 shows a Cauchy(0) distribution together with a  $n(0, 1)$ , where we see the similarity in shape, but the much thicker tails of the Cauchy.

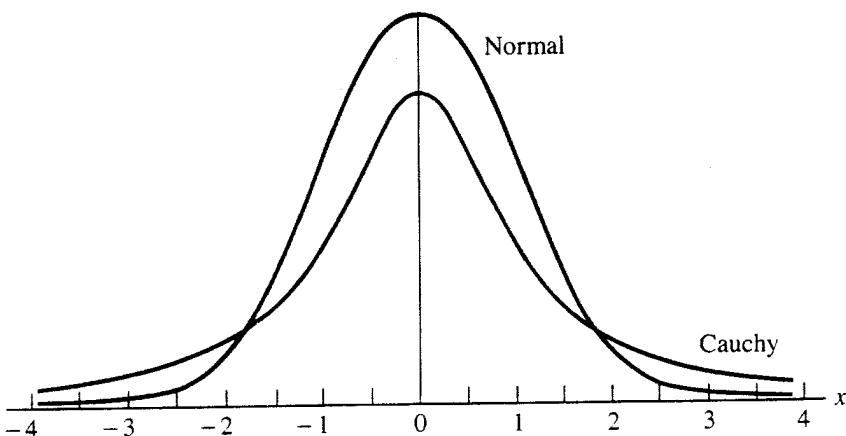


FIGURE 3.2.5 Standard normal density and Cauchy density

The Cauchy distribution plays a special role in the theory of statistics. It represents an extreme case against which conjectures can be tested. But do not make the mistake of considering the Cauchy distribution to be only a pathological case, for it has a way of turning up when you least expect it. For example, it is common practice for experimenters to calculate ratios of observations, that is, ratios of random variables. (In measuring growth, it is common to combine weight and height into one measurement weight-for-height, that is, weight/height.) A surprising fact is that the ratio of two standard normals has a Cauchy distribution (See Example 4.3.4). Taking ratios can lead to ill-behaved distributions.

#### *Lognormal Distribution*

If  $X$  is a random variable whose logarithm is normally distributed (that is,  $\log X \sim n(\mu, \sigma^2)$ ), then  $X$  has a lognormal distribution. The pdf of  $X$  can be obtained by straightforward transformation of the normal pdf using Theorem 2.1.2, yielding

(3.2.21)

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{x} e^{-(\log x - \mu)^2/(2\sigma^2)}, \quad 0 < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0,$$

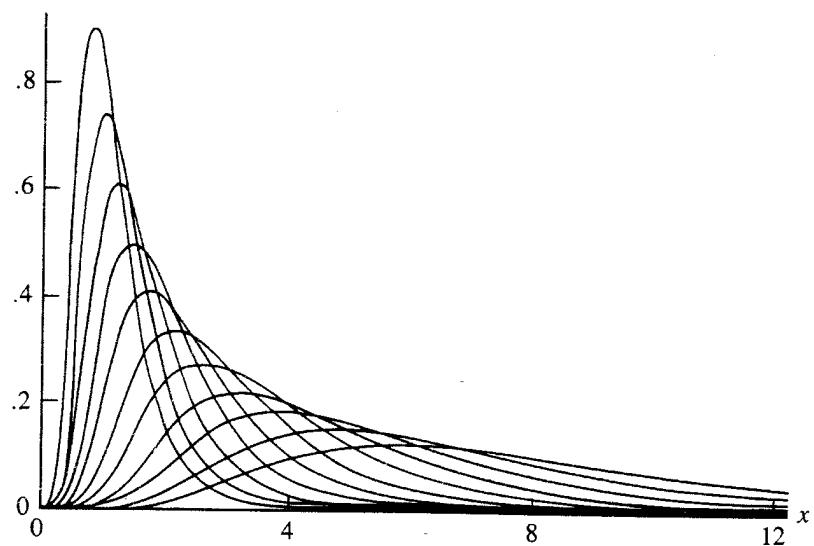
for the *lognormal pdf*. The moments of  $X$  can be calculated directly, using (3.2.21), or by exploiting the relationship to the normal and writing

$$\begin{aligned} EX &= Ee^{\log X} \\ &= Ee^Y && (Y = \log X \sim n(\mu, \sigma^2)) \\ &= e^{\mu + (\sigma^2/2)}. \end{aligned} \tag{3.2.22}$$

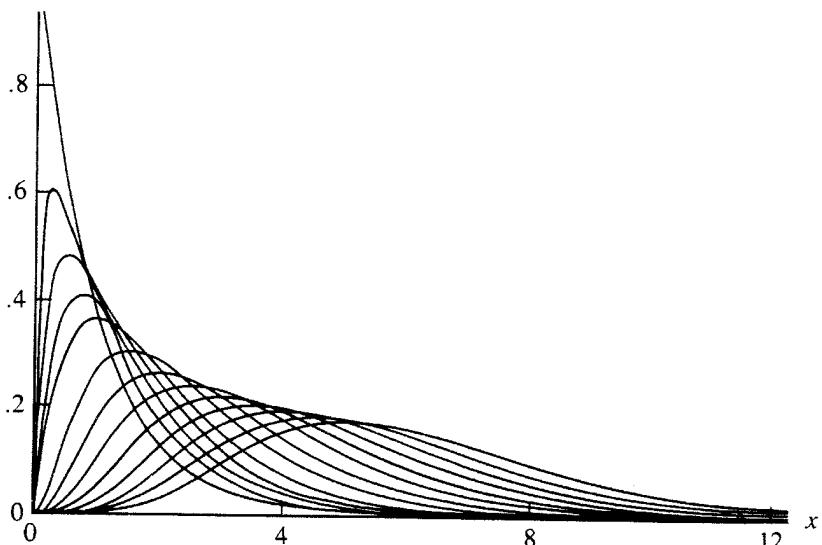
The last equality is obtained by recognizing the mgf of the normal distribution (set  $t = 1$ , see Exercise 2.37). We can use a similar technique to calculate  $EX^2$  and get

$$\text{Var } X = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}.$$

The lognormal distribution is similar in appearance to the gamma distribution, as Figure 3.2.6 shows. The distribution is very popular in modeling applications, when the variable of interest is skewed to the right. For example, incomes are necessarily skewed to the right and modeling with a lognormal allows the use of normal-theory statistics on  $\log(\text{income})$ , a very convenient circumstance.



a.



b.

**FIGURE 3.2.6** a. Some lognormal densities; b. Some gamma densities

### Double Exponential Distribution

The *double exponential distribution* is formed by reflecting the exponential distribution around its mean. The pdf is given by

$$(3.2.23) \quad f(x|\mu, \sigma) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0.$$

The double exponential provides a symmetric distribution with “fat” tails (much fatter than the normal), but still retains all of its moments. It is straightforward to calculate

$$EX = \mu \quad \text{and} \quad \text{Var } X = 2\sigma^2.$$

The double exponential distribution is not bell-shaped. In fact, it has a peak (or more formally, a point of nondifferentiability) at  $x = \mu$ . When dealing with this distribution analytically, it is important to remember this point. The absolute value signs can also be troublesome when performing integrations and it is best to divide the integral into regions around  $x = \mu$ :

$$(3.2.24) \quad \begin{aligned} EX &= \int_{-\infty}^{\infty} \frac{x}{2\sigma} e^{-|x-\mu|/\sigma} dx \\ &= \int_{-\infty}^{\mu} \frac{x}{2\sigma} e^{(x-\mu)/\sigma} dx + \int_{\mu}^{\infty} \frac{x}{2\sigma} e^{-(x-\mu)/\sigma} dx. \end{aligned}$$

Notice that we can remove the absolute value signs over the two regions of integration. (This strategy is useful, in general, in dealing with integrals containing absolute values; divide up the region of integration so the absolute value signs can be removed.) Evaluation of (3.2.24) can be completed by performing integration by parts on each integral.

There are many other continuous distributions that have uses in different statistical applications, many of which will appear throughout the rest of the book. The comprehensive work by Johnson and Kotz (1969, 1970a, 1970b) is a valuable reference for most useful statistical distributions.

## 3.3 Exponential Families

A family of pdfs or pmfs is called an *exponential family* if it can be expressed as

$$(3.3.1) \quad f(x|\theta) = h(x)c(\theta) \exp \left( \sum_{i=1}^k w_i(\theta)t_i(x) \right).$$

Here  $h(x) \geq 0$  and  $t_1(x), \dots, t_k(x)$  are real-valued functions of the observation  $x$  (they cannot depend on  $\theta$ ), and  $c(\theta) \geq 0$  and  $w_1(\theta), \dots, w_k(\theta)$  are real-valued functions of the possibly vector-valued parameter  $\theta$  (they cannot depend on  $x$ ). Many common families introduced in the previous section are exponential families. These include the

continuous families—normal, gamma, and beta; and the discrete families—binomial, Poisson, and negative binomial.

The specific form of (3.3.1) implies that exponential families have many nice mathematical properties. But more importantly for a statistical model, the form of (3.3.1) implies many nice statistical properties, which will be discussed throughout the remainder of the text. For example, suppose we have a large number of data values from a population that has a pdf or pmf of the form (3.3.1). Then only  $k$  numbers ( $k = \text{number of terms in the sum in (3.3.1)}$ ) that can be calculated from the data summarize all the information about  $\theta$  that is in the data. This “data reduction” property is treated in more detail in Chapter 6, where we discuss sufficient statistics.

To verify that a family of pdfs or pmfs is an exponential family, we must identify the functions  $h(x)$ ,  $c(\theta)$ ,  $w_i(\theta)$ , and  $t_i(x)$  and show that the family has the form (3.3.1). The next two examples illustrate this.

**Example 3.3.1:** Let  $n$  be a positive integer and consider the binomial( $n, p$ ) family with  $0 < p < 1$ . Then the pmf for this family, for  $x = 0, \dots, n$ , and  $0 < p < 1$ , is

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} \\ (3.3.2) \quad &= \binom{n}{x} (1-p)^n \left( \frac{p}{1-p} \right)^x \\ &= \binom{n}{x} (1-p)^n \exp \left( \log \left( \frac{p}{1-p} \right) x \right). \end{aligned}$$

Define

$$h(x) = \begin{cases} \binom{n}{x} & x = 0, \dots, n \\ 0 & \text{otherwise} \end{cases}, \quad c(p) = \begin{cases} (1-p)^n & 0 < p < 1 \\ 0 & \text{otherwise} \end{cases},$$

$$w_1(p) = \begin{cases} \log \left( \frac{p}{1-p} \right) & 0 < p < 1 \\ 0 & \text{otherwise} \end{cases}, \quad \text{and} \quad t_1(x) = x.$$

Then we have

$$(3.3.3) \quad f(x|p) = h(x)c(p) \exp[w_1(p)t_1(x)],$$

which is of the form (3.3.1) with  $k = 1$ . In particular, note that  $h(x) > 0$  only if  $x = 0, \dots, n$  and  $c(p) > 0$  only if  $0 < p < 1$ . This is important, as (3.3.3) must match (3.3.2) for *all* values of  $x$  and  $p$ . Also, the parameter values  $p = 0$  and  $1$  are sometimes included in the binomial model, but we have not included them here because the set of  $x$  values for which  $f(x|p) > 0$  is different for  $p = 0$  and  $1$  than for other  $p$  values. ||

**Example 3.3.2:** Let  $f(x|\mu, \sigma^2)$  be the normal( $\mu, \sigma^2$ ) family of pdfs where  $\theta = (\mu, \sigma)$ ,  $-\infty < \mu < \infty$ ,  $\sigma > 0$ . Then

$$\begin{aligned}
 f(x|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
 (3.3.4) \quad &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right).
 \end{aligned}$$

Define

$$\begin{aligned}
 h(x) &= 1 \quad \text{for all } x, \\
 c(\theta) = c(\mu, \sigma) &= \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\mu^2}{2\sigma^2}\right) & -\infty < \mu < \infty, \sigma > 0 \\ 0 & \text{otherwise} \end{cases}, \\
 w_1(\mu, \sigma) &= \begin{cases} \frac{1}{\sigma^2} & \sigma > 0 \\ 0 & \sigma \leq 0 \end{cases}, \quad w_2(\mu, \sigma) = \begin{cases} \frac{\mu}{\sigma^2} & \sigma > 0 \\ 0 & \sigma \leq 0 \end{cases}, \\
 t_1(x) &= -x^2/2, \quad \text{and} \quad t_2(x) = x.
 \end{aligned}$$

Then

$$f(x|\mu, \sigma^2) = h(x)c(\mu, \sigma) \exp[w_1(\mu, \sigma)t_1(x) + w_2(\mu, \sigma)t_2(x)],$$

which is the form (3.3.1) with  $k = 2$ . ||

In general, the set of  $x$  values for which  $f(x|\theta) > 0$  cannot depend on  $\theta$  in an exponential family. The entire definition of the pdf or pmf must be incorporated into the form (3.3.1). This is most easily accomplished by incorporating the range of  $x$  into the expression for  $f(x|\theta)$  through the use of an indicator function.

**DEFINITION 3.3.1:** The *indicator function* of a set  $A$ , most often denoted by  $I_A(x)$ , is the function

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}.$$

Thus, the normal pdf of Example 3.3.2 would be written

$$f(x|\mu, \sigma^2) = h(x)c(\mu, \sigma) \exp[w_1(\mu, \sigma)t_1(x) + w_2(\mu, \sigma)t_2(x)]I_{(-\infty, \infty)}(x).$$

Since the indicator function is a function only of  $x$ , it can be incorporated into the function  $h(x)$ , showing that this pdf is of the form (3.3.1).

From (3.3.1), since the factor  $\exp(\cdot)$  is always positive, it can be seen that for any  $\theta \in \Theta$ , that is, for any  $\theta$  for which  $c(\theta) > 0$ ,  $\{x : f(x|\theta) > 0\} = \{x : h(x) > 0\}$  and this set does not depend on  $\theta$ . So, for example, the set of pdfs given by  $f(x|\theta) = \theta^{-1} \exp(1 - (x/\theta)), 0 < \theta < x < \infty$ , is not an exponential family even though we can write  $\theta^{-1} \exp(1 - (x/\theta)) = h(x)c(\theta) \exp(w(\theta)t(x))$  where  $h(x) = e^1$ ,

$c(\theta) = \theta^{-1}$ ,  $w(\theta) = \theta^{-1}$ , and  $t(x) = -x$ . Writing the pdf with indicator functions makes this very clear. We have

$$f(x|\theta) = \theta^{-1} \exp\left(1 - \left(\frac{x}{\theta}\right)\right) I_{[\theta,\infty)}(x).$$

The indicator function cannot be incorporated into any of the functions of (3.3.1) since it is not a function of  $x$  alone, not a function of  $\theta$  alone, and cannot be expressed as an exponential. Thus, this is not an exponential family.

An exponential family is sometimes reparameterized as

$$f(x|\eta) = h(x)c^*(\eta) \exp\left(\sum_{i=1}^k \eta_i t_i(x)\right).$$

Here the  $h(x)$  and  $t_i(x)$  functions are the same as in the original parameterization (3.3.1). The set  $\mathcal{H} = \{\eta = (\eta_1, \dots, \eta_k) : \int_{-\infty}^{\infty} h(x) \exp\left(\sum_{i=1}^k \eta_i t_i(x)\right) dx < \infty\}$  is called the *natural parameter space* for the family. (The integral is replaced by a sum over the values of  $x$  for which  $h(x) > 0$  if  $X$  is discrete.) For the values of  $\eta \in \mathcal{H}$ , we must have  $c^*(\eta) = \left[\int_{-\infty}^{\infty} h(x) \exp\left(\sum_{i=1}^k \eta_i t_i(x)\right) dx\right]^{-1}$  to ensure that the pdf integrates to 1. Since the original  $f(x|\theta)$  in (3.3.1) is a pdf or pmf, the set  $\{\eta = (w_1(\theta), \dots, w_k(\theta)) : \theta \in \Theta\}$  must be a subset of the natural parameter space. But there may be other values of  $\eta \in \mathcal{H}$  also. The natural parameterization and the natural parameter space have many useful mathematical properties. For example,  $\mathcal{H}$  is convex. A description of these properties may be found in more advanced texts such as Lehmann (1986) or Brown (1986).

**Example 3.3.2 (Continued):** To determine the natural parameter space for the normal family of distributions, replace  $w_i(\mu, \sigma)$  with  $\eta_i$  in (3.3.4) to obtain

$$(3.3.5) \quad f(x|\eta_1, \eta_2) = \frac{\sqrt{\eta_1}}{\sqrt{2\pi}} \exp\left(-\frac{\eta_2^2}{2\eta_1}\right) \exp\left(-\frac{\eta_1 x^2}{2} + \eta_2 x\right).$$

The integral will be finite if and only if the coefficient on  $x^2$  is negative. This means  $\eta_1$  must be positive. If  $\eta_1 > 0$ , the integral will be finite regardless of the value of  $\eta_2$ . Thus the natural parameter space is  $\{(\eta_1, \eta_2) : \eta_1 > 0, -\infty < \eta_2 < \infty\}$ . Identifying (3.3.5) with (3.3.4), we see that  $\eta_2 = \mu/\sigma^2$ ,  $\eta_1 = 1/\sigma^2$ . Although natural parameters provide a convenient mathematical formulation, they sometimes lack simple interpretations like the mean and variance. ||

## 3.4 Location and Scale Families

In Sections 3.2 and 3.3, we discussed several common families of continuous distributions. In this section we discuss three techniques for constructing families of

distributions. The resulting families have ready physical interpretations that make them useful for modeling, as well as convenient mathematical properties.

The three types of families are called location families, scale families, and location-scale families. Each of the families is constructed by specifying a single pdf, say  $f(x)$ , called the *standard pdf* for the family. Then all other pdfs in the family are generated by transforming the standard pdf in a prescribed way. We start with a simple theorem about pdfs.

**THEOREM 3.4.1:** Let  $f(x)$  be any pdf and let  $\mu$  and  $\sigma > 0$  be any given constants. Then the function

$$g(x|\mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

is a pdf.

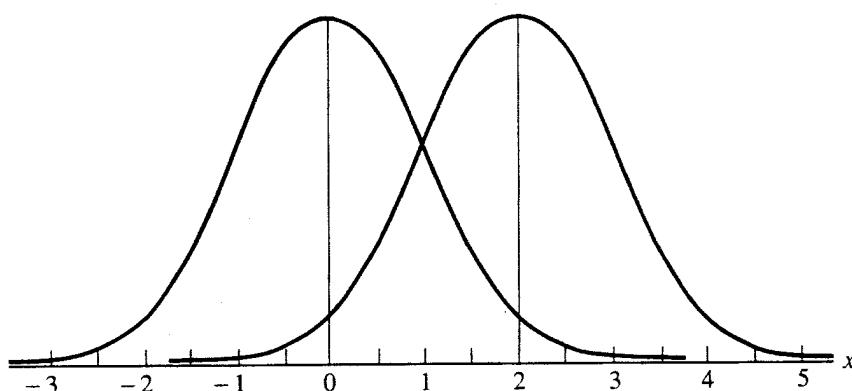
*Proof:* To verify that the transformation has produced a legitimate pdf, we need to check that  $(1/\sigma)f((x-\mu)/\sigma)$ , as a function of  $x$ , is a pdf for every value of  $\mu$  and  $\sigma$  we might substitute into the formula. That is, we must check that  $(1/\sigma)f((x-\mu)/\sigma)$  is nonnegative and integrates to 1. Since  $f(x)$  is a pdf,  $f(x) \geq 0$  for all values of  $x$ . So,  $(1/\sigma)f((x-\mu)/\sigma) \geq 0$  for all values of  $x, \mu$ , and  $\sigma$ . Next we note that

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx &= \int_{-\infty}^{\infty} f(y) dy \quad (\text{substitute } y = \frac{x-\mu}{\sigma}) \\ &= 1, \quad (\text{since } f(y) \text{ is a pdf}) \end{aligned}$$

as was to be verified. □

We now turn to the first of our constructions, that of location families.

**DEFINITION 3.4.1:** Let  $f(x)$  be any pdf. Then the family of pdfs  $f(x-\mu)$ , indexed by the parameter  $\mu$ ,  $-\infty < \mu < \infty$ , is called the *location family with standard pdf*  $f(x)$  and  $\mu$  is called the *location parameter* for the family.



**FIGURE 3.4.1** Two members of the same location family: Means at 0 and 2

To see the effect of introducing the location parameter  $\mu$ , consider Figure 3.4.1. At  $x = \mu$ ,  $f(x - \mu) = f(0)$ ; at  $x = \mu + 1$ ,  $f(x - \mu) = f(1)$ ; and, in general, at  $x = \mu + a$ ,  $f(x - \mu) = f(a)$ . Of course,  $f(x - \mu)$  for  $\mu = 0$  is just  $f(x)$ . Thus the location parameter  $\mu$  simply shifts the pdf  $f(x)$  so that the shape of the graph is unchanged but the point on the graph that was above  $x = 0$  for  $f(x)$  is above  $x = \mu$  for  $f(x - \mu)$ . It is clear from Figure 3.4.1 that the area under the graph of  $f(x)$  between  $x = -1$  and  $x = 2$  is the same as the area under the graph of  $f(x - \mu)$  between  $x = \mu - 1$  and  $x = \mu + 2$ . Thus if  $X$  is a random variable with pdf  $f(x - \mu)$  we can write

$$P(-1 \leq X \leq 2|0) = P(\mu - 1 \leq X \leq \mu + 2|\mu),$$

where the random variable  $X$  has pdf  $f(x - 0) = f(x)$  on the left of the equality and pdf  $f(x - \mu)$  on the right.

Several of the families introduced in Section 3.2 are, or have as subfamilies, location families. For example, if  $\sigma > 0$  is a specified, known number and we define

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}, \quad -\infty < x < \infty,$$

then the location family with standard pdf  $f(x)$  is the set of normal distributions with unknown mean  $\mu$  and known variance  $\sigma^2$ . To see this, check that replacing  $x$  by  $x - \mu$  in the above formula yields pdfs of the form defined in (3.2.13). Similarly, the Cauchy family and the double exponential family, with  $\sigma$  a specified value and  $\mu$  a parameter, are examples of location families. But the point of Definition 3.4.1 is that we can start with *any* pdf  $f(x)$  and generate a family of pdfs by introducing a location parameter.

If  $X$  is a random variable with pdf  $f(x - \mu)$ , then  $X$  may be represented as  $X = Z + \mu$ , where  $Z$  is a random variable with pdf  $f(z)$ . This representation is a consequence of Theorem 3.4.2 (with  $\sigma = 1$ ) which will be proved later. Consideration of this representation indicates when a location family might be an appropriate model for an observed variable  $X$ . We will describe two such situations.

First, suppose an experiment is designed to measure some physical constant  $\mu$ , say the temperature of a solution. But there is some measurement error involved in the observation. So the actual observed value  $X$  is  $Z + \mu$  where  $Z$  is the measurement error.  $X$  will be greater than  $\mu$  if  $Z > 0$  for this observation and less than  $\mu$  if  $Z < 0$ . The distribution of the random measurement error might be well known from previous experience in using this measuring device to measure other solutions. If this distribution has pdf  $f(z)$  then the pdf of the observed value  $X$  is  $f(x - \mu)$ .

As another example, suppose the distribution of reaction times of drivers on a coordination test is known from previous experimentation. Denote the reaction time for a randomly chosen driver by the random variable  $Z$ . Let the pdf of  $Z$  describing the known distribution be  $f(z)$ . Now, consider "applying a treatment" to the population. For example, consider what would happen if everyone drank three glasses of beer. We might assume that everyone's reaction time would change by some unknown amount  $\mu$ . (This very simple model, in which everyone's reaction time changes by the same

amount  $\mu$ , is probably not the best model. For example, it is known that the effect of alcohol is weight-dependent, so heavier people are likely to be less affected by the beers.) Being open-minded scientists, we might even allow the possibility that  $\mu < 0$ , that is, that the reaction times decrease. Then, if we observe the reaction time of a randomly selected driver after “treatment,” the reaction time would be  $X = Z + \mu$  and the family of possible distributions for  $X$  would be given by  $f(x - \mu)$ .

If the set of  $x$  for which  $f(x) > 0$  is not the whole real line, then the set of  $x$  for which  $f(x - \mu) > 0$  will depend on  $\mu$ . Example 3.4.1 illustrates this.

**Example 3.4.1:** Let  $f(x) = e^{-x}$ ,  $x \geq 0$ , and  $f(x) = 0$ ,  $x < 0$ . To form a location family we replace  $x$  with  $x - \mu$  to obtain

$$\begin{aligned} f(x|\mu) &= \begin{cases} e^{-(x-\mu)} & x - \mu \geq 0 \\ 0 & x - \mu < 0 \end{cases} \\ &= \begin{cases} e^{-(x-\mu)} & x \geq \mu \\ 0 & x < \mu \end{cases}. \end{aligned}$$

Graphs of  $f(x|\mu)$ , for various values of  $\mu$ , are shown in Figure 3.4.2. As in Figure 3.4.1, the graph has been shifted. Now the positive part of the graph starts at  $\mu$  rather than at zero. If  $X$  measures time, then  $\mu$  might be restricted to be nonnegative so that  $X$  will be positive with probability 1 for every value of  $\mu$ . In this type of model, where  $\mu$  denotes a bound on the range of  $X$ ,  $\mu$  is sometimes called a *threshold parameter*.

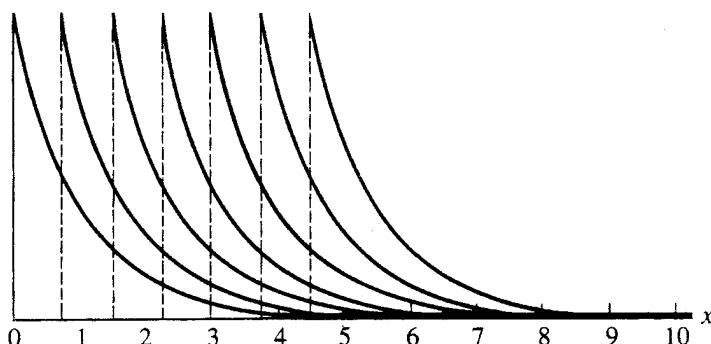


FIGURE 3.4.2 Exponential location densities

The other two types of families to be discussed in this section are scale families and location-scale families.

**DEFINITION 3.4.2:** Let  $f(x)$  be any pdf. Then for any  $\sigma > 0$ , the family of pdfs  $(1/\sigma)f(x/\sigma)$ , indexed by the parameter  $\sigma$ , is called the *scale family with standard pdf*  $f(x)$  and  $\sigma$  is called the *scale parameter* of the family.

The effect of introducing the scale parameter  $\sigma$  is either to stretch ( $\sigma > 1$ ) or to contract ( $\sigma < 1$ ) the graph of  $f(x)$  while still maintaining the same basic shape of the graph. This is illustrated in Figure 3.4.3. Most often when scale parameters are

used,  $f(x)$  is either symmetric about 0 or positive only for  $x > 0$ . In these cases the stretching is either symmetric about 0 or only in the positive direction. But, in the definition, any pdf may be used as the standard.

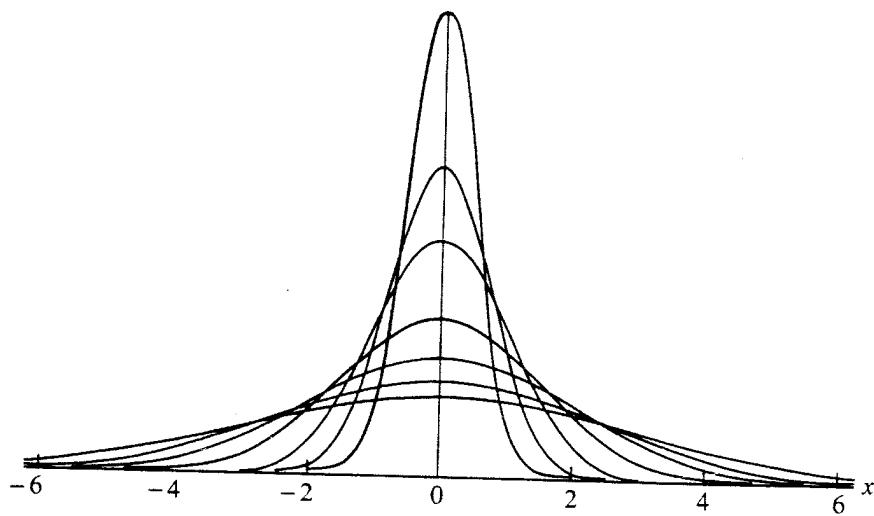


FIGURE 3.4.3 Members of the same scale family

Several of the families introduced in Section 3.2 either are scale families or have scale families as subfamilies. These are the gamma family if  $\alpha$  is a fixed value and  $\beta$  is the scale parameter, the normal family if  $\mu = 0$  and  $\sigma$  is the scale parameter, the exponential family, and the double exponential family if  $\mu = 0$  and  $\sigma$  is the scale parameter. In each case the standard pdf is the pdf obtained by setting the scale parameter equal to 1. Then all other members of the family can be shown to be of the form in Definition 3.4.2.

**DEFINITION 3.4.3:** Let  $f(x)$  be any pdf. Then for any  $\mu$ ,  $-\infty < \mu < \infty$ , and any  $\sigma > 0$ , the family of pdfs  $(1/\sigma)f((x - \mu)/\sigma)$ , indexed by the parameter  $(\mu, \sigma)$ , is called the *location-scale family with standard pdf  $f(x)$* ;  $\mu$  is called the *location parameter* and  $\sigma$  is called the *scale parameter*.

The effect of introducing both the location and scale parameters is to stretch ( $\sigma > 1$ ) or contract ( $\sigma < 1$ ) the graph with the scale parameter and then shift the graph so that the point that was above 0 is now above  $\mu$ . Figure 3.4.4 (page 120) illustrates this transformation of  $f(x)$ . The normal and double exponential families are examples of location-scale families. Exercise 3.34 presents the Cauchy as a location-scale family.

The following theorem relates the transformation of the pdf  $f(x)$  that defines a location-scale family to the transformation of a random variable  $Z$  with pdf  $f(z)$ . As mentioned earlier in the discussion of location families, the representation in terms of  $Z$  is a useful mathematical tool and can help us understand when a location-scale family might be appropriate in a modeling context. Setting  $\sigma = 1$  in Theorem 3.4.2 yields a result for location (only) families and setting  $\mu = 0$  yields a result for scale (only) families.

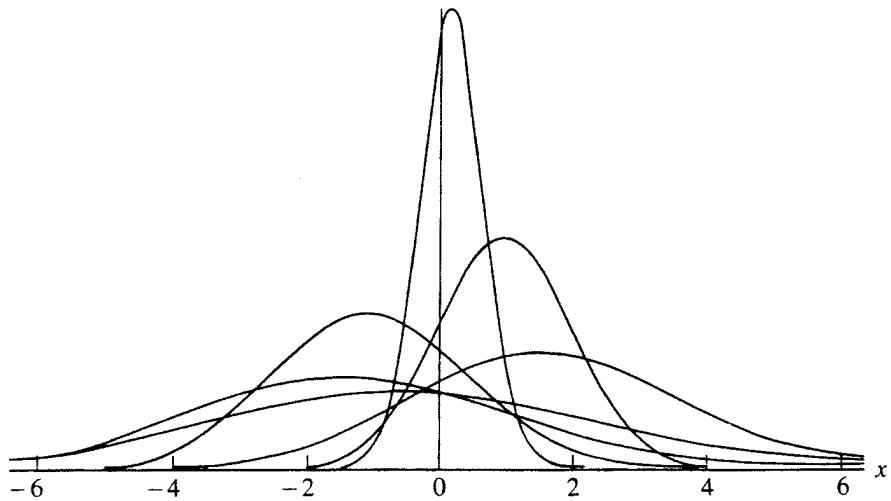


FIGURE 3.4.4 Members of the same location-scale family

**THEOREM 3.4.2:** Let  $f(\cdot)$  be any pdf. Let  $\mu$  be any real number and let  $\sigma$  be any positive real number. Then  $X$  is a random variable with pdf  $(1/\sigma)f((x - \mu)/\sigma)$  if and only if there exists a random variable  $Z$  with pdf  $f(z)$  and  $X = \sigma Z + \mu$ .

*Proof:* To prove the “if” part, define  $g(z) = \sigma z + \mu$ . Then  $X = g(Z)$ ,  $g$  is a monotone function,  $g^{-1}(x) = (x - \mu)/\sigma$ , and  $|(d/dx)g^{-1}(x)| = 1/\sigma$ . Thus by Theorem 2.1.2, the pdf of  $X$  is

$$f_X(x) = f_Z(g^{-1}(x)) \left| \frac{d}{dx} g^{-1}(x) \right| = f\left(\frac{x - \mu}{\sigma}\right) \frac{1}{\sigma}.$$

To prove the “only if” part, define  $g(x) = (x - \mu)/\sigma$  and let  $Z = g(X)$ . Theorem 2.1.2 again applies,  $g^{-1}(z) = \sigma z + \mu$ ,  $|(d/dz)g^{-1}(z)| = \sigma$ , and the pdf of  $Z$  is

$$f_Z(z) = f_X(g^{-1}(z)) \left| \frac{d}{dz} g^{-1}(z) \right| = \frac{1}{\sigma} f\left(\frac{(\sigma z + \mu) - \mu}{\sigma}\right) \sigma = f(z).$$

Also,

$$\sigma Z + \mu = \sigma g(X) + \mu = \sigma \left( \frac{X - \mu}{\sigma} \right) + \mu = X. \quad \square$$

An important fact to extract from Theorem 3.4.2 is that the random variable  $Z = (X - \mu)/\sigma$  has pdf

$$f_Z(z) = \frac{1}{1} f\left(\frac{z - 0}{1}\right) = f(z).$$

That is, the distribution of  $Z$  is that member of the location-scale family corresponding to  $\mu = 0, \sigma = 1$ . This was already proved for the special case of the normal family in Section 3.2.

Often, calculations can be carried out for the “standard” random variable  $Z$  with pdf  $f(z)$  and then the corresponding result for the random variable  $X$  with pdf  $(1/\sigma)f((x - \mu)/\sigma)$  can be easily derived. An example is given in the following, which is a generalization of a computation done in Section 3.2 for the normal family.

**THEOREM 3.4.3:** Let  $Z$  be a random variable with pdf  $f(z)$ . Suppose  $EZ$  and  $\text{Var } Z$  exist. If  $X$  is a random variable with pdf  $(1/\sigma)f((x - \mu)/\sigma)$ , then

$$EX = \sigma EZ + \mu \quad \text{and} \quad \text{Var } X = \sigma^2 \text{Var } Z.$$

In particular, if  $EZ = 0$  and  $\text{Var } Z = 1$ , then  $EX = \mu$  and  $\text{Var } X = \sigma^2$ .

*Proof:* By Theorem 3.4.2, there is a random variable  $Z^*$  with pdf  $f(z)$  and  $X = \sigma Z^* + \mu$ . So  $EX = \sigma EZ^* + \mu = \sigma EZ + \mu$  and  $\text{Var } X = \sigma^2 \text{Var } Z^* = \sigma^2 \text{Var } Z$ .  $\square$

For any location-scale family with a finite mean and variance, the standard pdf  $f(z)$  can be chosen in such a way that  $EZ = 0$  and  $\text{Var } Z = 1$ . (The proof that this choice can be made is left as Exercise 3.35.) This results in the convenient interpretation of  $\mu$  and  $\sigma^2$  as the mean and variance of  $X$ , respectively. This is the case for the usual definition of the normal family as given in Section 3.2. However, this is not the choice for the usual definition of the double exponential family as given in Section 3.2. There,  $\text{Var } Z = 2$ .

Probabilities for any member of a location-scale family may be computed in terms of the standard variable  $Z$  because

$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P\left(Z \leq \frac{x - \mu}{\sigma}\right).$$

Thus, if  $P(Z \leq z)$  is tabulated or easily calculable for the standard variable  $Z$ , then probabilities for  $X$  may be obtained. Calculations of normal probabilities using the standard normal table are examples of this.

## EXERCISES

---

- 3.1 Find expressions for  $EX$  and  $\text{Var } X$  if  $X$  is a random variable with the general discrete uniform( $N_0, N_1$ ) distribution that puts equal probability on each of the values  $N_0, N_0 + 1, \dots, N_1$ . Here  $N_0 \leq N_1$  and both are integers.
- 3.2 A manufacturer receives a lot of 100 parts from a vendor. The lot will be unacceptable if more than five of the parts are defective. The manufacturer is going to select randomly  $K$  parts from the lot for inspection and the lot will be accepted if no defective parts are found in the sample.
  - a. How large does  $K$  have to be to ensure that the probability that the manufacturer accepts an unacceptable lot is less than .10?
  - b. Suppose the manufacturer decides to accept the lot if there is at most one defective in the sample. How large does  $K$  have to be to ensure that the probability that the manufacturer accepts an unacceptable lot is less than .10?

- 3.3 The flow of traffic at certain street corners can sometimes be modeled as a sequence of Bernoulli trials by assuming that the probability of a car passing during any given second is a constant  $p$  and that there is no interaction between the passing of cars at different seconds. Treating seconds as indivisible time units (trials), the Bernoulli model applies. Suppose a pedestrian can cross the street only if no car is to pass during the next 3 seconds. Find the probability that the pedestrian has to wait for exactly 4 seconds before starting to cross.
- 3.4 A man with  $n$  keys wants to open his door and tries the keys at random. Exactly one key will open the door. Find the mean number of trials if
- unsuccessful keys are not eliminated from further selections
  - unsuccessful keys are eliminated.
- 3.5 A standard drug is known to be effective in 80% of the cases in which it is used. A new drug is tested on 100 patients and found to be effective in 85 cases. Is the new drug superior? (*Hint:* Evaluate the probability of observing 85 or more successes assuming that the new and old drugs are equally effective.)
- 3.6 A large number of insects are expected to be attracted to a certain variety of rose plant. A commercial insecticide is advertised as being 99% effective. Suppose 2,000 insects infest a rose garden where the insecticide has been applied, and let  $X = \#$  of surviving insects.
- What probability distribution might provide a reasonable model for this experiment?
  - Write down, but do not evaluate, an expression for the probability that fewer than 100 insects survive, using the model in part (a).
  - Evaluate an approximation to the probability in part (b).
- 3.7 Let the number of chocolate chips in a certain type of cookie have a Poisson distribution. We want the probability that a randomly chosen cookie has at least two chocolate chips to be greater than .99. Find the smallest value of the mean of the distribution that ensures this probability.
- 3.8 Two movie theaters compete for the business of 1,000 customers. Assume that each customer chooses between the movie theaters independently and with “indifference.” Let  $N$  denote the number of seats in each theater.
- Using a binomial model, find an expression for  $N$  that will guarantee that the probability of turning away a customer (because of a full house) is less than 1%.
  - Use the normal approximation to get a numerical value for  $N$ .
- 3.9 The hypergeometric distribution can be approximated by either the binomial or the Poisson distribution. (Of course, it can be approximated by other distributions, but in this exercise we will concentrate on only these two.) Let  $X$  have the hypergeometric distribution

$$P(X = x|N, M, K) = \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}}, \quad x = 0, 1, \dots, K.$$

- a. Show that as  $N \rightarrow \infty$ ,  $M \rightarrow \infty$ , and  $M/N \rightarrow p$ ,

$$P(X = x|N, M, K) \rightarrow \binom{K}{x} p^x (1-p)^{K-x}, \quad x = 0, 1, \dots, K.$$

(Stirling’s formula (Exercise 1.23) may be helpful.)

- b. Use the fact that the binomial can be approximated by the Poisson to show that if  $N \rightarrow \infty$ ,  $M \rightarrow \infty$ ,  $K \rightarrow \infty$ ,  $M/N \rightarrow 0$ , and  $KM/N \rightarrow \lambda$ , then

$$P(X = x | N, M, K) \rightarrow \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, \dots$$

c. Verify the approximation in part (b) directly, without using the Poisson approximation to the binomial. (Lemma 2.3.1 is helpful.)

- 3.10** Suppose  $X$  has a binomial( $n, p$ ) distribution and let  $Y$  have a negative binomial( $r, p$ ) distribution. Show that  $F_X(r - 1) = 1 - F_Y(n - r)$ .

- 3.11** A *truncated* discrete distribution is one in which a particular class cannot be observed, and is eliminated from the sample space. In particular, if  $X$  has range  $0, 1, 2, \dots$ , and the 0 class cannot be observed (as is usually the case), the *0-truncated* random variable  $X_T$  has pmf

$$P(X_T = x) = \frac{P(X = x)}{P(X > 0)}, \quad x = 1, 2, \dots$$

Find the pmf, mean, and variance of the 0-truncated random variable starting from

- a.  $X \sim \text{Poisson}(\lambda)$
- b.  $X \sim \text{negative binomial}(r, p)$ , as in (3.1.10)

- 3.12** Starting from the 0-truncated negative binomial (refer to Exercise 3.11), if we let  $r \rightarrow 0$  we get an interesting distribution, the *logarithmic series distribution*. A random variable  $X$  has a logarithmic series distribution with parameter  $p$  if

$$P(X = x) = \frac{-(1-p)^x}{x \log p}, \quad x = 1, 2, \dots, \quad 0 < p < 1.$$

- a. Verify that this defines a legitimate probability function.
- b. Find the mean and variance of  $X$ .

(The logarithmic series distribution has proved useful in modeling species abundance. See Stuart and Ord (1987) for a more detailed discussion of this distribution.)

- 3.13** In Section 3.1 it was claimed that the Poisson( $\lambda$ ) distribution is the limit of the negative binomial( $r, p$ ) distribution as  $r \rightarrow \infty$ ,  $p \rightarrow 1$ , and  $r(1-p) \rightarrow \lambda$ . Show that under these conditions the mgf of the negative binomial converges to that of the Poisson.
- 3.14** Verify these two identities regarding the gamma function that were given in the text.

- a.  $\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$
- b.  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

- 3.15** Establish a formula similar to (3.2.18) for the gamma distribution. If  $X \sim \text{gamma}(\alpha, \beta)$ , then for any positive constant  $\nu$

$$EX^\nu = \frac{\beta^\nu \Gamma(\nu + \alpha)}{\Gamma(\alpha)}.$$

- 3.16** There is an interesting relationship between negative binomial and gamma random variables, which may sometimes provide a useful approximation. Let  $Y$  be a negative binomial random variable with parameters  $r$  and  $p$ , where  $p$  is the success probability. Show that as  $p \rightarrow 0$ , the mgf of the random variable  $pY$  converges to that of a gamma distribution with parameters  $r$  and 1.

- 3.17** Show that

$$\int_x^\infty \frac{1}{\Gamma(\alpha)} z^{\alpha-1} e^{-z} dz = \sum_{y=0}^{\alpha-1} \frac{x^y e^{-x}}{y!}, \quad \alpha = 1, 2, 3, \dots$$

(Hint: Use integration by parts.) Express this formula as a probabilistic relationship between Poisson and gamma random variables.

- 3.18 Let the random variable  $X$  have the pdf

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, \quad 0 < x < \infty.$$

- a. Find the mean and variance of  $X$ . (This distribution is sometimes called a *folded normal*.)
- b. If  $X$  has the folded normal distribution, find the transformation  $g(X) = Y$ , and values of  $\alpha$  and  $\beta$ , so that  $Y \sim \text{gamma}(\alpha, \beta)$ .
- 3.19 Let  $X \sim n(\mu, \sigma^2)$ . Find values of  $\mu$  and  $\sigma^2$  such that  $P(|X| < 2) = \frac{1}{2}$ . Prove or disprove that these values of  $\mu$  and  $\sigma^2$  are unique.
- 3.20 Express the integral formula given in Exercise 2.40 as a probabilistic relationship between binomial and beta random variables.
- 3.21 Write the integral that would define the mgf of the pdf

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}.$$

Is the integral finite? (Do you expect it to be?)

- 3.22 For each of the following distributions, verify the formulas for  $EX$  and  $\text{Var } X$  given in the text.
  - a. Verify  $\text{Var } X$  if  $X$  has a  $\text{Poisson}(\lambda)$  distribution. (Hint: Compute  $EX(X - 1) = EX^2 - EX$ .)
  - b. Verify  $\text{Var } X$  if  $X$  has a negative binomial( $r, p$ ) distribution.
  - c. Verify  $\text{Var } X$  if  $X$  has a  $\text{gamma}(\alpha, \beta)$  distribution.
  - d. Verify  $EX$  and  $\text{Var } X$  if  $X$  has a  $\text{beta}(\alpha, \beta)$  distribution.
  - e. Verify  $EX$  and  $\text{Var } X$  if  $X$  has a double exponential( $\mu, \sigma$ ) distribution.
- 3.23 The *Pareto distribution*, with parameters  $\alpha$  and  $\beta$ , has pdf

$$f(x) = \frac{\beta\alpha^\beta}{x^{\beta+1}}, \quad \alpha < x < \infty, \quad \alpha > 0, \quad \beta > 0.$$

- a. Verify that  $f(x)$  is a pdf.
- b. Derive the mean and variance of this distribution.
- c. Prove that the variance does not exist if  $\beta \leq 2$ .
- 3.24 Many “named” distributions are special cases of the more common distributions already discussed. For each of the following named distributions derive the form of the pdf, verify that it is a pdf, and calculate the mean and variance.
  - a. If  $X \sim \text{exponential}(\beta)$ , then  $Y = X^{1/\gamma}$  has the *Weibull*( $\gamma, \beta$ ) *distribution*, where  $\gamma > 0$  is a constant.
  - b. If  $X \sim \text{exponential}(\beta)$ , then  $Y = (2X/\beta)^{1/2}$  has the *Rayleigh distribution*.
  - c. If  $X \sim \text{gamma}(\frac{3}{2}, \beta)$ , then  $Y = (X/\beta)^{1/2}$  has the *Maxwell distribution*.
  - d. If  $X \sim \text{exponential}(1)$ , then  $Y = \alpha - \gamma \log X$  has the *Gumbel*( $\alpha, \gamma$ ) *distribution*, where  $-\infty < \alpha < \infty$  and  $\gamma > 0$ . (The Gumbel distribution is also known as the *extreme value distribution*.)
- 3.25 Suppose the random variable  $T$  is the length of life of an object (possibly the lifetime of an electrical component or of a subject given a particular treatment). The *hazard function*  $h_T(t)$ , associated with the random variable  $T$ , is defined by

$$h_T(t) = \lim_{\delta \rightarrow 0} \frac{P(t \leq T < t + \delta | T \geq t)}{\delta}.$$

Thus, we can interpret  $h_T(t)$  as the rate of change of the probability that the object survives a little past time  $t$ , given that the object survives to time  $t$ . Show that if  $T$  is a continuous random variable, then

$$h_T(t) = \frac{f_T(t)}{1 - F_T(t)} = -\frac{d}{dt} \log(1 - F_T(t)).$$

**3.26** Verify that the following pdfs have the indicated hazard functions (see Exercise 3.25).

- a. If  $T \sim \text{exponential}(\beta)$ , then  $h_T(t) = 1/\beta$ .
- b. If  $T \sim \text{Weibull}(\gamma, \beta)$ , then  $h_T(t) = (\gamma/\beta)t^{\gamma-1}$ .
- c. If  $T \sim \text{logistic}(\mu, \beta)$ , that is,

$$F_T(t) = \frac{1}{1 + e^{-(t-\mu)/\beta}},$$

then  $h_T(t) = (1/\beta)F_T(t)$ .

**3.27** For each of the following families, show whether all the pdfs in the family are unimodal (see Exercise 2.28).

- a. uniform( $a, b$ )
- b. gamma( $\alpha, \beta$ )
- c. n( $\mu, \sigma^2$ )
- d. beta( $\alpha, \beta$ )

**3.28** Show that each of the following families is an exponential family.

- a. normal family with either parameter  $\mu$  or  $\sigma$  known
- b. gamma family with either parameter  $\alpha$  or  $\beta$  known or both unknown
- c. beta family with either parameter  $\alpha$  or  $\beta$  known or both unknown
- d. Poisson family
- e. negative binomial family with  $r$  known,  $0 < p < 1$

**3.29** For each family in Exercise 3.28, describe the natural parameter space.

**3.30** Consider an exponential family expressed in terms of its natural parameter space. Show that

$$E_\eta t_i(X) = -\frac{\partial}{\partial \eta_i} \log(c(\eta)).$$

You may use the fact that, for an exponential family,

$$\frac{\partial^j}{\partial \eta_i^j} \int_{-\infty}^{\infty} f_\eta(x) dx = \int_{-\infty}^{\infty} \frac{\partial^j}{\partial \eta_i^j} f_\eta(x) dx.$$

**3.31** Consider the pdf  $f(x) = \frac{63}{4}(x^6 - x^8)$ ,  $-1 < x < 1$ . Graph  $(1/\sigma)f((x - \mu)/\sigma)$  for each of the following on the same axes:

- a.  $\mu = 0, \sigma = 1$
- b.  $\mu = 3, \sigma = 1$
- c.  $\mu = 3, \sigma = 2$

**3.32** Show that if  $f(x)$  is a pdf, symmetric about 0, then  $\mu$  is the median of the location-scale pdf  $(1/\sigma)f((x - \mu)/\sigma)$ ,  $-\infty < x < \infty$ .

**3.33** Let  $Z$  be a random variable with pdf  $f(z)$ . Define  $z_\alpha$  to be a number that satisfies this relationship:

$$\alpha = P(Z > z_\alpha) = \int_{z_\alpha}^{\infty} f(z) dz.$$

Show that if  $X$  is a random variable with pdf  $(1/\sigma)f((x - \mu)/\sigma)$  and  $x_\alpha = \sigma z_\alpha + \mu$ , then  $P(X > x_\alpha) = \alpha$ . (Thus if a table of  $z_\alpha$  values were available, then values of  $x_\alpha$  could be easily computed for any member of the location-scale family.)

- 3.34 Consider the Cauchy family defined in Section 3.2. This family can be extended to a location-scale family yielding pdfs of the form

$$f(x|\mu, \sigma) = \frac{1}{\sigma\pi \left(1 + \left(\frac{x-\mu}{\sigma}\right)^2\right)}, \quad -\infty < x < \infty.$$

The mean and variance do not exist for the Cauchy distribution. So the parameters  $\mu$  and  $\sigma^2$  are not the mean and variance. But they do have important meaning. Show that if  $X$  is a random variable with a Cauchy distribution with parameters  $\mu$  and  $\sigma$  then

- a.  $\mu$  is the median of the distribution of  $X$ , that is,  $P(X \geq \mu) = P(X \leq \mu) = \frac{1}{2}$ .
- b.  $\mu + \sigma$  and  $\mu - \sigma$  are the quartiles of the distribution of  $X$ , that is,  $P(X \geq \mu + \sigma) = P(X \leq \mu - \sigma) = \frac{1}{4}$ .

(Hint: Prove this first for  $\mu = 0$  and  $\sigma = 1$  and then use Exercise 3.33.)

- 3.35 Let  $f(x)$  be any pdf with mean  $\mu$  and variance  $\sigma^2$ . Show how to create a location-scale family based on  $f(x)$  such that the standard pdf of the family, say  $f^*(x)$ , has mean 0 and variance 1.
- 3.36 A family of cdfs  $\{F(x|\theta), \theta \in \Theta\}$  is *stochastically increasing in  $\theta$*  if  $\theta_1 > \theta_2 \Rightarrow F(x|\theta_1)$  is stochastically greater than  $F(x|\theta_2)$ . (See Exercise 1.56 for the definition of stochastically greater.)
- a. Show that the  $n(\mu, \sigma^2)$  family is stochastically increasing in  $\mu$  for fixed  $\sigma^2$ .
  - b. Show that the gamma( $\alpha, \beta$ ) family of (3.2.6) is stochastically increasing in  $\beta$  (scale parameter) for fixed  $\alpha$  (shape parameter).
- 3.37 Refer to Exercise 3.36 for the definition of a stochastically increasing family.
- a. Show that a location family is stochastically increasing in its location parameter.
  - b. Show that a scale family is stochastically increasing in its scale parameter if the sample space is  $[0, \infty)$ .
- 3.38 A family of cdfs  $\{F(x|\theta), \theta \in \Theta\}$  is *stochastically decreasing in  $\theta$*  if  $\theta_1 > \theta_2 \Rightarrow F(x|\theta_2)$  is stochastically greater than  $F(x|\theta_1)$ . (See Exercises 3.36 and 3.37.)
- a. Prove that if  $X \sim F_X(x|\theta)$ , where the sample space of  $X$  is  $(0, \infty)$ , and  $F_X(x|\theta)$  is stochastically increasing in  $\theta$ , then  $F_Y(y|\theta)$  is stochastically decreasing in  $\theta$ , where  $Y = 1/X$ .
  - b. Prove that if  $X \sim F_X(x|\theta)$ , where  $F_X(x|\theta)$  is stochastically increasing in  $\theta$  and  $\theta > 0$ , then  $F_X(x|\frac{1}{\theta})$  is stochastically decreasing in  $\theta$ .

## Miscellanea

---

### The Poisson postulates

The Poisson distribution can be derived from a set of basic assumptions, sometimes called the Poisson postulates. These assumptions relate to the physical properties of the process under consideration. While, generally speaking, the assumptions are not very easy to verify, they

do provide an experimenter with a set of guidelines for considering whether the Poisson will provide a reasonable model. For a more complete treatment of the Poisson postulates, see the classic text by Feller (1968) or Barr and Zehna (1983).

*Theorem:* For each  $t \geq 0$ , let  $N_t$  be an integer-valued random variable with the following properties. (Think of  $N_t$  as denoting the number of arrivals in the time period from time 0 to time  $t$ .)

1.  $N_0 = 0$  (start with no arrivals)
2.  $s < t \Rightarrow N_s$  and  $N_t - N_s$  are independent. (arrivals in disjoint time periods are independent)
3.  $N_s$  and  $N_{t+s} - N_t$  are identically distributed. (number of arrivals depends only on period length)
4.  $\lim_{t \rightarrow 0} \frac{P(N_t = 1)}{t} = \lambda$  (arrival probability proportional to period length, if length is small)
5.  $\lim_{t \rightarrow 0} \frac{P(N_t > 1)}{t} = 0$  (no simultaneous arrivals)

If 1 – 5 hold, then for any integer  $n$ ,

$$P(N_t = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!},$$

that is,  $N_t \sim \text{Poisson}(\lambda t)$ . □

The postulates may also be interpreted as describing the behavior of objects spatially (for example, movement of insects), giving the Poisson application in spatial distributions.

# 4 Multiple Random Variables

*"Ah! my dear Watson, there we come into those realms  
of conjecture, where the most logical mind may be at fault."*  
**Sherlock Holmes**  
*The Adventure of the Empty House*

## 4.1 Joint and Marginal Distributions

In previous chapters, we have discussed probability models and computation of probability for events involving only one random variable. These are called *univariate models*. In this chapter we discuss probability models that involve more than one random variable—naturally enough, called *multivariate models*.

In an experimental situation, it would be very unusual to observe only the value of one random variable. That is, it would be an unusual experiment in which the total data collected consisted of just one numeric value. For example, consider an experiment designed to gain information about some health characteristics of a population of people. It would be a modest experiment indeed if the only datum collected was the body weight of one person. Rather, the body weights of several people in the population might be measured. These different weights would be observations on different random variables, one for each person measured. Multiple observations could also arise because several physical characteristics were measured on each person. For example, temperature, height, and blood pressure, in addition to weight, might be measured. These observations on different characteristics could also be modeled as observations on different random variables. Thus, we need to know how to describe and use probability models that deal with more than one random variable at a time. For the first several sections we will mainly discuss *bivariate models*, models involving two random variables.

Recall that, in Definition 1.4.1, a (univariate) random variable was defined to be a function from a sample space  $S$  into the real numbers. A random vector, consisting of several random variables, is defined similarly.

**DEFINITION 4.1.1:** An *n-dimensional random vector* is a function from a sample space  $S$  into  $\mathbb{R}^n$ ,  $n$ -dimensional Euclidean space.

Suppose, for example, that with each point in a sample space we associate an ordered pair of numbers, that is, a point  $(x, y) \in \mathbb{R}^2$ , where  $\mathbb{R}^2$  denotes the plane. Then we have defined a two-dimensional (or bivariate) random vector  $(X, Y)$ . Example 4.1.1 illustrates this.

**Example 4.1.1:** Consider the experiment of tossing two fair dice. The sample space for this experiment has 36 equally likely points and was introduced in Example 1.3.5. For example, the sample point  $(3, 3)$  denotes the outcome in which both dice show a 3; the sample point  $(4, 1)$  denotes the outcome in which the first die shows a 4 and the second die a 1; etc. Now, with each of these 36 points associate two numbers,  $X$  and  $Y$ . Let

$$X = \text{sum of the two dice} \quad \text{and} \quad Y = |\text{difference of the two dice}|.$$

For the sample point  $(3, 3)$ ,  $X = 3 + 3 = 6$  and  $Y = |3 - 3| = 0$ . For  $(4, 1)$ ,  $X = 5$  and  $Y = 3$ . These are also the values of  $X$  and  $Y$  for the sample point  $(1, 4)$ . For each of the 36 sample points we could compute the values of  $X$  and  $Y$ . In this way we have defined the bivariate random vector  $(X, Y)$ .

Having defined a random vector  $(X, Y)$ , we can now discuss probabilities of events that are defined in terms of  $(X, Y)$ . The probabilities of events defined in terms of  $X$  and  $Y$  are just defined in terms of the probabilities of the corresponding events in the sample space  $S$ . What is  $P(X = 5 \text{ and } Y = 3)$ ? You can verify that the only two sample points that yield  $X = 5$  and  $Y = 3$  are  $(4, 1)$  and  $(1, 4)$ . Thus the event “ $X = 5$  and  $Y = 3$ ” will occur if and only if the event  $\{(4, 1), (1, 4)\}$  occurs. Since each of the 36 sample points in  $S$  is equally likely,

$$P(\{(4, 1), (1, 4)\}) = \frac{2}{36} = \frac{1}{18}.$$

Thus,

$$P(X = 5 \text{ and } Y = 3) = \frac{1}{18}.$$

Henceforth, we will write  $P(X = 5, Y = 3)$  for  $P(X = 5 \text{ and } Y = 3)$ . Read the comma as “and.” Similarly,  $P(X = 6, Y = 0) = \frac{1}{36}$  because the only sample point that yields these values of  $X$  and  $Y$  is  $(3, 3)$ . For more complicated events, the technique is the same. For example,  $P(X = 7, Y \leq 4) = \frac{4}{36} = \frac{1}{9}$  because the only four sample points that yield  $X = 7$  and  $Y \leq 4$  are  $(4, 3), (3, 4), (5, 2)$ , and  $(2, 5)$ . ||

The random vector  $(X, Y)$  defined above is called a *discrete random vector* because it has only a countable (in this case, finite) number of possible values. For a discrete random vector, the function  $f(x, y)$  defined by  $f(x, y) = P(X = x, Y = y)$  can be used to compute any probabilities of events defined in terms of  $(X, Y)$ .

**DEFINITION 4.1.2:** Let  $(X, Y)$  be a discrete bivariate random vector. Then the function  $f(x, y)$  from  $\mathbb{R}^2$  into  $\mathbb{R}$  defined by  $f(x, y) = P(X = x, Y = y)$  is called the *joint probability mass function* or *joint pmf* of  $(X, Y)$ . If it is necessary to stress the fact that  $f$  is the joint pmf of the vector  $(X, Y)$  rather than some other vector, the notation  $f_{X,Y}(x, y)$  will be used.

**TABLE 4.1.1** Values of the joint pmf  $f(x, y)$ 

		$x$										
		2	3	4	5	6	7	8	9	10	11	12
$y$	0	$\frac{1}{36}$		$\frac{1}{36}$								
	1		$\frac{1}{18}$									
	2			$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$		
	3				$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$			
	4					$\frac{1}{18}$		$\frac{1}{18}$				
	5						$\frac{1}{18}$					

The joint pmf of  $(X, Y)$  completely defines the probability distribution of the random vector  $(X, Y)$ , just as the pmf of a discrete univariate random variable completely defines its distribution. For the  $(X, Y)$  defined in Example 4.1.1 in terms of the roll of a pair of dice, there are 21 possible values of  $(X, Y)$ . The value of  $f(x, y)$  for each of these 21 possible values is given in Table 4.1.1. Two of these values,  $f(5, 3) = \frac{1}{18}$  and  $f(6, 0) = \frac{1}{36}$ , were computed above and the rest are obtained by similar reasoning. The joint pmf  $f(x, y)$  is defined for all  $(x, y) \in \mathbb{R}^2$ , not just the 21 pairs in Table 4.1.1. For any other  $(x, y)$ ,  $f(x, y) = P(X = x, Y = y) = 0$ .

The joint pmf can be used to compute the probability of any event defined in terms of  $(X, Y)$ . Let  $A$  be any subset of  $\mathbb{R}^2$ . Then

$$P((X, Y) \in A) = \sum_{(x, y) \in A} f(x, y).$$

Since  $(X, Y)$  is discrete,  $f(x, y)$  is nonzero for at most a countable number of points  $(x, y)$ . Thus, the sum can be interpreted as a countable sum even if  $A$  contains an uncountable number of points. For example, let  $A = \{(x, y) : x = 7 \text{ and } y \leq 4\}$ . This is a half-infinite line in  $\mathbb{R}^2$ . But from Table 4.1.1 we see that the only  $(x, y) \in A$  for which  $f(x, y)$  is nonzero are  $(x, y) = (7, 1)$  and  $(x, y) = (7, 3)$ . Thus,

$$P(X = 7, Y \leq 4) = P((X, Y) \in A) = f(7, 1) + f(7, 3) = \frac{1}{18} + \frac{1}{18} = \frac{1}{9}.$$

This, of course, is the same value computed in Example 4.1.1 by considering the definition of  $(X, Y)$  and sample points in  $S$ . It is usually simpler to work with the joint pmf than it is to work with the fundamental definition.

Expectations of functions of random vectors are computed just as with univariate random variables. Let  $g(x, y)$  be a real-valued function defined for all possible values  $(x, y)$  of the discrete random vector  $(X, Y)$ . Then  $g(X, Y)$  is itself a random variable and its expected value  $Eg(X, Y)$  is given by

$$Eg(X, Y) = \sum_{(x, y) \in \mathbb{R}^2} g(x, y)f(x, y).$$

**Example 4.1.1 (Continued):** For the  $(X, Y)$  whose joint pmf is given in Table 4.1.1, what is the average value of  $XY$ ? Letting  $g(x, y) = xy$ , we compute  $\text{E}XY = \text{E}g(X, Y)$  by computing  $xyf(x, y)$  for each of the 21  $(x, y)$  points in Table 4.1.1 and summing these 21 terms. Thus,

$$\text{E}XY = (2)(0)\frac{1}{36} + (4)(0)\frac{1}{36} + \cdots + (8)(4)\frac{1}{18} + (7)(5)\frac{1}{18} = 13\frac{11}{18}. \quad ||$$

The expectation operator continues to have the properties listed in Theorem 2.2.1 when the random variable  $X$  is replaced by the random vector  $(X, Y)$ . For example, if  $g_1(x, y)$  and  $g_2(x, y)$  are two functions and  $a$ ,  $b$ , and  $c$  are constants, then

$$\text{E}(ag_1(X, Y) + bg_2(X, Y) + c) = a\text{E}g_1(X, Y) + b\text{E}g_2(X, Y) + c.$$

These properties follow from the properties of sums exactly as in the univariate case (see Exercise 4.2).

The joint pmf for any discrete bivariate random vector  $(X, Y)$  must have certain properties. For any  $(x, y)$ ,  $f(x, y) \geq 0$  since  $f(x, y)$  is a probability. Also, since  $(X, Y)$  is certain to be in  $\mathbb{R}^2$ ,

$$\sum_{(x,y) \in \mathbb{R}^2} f(x, y) = P((X, Y) \in \mathbb{R}^2) = 1.$$

It turns out that any nonnegative function from  $\mathbb{R}^2$  into  $\mathbb{R}$  that is nonzero for at most a countable number of  $(x, y)$  pairs and sums to 1 is the joint pmf for some bivariate discrete random vector  $(X, Y)$ . Thus, by defining  $f(x, y)$ , we can define a probability model for  $(X, Y)$  without ever working with the fundamental sample space  $S$ .

**Example 4.1.2:** Define  $f(x, y)$  by

$$f(0, 0) = f(0, 1) = \frac{1}{6},$$

$$f(1, 0) = f(1, 1) = \frac{1}{3},$$

$$f(x, y) = 0 \quad \text{for any other } (x, y).$$

Then  $f(x, y)$  is nonnegative and sums to 1 so  $f(x, y)$  is the joint pmf for some bivariate random vector  $(X, Y)$ . We can use  $f(x, y)$  to compute probabilities such as  $P(X = Y) = f(0, 0) + f(1, 1) = \frac{1}{2}$ . All this can be done without reference to the sample space  $S$ . Indeed, there are many sample spaces and functions thereon that lead to this joint pmf for  $(X, Y)$ . Here is one. Let  $S$  be the 36-point sample space for the experiment of tossing two fair dice. Let  $X = 0$  if the first die shows at most 2 and  $X = 1$  if the first die shows more than 2. Let  $Y = 0$  if the second die shows an odd number and  $Y = 1$  if the second die shows an even number. It is left as Exercise 4.3 to show that this definition leads to the above probability distribution for  $(X, Y)$ . ||

Even if we are considering a probability model for a random vector  $(X, Y)$ , there may be probabilities or expectations of interest that involve only one of the random variables in the vector. We may wish to know  $P(X = 2)$ , for instance. The variable  $X$  is itself a random variable, in the sense of Chapter 1, and its probability distribution is described by its pmf, namely,  $f_X(x) = P(X = x)$ . (As mentioned earlier, we now use the subscript to distinguish  $f_X(x)$  from the joint pmf  $f_{X,Y}(x, y)$ .) We now call  $f_X(x)$  the *marginal pmf of X* to emphasize the fact that it is the pmf of  $X$  but in the context of the probability model that gives the joint distribution of the vector  $(X, Y)$ . The marginal pmf of  $X$  or  $Y$  is easily calculated from the joint pmf of  $(X, Y)$  as Theorem 4.1.1 indicates.

**THEOREM 4.1.1:** Let  $(X, Y)$  be a discrete bivariate random vector with joint pmf  $f_{X,Y}(x, y)$ . Then the marginal pmfs of  $X$  and  $Y$ ,  $f_X(x) = P(X = x)$  and  $f_Y(y) = P(Y = y)$ , are given by

$$f_X(x) = \sum_{y \in \mathbb{R}} f_{X,Y}(x, y) \quad \text{and} \quad f_Y(y) = \sum_{x \in \mathbb{R}} f_{X,Y}(x, y).$$

*Proof:* We will prove the result for  $f_X(x)$ . The proof for  $f_Y(y)$  is similar. For any  $x \in \mathbb{R}$ , let  $A_x = \{(x, y) : -\infty < y < \infty\}$ . That is,  $A_x$  is the line in the plane with first coordinate equal to  $x$ . Then, for any  $x \in \mathbb{R}$ ,

$$\begin{aligned} f_X(x) &= P(X = x) \\ &= P(X = x, -\infty < Y < \infty) \quad - (P(-\infty < Y < \infty) = 1) \\ &= P((X, Y) \in A_x) \quad (\text{definition of } A_x) \\ &= \sum_{(x,y) \in A_x} f_{X,Y}(x, y) \\ &= \sum_{y \in \mathbb{R}} f_{X,Y}(x, y). \end{aligned} \quad \square$$

**Example 4.1.3:** Using the result of Theorem 4.1.1, we can compute the marginal distributions for  $X$  and  $Y$  from the joint distribution given in Table 4.1.1. To compute the marginal pmf of  $Y$ , for each possible value of  $Y$  we sum over the possible values of  $X$ . In this way we obtain

$$\begin{aligned} f_Y(0) &= f_{X,Y}(2, 0) + f_{X,Y}(4, 0) + f_{X,Y}(6, 0) \\ &\quad + f_{X,Y}(8, 0) + f_{X,Y}(10, 0) + f_{X,Y}(12, 0) \\ &= \frac{1}{6}. \end{aligned}$$

Similarly, we obtain

$$f_Y(1) = \frac{5}{18}, \quad f_Y(2) = \frac{2}{9}, \quad f_Y(3) = \frac{1}{6}, \quad f_Y(4) = \frac{1}{9}, \quad f_Y(5) = \frac{1}{18}.$$

Notice that  $f_Y(0) + f_Y(1) + f_Y(2) + f_Y(3) + f_Y(4) + f_Y(5) = 1$ , as it must, since these are the only six possible values of  $Y$ . ||

The marginal pmf of  $X$  or  $Y$  is the same as the pmf of  $X$  or  $Y$  defined in Chapter 1. The marginal pmf of  $X$  or  $Y$  can be used to compute probabilities or expectations that involve only  $X$  or  $Y$ . But to compute a probability or expectation that simultaneously involves both  $X$  and  $Y$ , we must use the joint pmf of  $X$  and  $Y$ .

**Example 4.1.4:** Using the marginal pmf of  $Y$  computed in Example 4.1.3, we can compute

$$P(Y < 3) = f_Y(0) + f_Y(1) + f_Y(2) = \frac{1}{6} + \frac{5}{18} + \frac{2}{9} = \frac{2}{3}.$$

Also,

$$EY^3 = 0^3 f_Y(0) + \dots + 5^3 f_Y(5) = 20\frac{11}{18}. \quad ||$$

The marginal distributions of  $X$  and  $Y$ , described by the marginal pmfs  $f_X(x)$  and  $f_Y(y)$ , do not completely describe the joint distribution of  $X$  and  $Y$ . Indeed, there are many different joint distributions that have the same marginal distributions. Thus, it is hopeless to try to determine the joint pmf,  $f_{X,Y}(x,y)$ , from knowledge of only the marginal pmfs,  $f_X(x)$  and  $f_Y(y)$ . The next example illustrates the point.

**Example 4.1.5:** Define a joint pmf by

$$\begin{aligned} f(0,0) &= \frac{1}{12}, & f(1,0) &= \frac{5}{12}, & f(0,1) &= f(1,1) = \frac{3}{12}, \\ f(x,y) &= 0 \quad \text{for all other values.} \end{aligned}$$

The marginal pmf of  $Y$  is  $f_Y(0) = f(0,0) + f(1,0) = \frac{1}{2}$  and  $f_Y(1) = f(0,1) + f(1,1) = \frac{1}{2}$ . The marginal pmf of  $X$  is  $f_X(0) = \frac{1}{3}$  and  $f_X(1) = \frac{2}{3}$ . Now check that for the joint pmf given in Example 4.1.2, which is obviously different from the one given here, the marginal pmfs of both  $X$  and  $Y$  are exactly the same as the ones just computed. Thus, we cannot determine what the joint pmf is if we know only the marginal pmfs. The joint pmf tells us additional information about the distribution of  $(X, Y)$  that is not found in the marginal distributions. ||

To this point we have discussed discrete bivariate random vectors. We can also consider random vectors whose components are continuous random variables. The probability distribution of a continuous random vector is usually described using a density function, as in the univariate case.

**DEFINITION 4.1.3:** A function  $f(x,y)$  from  $\mathbb{R}^2$  into  $\mathbb{R}$  is called a *joint probability density function* or *joint pdf of the continuous bivariate random vector  $(X, Y)$*  if, for every  $A \subset \mathbb{R}^2$ ,

$$P((X, Y) \in A) = \iint_A f(x, y) dx dy.$$

A joint pdf is used just like a univariate pdf except now the integrals are double integrals over sets in the plane. The notation  $\int \int_A$  simply means that the limits of integration are set so that the function is integrated over all  $(x, y) \in A$ . Expectations of functions of continuous random vectors are defined as in the discrete case with integrals replacing sums and the pdf replacing the pmf. That is, if  $g(x, y)$  is a real-valued function, then the *expected value* of  $g(X, Y)$  is defined to be

$$(4.1.1) \quad Eg(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy.$$

It is important to realize that the joint pdf *is* defined for all  $(x, y) \in \mathbb{R}^2$ . The pdf may equal 0 on a large set  $A$  if  $P((X, Y) \in A) = 0$  but the pdf *is* defined for the points in  $A$ .

The *marginal probability density functions* of  $X$  and  $Y$  are also defined as in the discrete case with integrals replacing sums. The marginal pdfs may be used to compute probabilities or expectations that involve only  $X$  or  $Y$ . Specifically, the marginal pdfs of  $X$  and  $Y$  are defined by

$$(4.1.2) \quad \begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x, y) dy, \quad -\infty < x < \infty, \\ f_Y(y) &= \int_{-\infty}^{\infty} f(x, y) dx, \quad -\infty < y < \infty. \end{aligned}$$

Any function  $f(x, y)$  satisfying  $f(x, y) \geq 0$  for all  $(x, y) \in \mathbb{R}^2$  and

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy$$

is the joint pdf of some continuous bivariate random vector  $(X, Y)$ . All of these concepts regarding joint pdfs are illustrated in the following two examples.

**Example 4.1.6:** Define a joint pdf by

$$f(x, y) = \begin{cases} 6xy^2 & \text{if } 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

(Henceforth, it will be understood that  $f(x, y) = 0$  for  $(x, y)$  values not specifically mentioned in the definition.) First, we might check that  $f(x, y)$  is indeed a joint pdf. That  $f(x, y) \geq 0$  for all  $(x, y)$  in the defined range is fairly obvious. To compute the integral of  $f(x, y)$  over the whole plane, note that, since  $f(x, y)$  is 0 except on the unit square, the integral over the plane is the same as the integral over the square. Thus we have

$$\begin{aligned}\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy &= \int_0^1 \int_0^1 6xy^2 dx dy = \int_0^1 3x^2y^2 \Big|_0^1 dy \\ &= \int_0^1 3y^2 dy = y^3 \Big|_0^1 = 1.\end{aligned}$$

Now, consider calculating a probability such as  $P(X + Y \geq 1)$ . Letting  $A = \{(x, y) : x + y \geq 1\}$ , we can re-express this as  $P((X, Y) \in A)$ . From Definition 4.1.3, to calculate the probability we integrate the joint pdf over the set  $A$ . But the joint pdf is 0 except on the unit square. So integrating over  $A$  is the same as integrating over only that part of  $A$  which is in the unit square. The set  $A$  is a half-plane in the northeast part of the plane and the part of  $A$  in the unit square is the triangular region bounded by the lines  $x = 1$ ,  $y = 1$ , and  $x + y = 1$ . We can write

$$\begin{aligned}A &= \{(x, y) : x + y \geq 1, 0 < x < 1, 0 < y < 1\} \\ &= \{(x, y) : x \geq 1 - y, 0 < x < 1, 0 < y < 1\} \\ &= \{(x, y) : 1 - y \leq x < 1, 0 < y < 1\}.\end{aligned}$$

This gives us the limits of integration we need to calculate the probability. We have

$$P(X + Y \geq 1) = \iint_A f(x, y) dx dy = \int_0^1 \int_{1-y}^1 6xy^2 dx dy = \frac{9}{10}.$$

Using (4.1.2), we can calculate the marginal pdf of  $X$  or  $Y$ . For example, to calculate  $f_X(x)$ , we note that for  $x \geq 1$  or  $x \leq 0$ ,  $f(x, y) = 0$  for all values of  $y$ . Thus for  $x \geq 1$  or  $x \leq 0$ ,

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = 0.$$

For  $0 < x < 1$ ,  $f(x, y)$  is nonzero only if  $0 < y < 1$ . Thus for  $0 < x < 1$ ,

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_0^1 6xy^2 dy = 2xy^3 \Big|_0^1 = 2x.$$

This marginal pdf of  $X$  can now be used to calculate probabilities involving only  $X$ . For example,

$$P\left(\frac{1}{2} < X < \frac{3}{4}\right) = \int_{\frac{1}{2}}^{\frac{3}{4}} 2x dx = \frac{5}{16}. \quad ||$$

**Example 4.1.7:** As another example of a joint pdf, let  $f(x, y) = e^{-y}$ ,  $0 < x < y < \infty$ . Although  $e^{-y}$  does not depend on  $x$ ,  $f(x, y)$  certainly is a function of  $x$  since the set where  $f(x, y)$  is nonzero depends on  $x$ . This is made more obvious by using an indicator function to write

$$f(x, y) = e^{-y} I_{\{(u,v):0 < u < v < \infty\}}(x, y).$$

To calculate  $P(X + Y \geq 1)$ , we could integrate the joint pdf over the region that is the intersection of the set  $A = \{(x, y) : x + y \geq 1\}$  and the set where  $f(x, y)$  is nonzero. Graph these sets and notice that this region is an unbounded region (lighter shading in Figure 4.1.1) with three sides given by the lines  $x = y$ ,  $x + y = 1$ , and  $x = 0$ . To integrate over this region we would have to break the region into at least two parts in order to write the appropriate limits of integration.

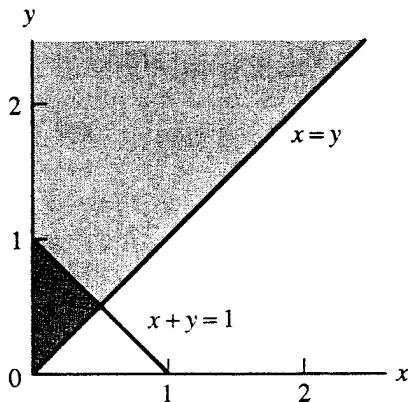


FIGURE 4.1.1 Regions for Example 4.1.7

The integration is easier over the intersection of the set  $B = \{(x, y) : x + y < 1\}$  and the set where  $f(x, y)$  is nonzero, the triangular region (darker shading in Figure 4.1.1) bounded by the lines  $x = y$ ,  $x + y = 1$ , and  $x = 0$ . Thus

$$\begin{aligned} P(X + Y \geq 1) &= 1 - P(X + Y < 1) = 1 - \int_0^{\frac{1}{2}} \int_x^{1-x} e^{-y} dy dx \\ &= 1 - \int_0^{\frac{1}{2}} (e^{-x} - e^{-(1-x)}) dx = 2e^{-1/2} - e^{-1}. \end{aligned}$$

This illustrates that it is almost always helpful to graph the sets of interest in determining the appropriate limits of integration for problems such as this. ||

The joint probability distribution of  $(X, Y)$  can be completely described with the *joint cdf* (cumulative distribution function) rather than with the joint pmf or joint pdf. The joint cdf is the function  $F(x, y)$  defined by

$$F(x, y) = P(X \leq x, Y \leq y)$$

for all  $(x, y) \in \mathbb{R}^2$ . The joint cdf is usually not very handy to use for a discrete random vector. But for a continuous bivariate random vector we have the important relationship, as in the univariate case,

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(s, t) dt ds.$$

From the bivariate Fundamental Theorem of Calculus, this implies that

$$(4.1.3) \quad \frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y),$$

at continuity points of  $f(x, y)$ . This relationship is useful in situations where an expression for  $F(x, y)$  can be found. The mixed partial derivative can be computed to find the joint pdf.

## 4.2 Conditional Distributions and Independence

Oftentimes when two random variables,  $(X, Y)$ , are observed, the values of the two variables are related. For example, suppose that, in sampling from a human population,  $X$  denotes a person's height and  $Y$  denotes the same person's weight. Surely we would think it more likely that  $Y > 200$  pounds if we were told that  $X = 73$  inches than if we were told that  $X = 41$  inches. Knowledge about the value of  $X$  gives us some information about the value of  $Y$  even if it does not tell us the value of  $Y$  exactly. Conditional probabilities regarding  $Y$  given knowledge of the  $X$  value can be computed using the joint distribution of  $(X, Y)$ . Sometimes, however, knowledge about  $X$  gives us no information about  $Y$ . We will discuss these topics concerning conditional probabilities in this section.

If  $(X, Y)$  is a discrete random vector, then a conditional probability of the form  $P(Y = y|X = x)$  is interpreted exactly as in Definition 1.3.1. For a countable (maybe finite) number of  $x$  values,  $P(X = x) > 0$ . For these values of  $x$ ,  $P(Y = y|X = x)$  is simply  $P(X = x, Y = y)/P(X = x)$ , according to the definition. The event  $\{Y = y\}$  is the event  $A$  in the formula and the event  $\{X = x\}$  is the event  $B$ . For a fixed value of  $x$ ,  $P(Y = y|X = x)$  could be computed for all possible values of  $y$ . In this way the probability of various values of  $y$  could be assessed given the knowledge that  $X = x$  was observed. This computation can be simplified by noting that in terms of the joint and marginal pmfs of  $X$  and  $Y$ , the above probabilities are  $P(X = x, Y = y) = f(x, y)$  and  $P(X = x) = f_X(x)$ . This leads to the following definition.

**DEFINITION 4.2.1:** Let  $(X, Y)$  be a discrete bivariate random vector with joint pmf  $f(x, y)$  and marginal pmfs  $f_X(x)$  and  $f_Y(y)$ . For any  $x$  such that  $P(X = x) = f_X(x) > 0$ , the *conditional pmf of  $Y$  given that  $X = x$*  is the function of  $y$  denoted by  $f(y|x)$  and defined by

$$f(y|x) = P(Y = y|X = x) = \frac{f(x, y)}{f_X(x)}.$$

For any  $y$  such that  $P(Y = y) = f_Y(y) > 0$ , the *conditional pmf of  $X$  given that  $Y = y$*  is the function of  $x$  denoted by  $f(x|y)$  and defined by

$$f(x|y) = P(X = x|Y = y) = \frac{f(x, y)}{f_Y(y)}.$$

Since we have called  $f(y|x)$  a pmf, we should verify that this function of  $y$  does indeed define a pmf for a random variable. First,  $f(y|x) \geq 0$  for every  $y$  since  $f(x,y) \geq 0$  and  $f_X(x) > 0$ . Second,

$$\sum_y f(y|x) = \frac{\sum_y f(x,y)}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1.$$

Thus,  $f(y|x)$  is indeed a pmf and can be used in the usual way to compute probabilities involving  $Y$  given the knowledge that  $X = x$  occurred.

**Example 4.2.1:** Define the joint pmf of  $(X, Y)$  by

$$\begin{aligned} f(0, 10) &= f(0, 20) = \frac{2}{18}, & f(1, 10) &= f(1, 30) = \frac{3}{18}, \\ f(1, 20) &= \frac{4}{18}, \quad \text{and} & f(2, 30) &= \frac{4}{18} \end{aligned}$$

We can use Definition 4.2.1 to compute the conditional pmf of  $Y$  given  $X$  for each of the possible values of  $X$ ,  $x = 0, 1, 2$ . First, the marginal pmf of  $X$  is

$$\begin{aligned} f_X(0) &= f(0, 10) + f(0, 20) = \frac{4}{18} \\ f_X(1) &= f(1, 10) + f(1, 20) + f(1, 30) = \frac{10}{18} \\ f_X(2) &= f(2, 30) = \frac{4}{18}. \end{aligned}$$

For  $x = 0$ ,  $f(0, y)$  is positive only for  $y = 10$  and  $y = 20$ . Thus  $f(y|0)$  is positive only for  $y = 10$  and  $y = 20$ , and

$$f(10|0) = \frac{f(0, 10)}{f_X(0)} = \frac{\frac{2}{18}}{\frac{4}{18}} = \frac{1}{2}$$

and

$$f(20|0) = \frac{f(0, 20)}{f_X(0)} = \frac{1}{2}.$$

That is, given the knowledge that  $X = 0$ , the conditional probability distribution for  $Y$  is the discrete distribution that assigns probability  $\frac{1}{2}$  to each of the two points  $y = 10$  and  $y = 20$ .

For  $x = 1$ ,  $f(y|1)$  is positive for  $y = 10, 20$ , and  $30$ , and

$$f(10|1) = f(30|1) = \frac{\frac{3}{18}}{\frac{10}{18}} = \frac{3}{10}$$

$$f(20|1) = \frac{\frac{4}{18}}{\frac{10}{18}} = \frac{4}{10};$$

and for  $x = 2$ ,

$$f(30|2) = \frac{\frac{4}{18}}{\frac{4}{18}} = 1.$$

The latter result reflects a fact that is also apparent from the joint pmf. If we know that  $X = 2$ , then we know that  $Y$  must be 30.

Other conditional probabilities can be computed using these conditional pmfs. For example,

$$P(Y > 10|X = 1) = f(20|1) + f(30|1) = \frac{7}{10}$$

or  $P(Y > 10|X = 0) = f(20|0) = \frac{1}{2}.$  ||

If  $X$  and  $Y$  are continuous random variables, then  $P(X = x) = 0$  for every value of  $x$ . To compute a conditional probability such as  $P(Y > 200|X = 73)$ , Definition 1.3.1 cannot be used since the denominator,  $P(X = 73)$ , is zero. Yet in actuality, a specific value such as  $X = 73$  is observed and, as the height and weight example at the beginning of this section indicated, the knowledge that  $X = 73$  might give us information about  $Y$ . It turns out that the appropriate way to define a conditional probability distribution for  $Y$  given  $X = x$ , when  $X$  and  $Y$  are both continuous, is analogous to the discrete case with pdfs replacing pmfs.

**DEFINITION 4.2.2:** Let  $(X, Y)$  be a continuous bivariate random vector with joint pdf  $f(x, y)$  and marginal pdfs  $f_X(x)$  and  $f_Y(y)$ . For any  $x$  such that  $f_X(x) > 0$ , the *conditional pdf of  $Y$  given that  $X = x$*  is the function of  $y$  denoted by  $f(y|x)$  and defined by

$$f(y|x) = \frac{f(x, y)}{f_X(x)}.$$

For any  $y$  such that  $f_Y(y) > 0$ , the *conditional pdf of  $X$  given that  $Y = y$*  is the function of  $x$  denoted by  $f(x|y)$  and defined by

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

To verify that  $f(x|y)$  and  $f(y|x)$  are indeed pdfs, the same steps can be used as in the earlier verification that Definition 4.2.1 had defined true pmfs with integrals now replacing sums.

In addition to their usefulness for calculating probabilities, the conditional pdfs or pmfs can also be used to calculate expected values. Just remember that  $f(y|x)$  as a function of  $y$  is a pdf or pmf and use it in the same way that we have previously used unconditional pdfs or pmfs. If  $g(Y)$  is a function of  $Y$ , then the *conditional expected value of  $g(Y)$  given that  $X = x$*  is denoted by  $E(g(Y)|x)$  and is given by

$$E(g(Y)|x) = \sum_y g(y)f(y|x) \quad \text{and} \quad E(g(Y)|x) = \int_{-\infty}^{\infty} g(y)f(y|x) dy$$

in the discrete and continuous cases, respectively. The conditional expected value has all of the properties of the usual expected value listed in Theorem 2.2.1. Moreover,  $E(Y|X)$  provides the best guess at  $Y$  based on knowledge of  $X$ , extending the result in Example 2.2.4. (See Exercise 4.13.)

**Example 4.2.2:** As in Example 4.1.7, let the continuous random vector  $(X, Y)$  have joint pdf  $f(x, y) = e^{-y}$ ,  $0 < x < y < \infty$ . Suppose we wish to compute the conditional pdf of  $Y$  given  $X = x$ . The marginal pdf of  $X$  is computed as follows. If  $x \leq 0$ ,  $f(x, y) = 0$  for all values of  $y$  so  $f_X(x) = 0$ . If  $x > 0$ ,  $f(x, y) > 0$  only if  $y > x$ . Thus

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_x^{\infty} e^{-y} dy = e^{-x}.$$

Thus, marginally,  $X$  has an exponential distribution. Using Definition 4.2.2, the conditional distribution of  $Y$  given  $X = x$  can be computed for any  $x \geq 0$  (since these are the values for which  $f_X(x) > 0$ ). For any such  $x$ ,

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{e^{-y}}{e^{-x}} = e^{-(y-x)}, \quad \text{if } y > x$$

and

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{0}{e^{-x}} = 0, \quad \text{if } y \leq x.$$

Thus, given  $X = x$ ,  $Y$  has an exponential distribution where  $x$  is the location parameter in the distribution of  $Y$  and  $\beta = 1$  is the scale parameter. The conditional distribution of  $Y$  is different for every value of  $x$ . It then follows that

$$E(Y|X = x) = \int_x^{\infty} ye^{-(y-x)} dy = 1 + x.$$

The variance of the probability distribution described by  $f(y|x)$  is called the *conditional variance of  $Y$  given  $X = x$* . Using the notation  $\text{Var}(Y|x)$  for this, we have, using the ordinary definition of variance,

$$\text{Var}(Y|x) = E(Y^2|x) - (E(Y|x))^2.$$

Applying this definition to our example, we obtain

$$\text{Var}(Y|x) = \int_x^{\infty} y^2 e^{-(y-x)} dy - \left( \int_x^{\infty} ye^{-(y-x)} dy \right)^2 = 1.$$

In this case the conditional variance of  $Y$  given  $X = x$  is the same for all values of  $x$ . In other situations, however, it may be different for different values of  $x$ . This conditional variance might be compared to the unconditional variance of  $Y$ .

The marginal distribution of  $Y$  is gamma(2, 1), which has  $\text{Var } Y = 2$ . Given the knowledge that  $X = x$ , the variability in  $Y$  is considerably reduced. ||

A physical situation for which the model in Example 4.2.2 might be used is this. Suppose we have two light bulbs. The lengths of time each will burn are random variables denoted by  $X$  and  $Z$ . The lifelengths  $X$  and  $Z$  are independent and both have pdf  $e^{-x}$ ,  $x > 0$ . The first bulb will be turned on. As soon as it burns out, the second bulb will be turned on. Now consider observing  $X$ , the time when the first bulb burns out, and  $Y = X + Z$ , the time when the second bulb burns out. Given that  $X = x$  is when the first burned out and the second is started,  $Y = Z + x$ . This is like Example 3.4.1. The value  $x$  is acting as a location parameter and the pdf of  $Y$ , in this case the *conditional* pdf of  $Y$  given  $X = x$ , is  $f(y|x) = f_Z(y-x) = e^{-(y-x)}$ ,  $y > x$ .

The conditional distribution of  $Y$  given  $X = x$  is possibly a different probability distribution for each value of  $x$ . Thus we really have a family of probability distributions for  $Y$ , one for each  $x$ . When we wish to describe this entire family, we will use the phrase “the distribution of  $Y|X$ .” If, for example,  $X$  is a positive integer-valued random variable and the conditional distribution of  $Y$  given  $X = x$  is binomial( $x, p$ ), then we might say the distribution of  $Y|X$  is binomial( $X, p$ ) or write  $Y|X \sim \text{binomial}(X, p)$ . Whenever we use the symbol  $Y|X$  or have a random variable as the parameter of a probability distribution, we are describing the family of conditional probability distributions. Joint pdfs or pmfs are sometimes defined by specifying the conditional  $f(y|x)$  and the marginal  $f_X(x)$ . Then the definition yields  $f(x, y) = f(y|x)f_X(x)$ . These types of models are discussed more in Section 4.4.

Notice also that  $E(g(Y)|x)$  is a function of  $x$ . That is, for each value of  $x$ ,  $E(g(Y)|x)$  is a real number obtained by computing the appropriate integral or sum. Thus,  $E(g(Y)|X)$  is a random variable whose value depends on the value of  $X$ . If  $X = x$ , the value of the random variable  $E(g(Y)|X)$  is  $E(g(Y)|x)$ . Thus, in Example 4.2.2, we can write  $E(Y|X) = 1 + X$ .

In all the previous examples, the conditional distribution of  $Y$  given  $X = x$  was different for different values of  $x$ . In some situations, the knowledge that  $X = x$  does not give us any more information about  $Y$  than what we already had. This important relationship between  $X$  and  $Y$  is called *independence*. Just as with independent events in Chapter 1, it is more convenient to define independence in a symmetric fashion and then derive conditional properties like those we just mentioned. This we now do.

**DEFINITION 4.2.3:** Let  $(X, Y)$  be a bivariate random vector with joint pdf or pmf  $f(x, y)$ , and marginal pdfs or pmfs  $f_X(x)$  and  $f_Y(y)$ . Then  $X$  and  $Y$  are called *independent random variables* if, for every  $x \in \mathfrak{R}$  and  $y \in \mathfrak{R}$ ,

$$(4.2.1) \quad f(x, y) = f_X(x)f_Y(y).$$

If  $X$  and  $Y$  are independent, the conditional pdf of  $Y$  given  $X = x$  is

$$f(y|x) = \frac{f(x, y)}{f_X(x)} \quad (\text{definition})$$

$$\begin{aligned}
 &= \frac{f_X(x)f_Y(y)}{f_X(x)} \quad (\text{from (4.2.1)}) \\
 &= f_Y(y),
 \end{aligned}$$

regardless of the value of  $x$ . Thus, for any  $A \subset \Re$  and  $x \in \Re$ ,  $P(Y \in A|x) = \int_A f(y|x)dy = \int_A f_Y(y)dy = P(Y \in A)$ . The knowledge that  $X = x$  gives us no additional information about  $Y$ .

Definition 4.2.3 is used in two different ways. We might start with a joint pdf or pmf and then check whether  $X$  and  $Y$  are independent. To do this we must verify that (4.2.1) is true for every value of  $x$  and  $y$ . Or we might wish to define a model in which  $X$  and  $Y$  are independent. Consideration of what  $X$  and  $Y$  represent might indicate that knowledge that  $X = x$  should give us no information about  $Y$ . In this case we could specify the marginal distributions of  $X$  and  $Y$  and then define the joint distribution as the product as given in (4.2.1).

**Example 4.2.3:** Consider the discrete bivariate random vector  $(X, Y)$ , with joint pmf given by

$$\begin{aligned}
 f(10, 1) &= f(20, 1) = f(20, 2) = \frac{1}{10}, \\
 f(10, 2) &= f(10, 3) = \frac{1}{5}, \quad \text{and} \quad f(20, 3) = \frac{3}{10}.
 \end{aligned}$$

The marginal pmfs are easily calculated to be

$$f_X(10) = f_X(20) = \frac{1}{2} \quad \text{and} \quad f_Y(1) = \frac{1}{5}, \quad f_Y(2) = \frac{3}{10}, \quad \text{and} \quad f_Y(3) = \frac{1}{2}.$$

The random variables  $X$  and  $Y$  are not independent because (4.2.1) is not true for every  $x$  and  $y$ . For example,

$$f(10, 3) = \frac{1}{5} \neq \frac{1}{2} \cdot \frac{1}{2} = f_X(10)f_Y(3).$$

The relationship (4.2.1) must hold for every choice of  $x$  and  $y$  if  $X$  and  $Y$  are to be independent. Note that  $f(10, 1) = \frac{1}{10} = \frac{1}{2} \cdot \frac{1}{5} = f_X(10)f_Y(1)$ . That (4.2.1) holds for some values of  $x$  and  $y$  does not ensure that  $X$  and  $Y$  are independent. All values must be checked. ||

The verification that  $X$  and  $Y$  are independent by direct use of (4.2.1) would require the knowledge of  $f_X(x)$  and  $f_Y(y)$ . The following lemma makes the verification somewhat easier.

**LEMMA 4.2.1:** Let  $(X, Y)$  be a bivariate random vector with joint pdf or pmf  $f(x, y)$ . Then  $X$  and  $Y$  are independent random variables if and only if there exist functions  $g(x)$  and  $h(y)$  such that, for every  $x \in \Re$  and  $y \in \Re$ ,

$$f(x, y) = g(x)h(y).$$

*Proof:* The “only if” part is proved by defining  $g(x) = f_X(x)$  and  $h(y) = f_Y(y)$  and using (4.2.1). To prove the “if” part for continuous random variables, suppose that  $f(x, y) = g(x)h(y)$ . Define

$$\int_{-\infty}^{\infty} g(x) dx = c \quad \text{and} \quad \int_{-\infty}^{\infty} h(y) dy = d,$$

where the constants  $c$  and  $d$  satisfy

$$\begin{aligned}
 cd &= \left( \int_{-\infty}^{\infty} g(x) dx \right) \left( \int_{-\infty}^{\infty} h(y) dy \right) \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y) dx dy \\
 (4.2.2) \quad &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy \\
 &= 1. \quad (f(x, y) \text{ is a joint pdf})
 \end{aligned}$$

Furthermore, the marginal pdfs are given by

$$\begin{aligned}
 (4.2.3) \quad f_X(x) &= \int_{-\infty}^{\infty} g(x)h(y) dy = g(x)d \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} g(x)h(y) dx = h(y)c.
 \end{aligned}$$

Thus, using (4.2.2) and (4.2.3), we have

$$f(x, y) = g(x)h(y) = g(x)h(y)cd = f_X(x)f_Y(y),$$

showing that  $X$  and  $Y$  are independent. Replacing integrals with sums proves the lemma for discrete random vectors.  $\square$

**Example 4.2.4:** Consider the joint pdf  $f(x, y) = \frac{1}{384}x^2y^4e^{-y-(x/2)}$ ,  $x > 0$  and  $y > 0$ . If we define

$$g(x) = \begin{cases} x^2e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases} \quad \text{and} \quad h(y) = \begin{cases} y^4e^{-y}/384 & y > 0 \\ 0 & y \leq 0 \end{cases},$$

then  $f(x, y) = g(x)h(y)$ , for all  $x \in \mathbb{R}$  and all  $y \in \mathbb{R}$ . By Lemma 4.2.1, we conclude that  $X$  and  $Y$  are independent random variables. We do not have to compute marginal pdfs.  $\parallel$

If  $X$  and  $Y$  are independent random variables, then from (4.2.1) it is clear that  $f(x, y) > 0$  on the set  $\{(x, y) : x \in A \text{ and } y \in B\}$  where  $A = \{x : f_X(x) > 0\}$  and  $B = \{y : f_Y(y) > 0\}$ . A set of this form is called a cross-product and is usually denoted by  $A \times B$ . Membership in a cross-product can be checked by considering the  $x$  and  $y$  values separately. If  $f(x, y)$  is a joint pdf or pmf and the set where

$f(x, y) > 0$  is not a cross-product, then the random variables  $X$  and  $Y$  with joint pdf or pmf  $f(x, y)$  are not independent. In Example 4.2.2, the set  $0 < x < y < \infty$  is not a cross-product. To check membership in this set we must not only check that  $0 < x < \infty$  and  $0 < y < \infty$  but also  $x < y$ . Thus the random variables in Example 4.2.2 are not independent. Example 4.2.1 gives an example of a joint pmf which is positive on a set which is not a cross-product.

**Example 4.2.5:** As an example of using independence to define a joint probability model, consider this situation. A student from an elementary school in Kansas City is randomly selected and  $X =$  the number of living parents of the student is recorded. Suppose the marginal distribution of  $X$  is

$$f_X(0) = .01, \quad f_X(1) = .09, \quad \text{and} \quad f_X(2) = .90.$$

A retiree from Sun City is randomly selected and  $Y =$  the number of living parents of the retiree is recorded. Suppose the marginal distribution of  $Y$  is

$$f_Y(0) = .70, \quad f_Y(1) = .25, \quad \text{and} \quad f_Y(2) = .05.$$

It seems reasonable to assume that these two random variables are independent. Knowledge of the number of parents of the student tells us nothing about the number of parents of the retiree. The only joint distribution of  $X$  and  $Y$  that reflects this independence is the one defined by (4.2.1). Thus, for example,

$$f(0, 0) = f_X(0)f_Y(0) = .0070 \quad \text{and} \quad f(0, 1) = f_X(0)f_Y(1) = .0025.$$

This joint distribution can be used to calculate quantities such as

$$\begin{aligned} P(X = Y) &= f(0, 0) + f(1, 1) + f(2, 2) \\ &= (.01)(.70) + (.09)(.25) + (.90)(.05) = .0745. \end{aligned}$$

Certain probabilities and expectations are easy to calculate if  $X$  and  $Y$  are independent, as the next theorem indicates.

**THEOREM 4.2.1:** Let  $X$  and  $Y$  be independent random variables.

- a. For any  $A \subset \mathbb{R}$  and  $B \subset \mathbb{R}$ ,  $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ , that is, the events  $\{X \in A\}$  and  $\{Y \in B\}$  are independent events.
- b. Let  $g(x)$  be a function only of  $x$  and  $h(y)$  be a function only of  $y$ . Then

$$E(g(X)h(Y)) = (Eg(X))(Eh(Y)).$$

*Proof:* For continuous random variables, part (b) is proved by noting that

$$\begin{aligned} E(g(X)h(Y)) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y) dx dy \quad (\text{by (4.2.1)}) \end{aligned}$$

$$\begin{aligned}
&= \int_{-\infty}^{\infty} h(y)f_Y(y) \int_{-\infty}^{\infty} g(x)f_X(x) dx dy \\
&= \left( \int_{-\infty}^{\infty} g(x)f_X(x) dx \right) \left( \int_{-\infty}^{\infty} h(y)f_Y(y) dy \right) \\
&= (\mathbb{E}g(X))(\mathbb{E}h(Y)).
\end{aligned}$$

The result for discrete random variables is proved by replacing integrals by sums. Part (a) can be proved by a series of steps similar to those above or by the following argument. Let  $g(x)$  be the indicator function of the set  $A$ . Let  $h(y)$  be the indicator function of the set  $B$ . Note that  $g(x)h(y)$  is the indicator function of the set  $C \subset \mathbb{R}^2$  defined by  $C = \{(x, y) : x \in A, y \in B\}$ . Also note that for an indicator function such as  $g(x)$ ,  $\mathbb{E}g(X) = P(X \in A)$ . Thus using the expectation equality just proved, we have

$$\begin{aligned}
P(X \in A, Y \in B) &= P((X, Y) \in C) = \mathbb{E}(g(X)h(Y)) \\
&= (\mathbb{E}g(X))(\mathbb{E}h(Y)) = P(X \in A)P(Y \in B).
\end{aligned}$$
□

**Example 4.2.6:** Let  $X$  and  $Y$  be independent exponential(1) random variables. From Theorem 4.2.1 we have

$$P(X \geq 4, Y < 3) = P(X \geq 4)P(Y < 3) = e^{-4}(1 - e^{-3}).$$

Letting  $g(x) = x^2$  and  $h(y) = y$ , we see that

$$\mathbb{E}(X^2Y) = (\mathbb{E}X^2)(\mathbb{E}Y) = (\text{Var } X + (\mathbb{E}X)^2)\mathbb{E}Y = (1 + 1^2)1 = 2.$$
||

The following result concerning sums of independent random variables is a simple consequence of Theorem 4.2.1.

**THEOREM 4.2.2:** Let  $X$  and  $Y$  be independent random variables with moment generating functions  $M_X(t)$  and  $M_Y(t)$ . Then the moment generating function of the random variable  $Z = X + Y$  is given by

$$M_Z(t) = M_X(t)M_Y(t).$$

*Proof:* Using the definition of the mgf and Theorem 4.2.1, we have

$$M_Z(t) = \mathbb{E}e^{tZ} = \mathbb{E}e^{t(X+Y)} = \mathbb{E}(e^{tX}e^{tY}) = (\mathbb{E}e^{tX})(\mathbb{E}e^{tY}) = M_X(t)M_Y(t).$$
□

**Example 4.2.7:** Sometimes Theorem 4.2.2 can be used to easily derive the distribution of  $Z$  from knowledge of the distribution of  $X$  and  $Y$ . For example, let  $X \sim n(\mu, \sigma^2)$  and  $Y \sim n(\gamma, \tau^2)$  be independent normal random variables. From Exercise 2.37, the mgfs of  $X$  and  $Y$  are

$$M_X(t) = \exp(\mu t + \sigma^2 t^2/2) \quad \text{and} \quad M_Y(t) = \exp(\gamma t + \tau^2 t^2/2).$$

Thus, from Theorem 4.2.2, the mgf of  $Z = X + Y$  is

$$M_Z(t) = M_X(t)M_Y(t) = \exp((\mu + \gamma)t + (\sigma^2 + \tau^2)t^2/2).$$

This is the mgf of a normal random variable with mean  $\mu + \gamma$  and variance  $\sigma^2 + \tau^2$ . This result is important enough to be stated as a theorem. ||

**THEOREM 4.2.3:** Let  $X \sim n(\mu, \sigma^2)$  and  $Y \sim n(\gamma, \tau^2)$  be independent normal random variables. Then the random variable  $Z = X + Y$  has a  $n(\mu + \gamma, \sigma^2 + \tau^2)$  distribution. □

If  $f(x, y)$  is the joint pdf for the continuous random vector  $(X, Y)$ , (4.2.1) may fail to hold on a set  $A$  of  $(x, y)$  values for which  $\int_A \int dx dy = 0$ . In such a case  $X$  and  $Y$  are still called independent random variables. This reflects the fact that two pdfs that differ only on a set such as  $A$  define the same probability distribution for  $(X, Y)$ . To see this, suppose  $f(x, y)$  and  $f^*(x, y)$  are two pdfs that are equal everywhere except on a set  $A$  for which  $\int_A \int dx dy = 0$ . Let  $(X, Y)$  have pdf  $f(x, y)$ , let  $(X^*, Y^*)$  have pdf  $f^*(x, y)$ , and let  $B$  be any subset of  $\mathbb{R}^2$ . Then

$$\begin{aligned} P((X, Y) \in B) &= \iint_B f(x, y) dx dy = \iint_{B \cap A^c} f(x, y) dx dy = \iint_{B \cap A^c} f^*(x, y) dx dy \\ &= \iint_B f^*(x, y) dx dy = P((X^*, Y^*) \in B). \end{aligned}$$

Thus  $(X, Y)$  and  $(X^*, Y^*)$  have the same probability distribution. So, for example,  $f(x, y) = e^{-x-y}$ ,  $x > 0$  and  $y > 0$ , is a pdf for two independent exponential random variables and satisfies (4.2.1). But,  $f^*(x, y)$ , which is equal to  $f(x, y)$  except that  $f^*(x, y) = 0$  if  $x = y$ , is also the pdf for two independent exponential random variables even though (4.2.1) is not true on the set  $A = \{(x, x) : x > 0\}$ .

### 4.3 Bivariate Transformations

In Section 2.1, methods of finding the distribution of a function of a random variable were discussed. In this section we extend these ideas to the case of bivariate random vectors.

Let  $(X, Y)$  be a bivariate random vector with a known probability distribution. Now consider a new bivariate random vector  $(U, V)$  defined by  $U = g_1(X, Y)$  and  $V = g_2(X, Y)$  where  $g_1(x, y)$  and  $g_2(x, y)$  are some specified functions. If  $B$  is any subset of  $\mathbb{R}^2$ , then  $(U, V) \in B$  if and only if  $(X, Y) \in A$ , where  $A = \{(x, y) : (g_1(x, y), g_2(x, y)) \in B\}$ . Thus  $P((U, V) \in B) = P((X, Y) \in A)$  and the probability distribution of  $(U, V)$  is completely determined by the probability distribution of  $(X, Y)$ .

If  $(X, Y)$  is a discrete bivariate random vector, then there is only a countable set of values for which the joint pmf of  $(X, Y)$  is positive. Call this set  $\mathcal{A}$ . Define the set  $\mathcal{B} = \{(u, v) : u = g_1(x, y) \text{ and } v = g_2(x, y) \text{ for some } (x, y) \in \mathcal{A}\}$ . Then  $\mathcal{B}$  is the countable set of possible values for the discrete random vector  $(U, V)$ . And if, for

any  $(u, v) \in \mathcal{B}$ ,  $A_{uv}$  is defined to be  $\{(x, y) \in \mathcal{A} : g_1(x, y) = u \text{ and } g_2(x, y) = v\}$  then the joint pmf of  $(U, V)$ ,  $f_{U,V}(u, v)$ , can be computed from the joint pmf of  $(X, Y)$  by

$$(4.3.1) \quad f_{U,V}(u, v) = P(U = u, V = v) = P((X, Y) \in A_{uv}) = \sum_{(x,y) \in A_{uv}} f_{X,Y}(x, y).$$

**Example 4.3.1:** Let  $X$  and  $Y$  be independent Poisson random variables with parameters  $\theta$  and  $\lambda$ , respectively. Thus the joint pmf of  $(X, Y)$  is

$$f_{X,Y}(x, y) = \frac{\theta^x e^{-\theta}}{x!} \frac{\lambda^y e^{-\lambda}}{y!}, \quad x = 0, 1, 2, \dots, y = 0, 1, 2, \dots$$

The set  $\mathcal{A}$  is  $\{(x, y) : x = 0, 1, 2, \dots \text{ and } y = 0, 1, 2, \dots\}$ . Now define  $U = X + Y$  and  $V = Y$ . That is,  $g_1(x, y) = x + y$  and  $g_2(x, y) = y$ . We will describe the set  $\mathcal{B}$ , the set of possible  $(u, v)$  values. The possible values for  $v$  are the nonnegative integers. The variable  $v = y$  and thus has the same set of possible values. For a given value of  $v$ ,  $u = x + y = x + v$  must be an integer greater than or equal to  $v$  since  $x$  is a nonnegative integer. The set of all possible  $(u, v)$  values is thus given by  $\mathcal{B} = \{(u, v) : v = 0, 1, 2, \dots \text{ and } u = v, v+1, v+2, \dots\}$ . For any  $(u, v) \in \mathcal{B}$ , the only  $(x, y)$  value satisfying  $x + y = u$  and  $y = v$  is  $x = u - v$  and  $y = v$ . Thus, in this example,  $A_{uv}$  always consists of only the single point  $(u - v, v)$ . From (4.3.1) we thus obtain the joint pmf of  $(U, V)$  as

$$f_{U,V}(u, v) = f_{X,Y}(u - v, v) = \frac{\theta^{u-v} e^{-\theta}}{(u-v)!} \frac{\lambda^v e^{-\lambda}}{v!}, \quad v = 0, 1, 2, \dots, \quad u = v, v+1, v+2, \dots$$

In this example it is interesting to compute the marginal pmf of  $U$ . For any fixed nonnegative integer  $u$ ,  $f_{U,V}(u, v) > 0$  only for  $v = 0, 1, \dots, u$ . This gives the set of  $v$  values to sum over to obtain the marginal pmf of  $U$ . It is

$$f_U(u) = \sum_{v=0}^u \frac{\theta^{u-v} e^{-\theta}}{(u-v)!} \frac{\lambda^v e^{-\lambda}}{v!} = e^{-(\theta+\lambda)} \sum_{v=0}^u \frac{\theta^{u-v}}{(u-v)!} \frac{\lambda^v}{v!}, \quad u = 0, 1, 2, \dots$$

This can be simplified by noting that, if we multiply and divide each term by  $u!$ , we can use the Binomial Theorem to obtain

$$f_U(u) = \frac{e^{-(\theta+\lambda)}}{u!} \sum_{v=0}^u \binom{u}{v} \lambda^v \theta^{u-v} = \frac{e^{-(\theta+\lambda)}}{u!} (\theta + \lambda)^u, \quad u = 0, 1, 2, \dots$$

This is the pmf of a Poisson random variable with parameter  $\theta + \lambda$ . This result is significant enough to be stated as a theorem. ||

**THEOREM 4.3.1:** If  $X \sim \text{Poisson}(\theta)$  and  $Y \sim \text{Poisson}(\lambda)$  and  $X$  and  $Y$  are independent, then  $X + Y \sim \text{Poisson}(\theta + \lambda)$ . □

If  $(X, Y)$  is a continuous random vector with joint pdf  $f_{X,Y}(x, y)$ , then the joint pdf of  $(U, V)$  can be expressed in terms of  $f_{X,Y}(x, y)$  in a manner analogous to (2.1.8). As before,  $\mathcal{A} = \{(x, y) : f_{X,Y}(x, y) > 0\}$  and  $\mathcal{B} = \{(u, v) : u = g_1(x, y) \text{ and } v = g_2(x, y) \text{ for some } (x, y) \in \mathcal{A}\}$ . The joint pdf  $f_{U,V}(u, v)$  will be positive on the set  $\mathcal{B}$ . For the simplest version of this result we assume that the transformation  $u = g_1(x, y)$  and  $v = g_2(x, y)$  defines a one-to-one transformation of  $\mathcal{A}$  onto  $\mathcal{B}$ . The transformation is onto because of the definition of  $\mathcal{B}$ . We are assuming that for each  $(u, v) \in \mathcal{B}$  there is only one  $(x, y) \in \mathcal{A}$  such that  $(u, v) = (g_1(x, y), g_2(x, y))$ . For such a one-to-one, onto transformation, we can solve the equations  $u = g_1(x, y)$  and  $v = g_2(x, y)$  for  $x$  and  $y$  in terms of  $u$  and  $v$ . We will denote this inverse transformation by  $x = h_1(u, v)$  and  $y = h_2(u, v)$ . The role played by a derivative in the univariate case is now played by a quantity called the *Jacobian of the transformation*. This function of  $(u, v)$ , denoted by  $J$ , is the *determinant of a matrix of partial derivatives*. It is defined by

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial y}{\partial u} \frac{\partial x}{\partial v},$$

where

$$\frac{\partial x}{\partial u} = \frac{\partial h_1(u, v)}{\partial u}, \quad \frac{\partial x}{\partial v} = \frac{\partial h_1(u, v)}{\partial v}, \quad \frac{\partial y}{\partial u} = \frac{\partial h_2(u, v)}{\partial u}, \quad \text{and} \quad \frac{\partial y}{\partial v} = \frac{\partial h_2(u, v)}{\partial v}.$$

We assume that  $J$  is not identically zero on  $\mathcal{B}$ . Then the joint pdf of  $(U, V)$  is 0 outside the set  $\mathcal{B}$  and on the set  $\mathcal{B}$  is given by

$$(4.3.2) \quad f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v))|J|,$$

where  $|J|$  is the absolute value of  $J$ . When using (4.3.2), it is sometimes just as difficult to determine the set  $\mathcal{B}$  and verify that the transformation is one-to-one as it is to substitute into formula (4.3.2). Note these parts of the explanations in the following examples.

**Example 4.3.2:** Let  $X \sim \text{beta}(\alpha, \beta)$  and  $Y \sim \text{beta}(\alpha + \beta, \gamma)$  be independent random variables. The joint pdf of  $(X, Y)$  is

$$f_{X,Y}(x, y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \frac{\Gamma(\alpha + \beta + \gamma)}{\Gamma(\alpha + \beta)\Gamma(\gamma)} y^{\alpha+\beta-1} (1-y)^{\gamma-1},$$

$$0 < x < 1, \quad 0 < y < 1.$$

Consider the transformation  $U = XY$  and  $V = X$ . The set of possible values for  $V$  is  $0 < v < 1$  since  $V = X$ . For a fixed value of  $V = v$ ,  $U$  must be between 0 and  $v$  since  $X = V = v$  and  $Y$  is between 0 and 1. Thus, this transformation maps the set  $\mathcal{A}$  onto the set  $\mathcal{B} = \{(u, v) : 0 < u < v < 1\}$ . For any  $(u, v) \in \mathcal{B}$ , the equations  $u = xy$  and  $v = x$  can be uniquely solved for  $x = h_1(u, v) = v$  and  $y = h_2(u, v) = u/v$ . Note that if considered as a transformation defined on all of  $\mathbb{R}^2$ ,

this transformation is not one-to-one. Any point  $(0, y)$  is mapped into the point  $(0, 0)$ . But as a function defined only on  $\mathcal{A}$ , it is a one-to-one transformation onto  $\mathcal{B}$ . The Jacobian is given by

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ \frac{1}{v} & -\frac{u}{v^2} \end{vmatrix} = -\frac{1}{v}.$$

Thus, from (4.3.2) we obtain the joint pdf as

$$(4.3.3) \quad f_{U,V}(u, v) = \frac{\Gamma(\alpha + \beta + \gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} v^{\alpha-1} (1-v)^{\beta-1} \left(\frac{u}{v}\right)^{\alpha+\beta-1} \left(1 - \frac{u}{v}\right)^{\gamma-1} \frac{1}{v},$$

$$0 < u < v < 1.$$

The marginal distribution of  $V = X$  is, of course, a beta( $\alpha, \beta$ ) distribution. But the distribution of  $U$  is also a beta distribution:

$$\begin{aligned} f_U(u) &= \int_u^1 f_{U,V}(u, v) dv \\ &= \frac{\Gamma(\alpha + \beta + \gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} u^{\alpha-1} \int_u^1 \left(\frac{u}{v} - u\right)^{\beta-1} \left(1 - \frac{u}{v}\right)^{\gamma-1} \left(\frac{u}{v^2}\right) dv. \end{aligned}$$

The expression (4.3.3) was used but some terms have been rearranged. Now make the univariate change of variable  $y = (u/v - u)/(1 - u)$  so that  $dy = -u/[v^2(1 - u)]dv$  to obtain

$$\begin{aligned} f_U(u) &= \frac{\Gamma(\alpha + \beta + \gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} u^{\alpha-1} (1-u)^{\beta+\gamma-1} \int_0^1 y^{\beta-1} (1-y)^{\gamma-1} dy \\ &= \frac{\Gamma(\alpha + \beta + \gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} u^{\alpha-1} (1-u)^{\beta+\gamma-1} \frac{\Gamma(\beta)\Gamma(\gamma)}{\Gamma(\beta + \gamma)} \\ &= \frac{\Gamma(\alpha + \beta + \gamma)}{\Gamma(\alpha)\Gamma(\beta + \gamma)} u^{\alpha-1} (1-u)^{\beta+\gamma-1}, \quad 0 < u < 1. \end{aligned}$$

To obtain the second identity we recognized the integrand as the kernel of a beta pdf and used (3.2.17). Thus we see that the marginal distribution of  $U$  is beta( $\alpha, \beta + \gamma$ ). ||

**Example 4.3.3:** Let  $X$  and  $Y$  be independent, standard normal random variables. Consider the transformation  $U = X + Y$  and  $V = X - Y$ . In the notation used above,  $U = g_1(X, Y)$  where  $g_1(x, y) = x + y$  and  $V = g_2(X, Y)$  where  $g_2(x, y) = x - y$ . The joint pdf of  $X$  and  $Y$  is, of course,  $f_{X,Y}(x, y) = (2\pi)^{-1} \exp(-x^2/2) \exp(-y^2/2)$ ,  $-\infty < x < \infty, -\infty < y < \infty$ . So the set  $\mathcal{A} = \mathbb{R}^2$ . To determine the set  $\mathcal{B}$  on which  $f_{U,V}(u, v)$  is positive, we must determine all the values that

$$(4.3.4) \quad u = x + y \quad \text{and} \quad v = x - y$$

take on as  $(x, y)$  range over the set  $\mathcal{A} = \mathbb{R}^2$ . But we can set  $u$  to be any number and  $v$  to be any number and uniquely solve equations (4.3.4) for  $x$  and  $y$  to obtain

$$(4.3.5) \quad x = h_1(u, v) = \frac{u + v}{2} \quad \text{and} \quad y = h_2(u, v) = \frac{u - v}{2}.$$

This shows two things. For any  $(u, v) \in \mathbb{R}^2$  there is an  $(x, y) \in \mathcal{A}$  (defined by (4.3.5)) such that  $u = x + y$  and  $v = x - y$ . So  $\mathcal{B}$ , the set of all possible  $(u, v)$  values, is  $\mathbb{R}^2$ . Since the solution (4.3.5) is unique, this also shows that the transformation we have considered is one-to-one. Only the  $(x, y)$  given in (4.3.5) will yield  $u = x + y$  and  $v = x - y$ . From (4.3.5) the partial derivatives of  $x$  and  $y$  are easy to compute. We obtain

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

Substituting the expressions (4.3.5) for  $x$  and  $y$  into  $f_{X,Y}(x, y)$  and using  $|J| = \frac{1}{2}$ , we obtain the joint pdf of  $(U, V)$  from (4.3.2) as

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v))|J| = \frac{1}{2\pi} e^{-((u+v)/2)^2/2} e^{-((u-v)/2)^2/2} \frac{1}{2}$$

for  $-\infty < u < \infty$  and  $-\infty < v < \infty$ . Multiplying out the squares in the exponentials, we see that the terms involving  $uv$  cancel. Thus after some simplification and rearrangement we obtain

$$f_{U,V}(u, v) = \left( \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-u^2/4} \right) \left( \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-v^2/4} \right).$$

The joint pdf has factored into a function of  $u$  and a function of  $v$ . By Lemma 4.2.1,  $U$  and  $V$  are independent. From Theorem 4.2.3, the marginal distribution of  $U = X + Y$  is  $n(0, 2)$ . Similarly, Theorem 4.2.2 could be used to find that the marginal distribution of  $V$  is also  $n(0, 2)$ . This important fact, that sums and differences of independent normal random variables are independent normal random variables, is true regardless of the means of  $X$  and  $Y$ , so long as  $\text{Var } X = \text{Var } Y$ . This result is left as Exercise 4.27. Theorems 4.2.2 and 4.2.3 give us the marginal distributions of  $U$  and  $V$ . But the more involved analysis here is required to determine that  $U$  and  $V$  are independent. ||

In Example 4.3.3, we found that  $U$  and  $V$  are independent random variables. There is a much simpler, but very important, situation in which new variables  $U$  and  $V$ , defined in terms of original variables  $X$  and  $Y$ , are independent. Theorem 4.3.2 describes this.

**THEOREM 4.3.2:** Let  $X$  and  $Y$  be independent random variables. Let  $g(x)$  be a function only of  $x$  and  $h(y)$  be a function only of  $y$ . Then the random variables  $U = g(X)$  and  $V = h(Y)$  are independent.

*Proof:* We will prove the theorem assuming  $U$  and  $V$  are continuous random variables. For any  $u \in \mathbb{R}$  and  $v \in \mathbb{R}$ , define

$$A_u = \{x : g(x) \leq u\} \quad \text{and} \quad B_v = \{y : h(y) \leq v\}.$$

Then the joint cdf of  $(U, V)$  is

$$\begin{aligned} F_{U,V}(u, v) &= P(U \leq u, V \leq v) && \text{(definition of cdf)} \\ &= P(X \in A_u, Y \in B_v) && \text{(definition of } U \text{ and } V) \\ &= P(X \in A_u)P(Y \in B_v). && \text{(Theorem 4.2.1)} \end{aligned}$$

The joint pdf of  $(U, V)$  is

$$\begin{aligned} f_{U,V}(u, v) &= \frac{\partial^2}{\partial u \partial v} F_{U,V}(u, v) && \text{(equation (4.1.3))} \\ &= \left( \frac{d}{du} P(X \in A_u) \right) \left( \frac{d}{dv} P(Y \in B_v) \right) \end{aligned}$$

where, as the notation indicates, the first factor is a function only of  $u$  and the second factor is a function only of  $v$ . Hence, by Lemma 4.2.1,  $U$  and  $V$  are independent.  $\square$

It may be that there is only one function, say  $U = g_1(X, Y)$ , of interest. In such cases, this method may still be used to find the distribution of  $U$ . If another convenient function,  $V = g_2(X, Y)$ , can be chosen so that the resulting transformation from  $(X, Y)$  to  $(U, V)$  is one-to-one on  $\mathcal{A}$ , then the joint pdf of  $(U, V)$  can be derived using (4.3.2) and the marginal pdf of  $U$  can be obtained from the joint pdf. In the previous example, perhaps we were interested only in  $U = XY$ . We could choose to define  $V = X$ , recognizing that the resulting transformation is one-to-one on  $\mathcal{A}$ . Then proceed as in the example to obtain the marginal pdf of  $U$ . But other choices, such as  $V = Y$ , would work as well (see Exercise 4.23).

Of course, in many situations, the transformation of interest is not one-to-one. Just as Theorem 2.1.3 generalized the univariate method to many-to-one functions, the same can be done here. As before,  $\mathcal{A} = \{(x, y) : f_{X,Y}(x, y) > 0\}$ . Suppose  $A_0, A_1, \dots, A_k$  form a partition of  $\mathcal{A}$  with these properties. The set  $A_0$ , which may be empty, satisfies  $P((X, Y) \in A_0) = 0$ . The transformation  $U = g_1(X, Y)$  and  $V = g_2(X, Y)$  is a one-to-one transformation from  $A_i$  onto  $\mathcal{B}$  for each  $i = 1, 2, \dots, k$ . Then for each  $i$ , the inverse functions from  $\mathcal{B}$  to  $A_i$  can be found. Denote the  $i$ th inverse by  $x = h_{1i}(u, v)$  and  $y = h_{2i}(u, v)$ . This  $i$ th inverse gives, for  $(u, v) \in \mathcal{B}$ , the unique  $(x, y) \in A_i$  such that  $(u, v) = (g_1(x, y), g_2(x, y))$ . Let  $J_i$  denote the Jacobian computed from the  $i$ th inverse. Then assuming that these Jacobians do not vanish identically on  $\mathcal{B}$ , we have the following representation of the joint pdf,  $f_{U,V}(u, v)$ :

$$(4.3.6) \quad f_{U,V}(u, v) = \sum_{i=1}^k f_{X,Y}(h_{1i}(u, v), h_{2i}(u, v)) |J_i|.$$

**Example 4.3.4:** Let  $X$  and  $Y$  be independent  $n(0, 1)$  random variables. Consider the transformation  $U = X/Y$  and  $V = |Y|$ . ( $U$  and  $V$  can be defined to be any value, say  $(1, 1)$ , if  $Y = 0$  since  $P(Y = 0) = 0$ .) This transformation is not one-to-one since the points  $(x, y)$  and  $(-x, -y)$  are both mapped into the same  $(u, v)$  point. But if we restrict consideration to either positive or negative values of  $y$ , then the transformation is one-to-one. In the above notation, let

$$A_1 = \{(x, y) : y > 0\}, \quad A_2 = \{(x, y) : y < 0\}, \quad \text{and} \quad A_0 = \{(x, y) : y = 0\}.$$

$A_0, A_1$ , and  $A_2$  form a partition of  $\mathcal{A} = \mathbb{R}^2$  and  $P((X, Y) \in A_0) = P(Y = 0) = 0$ . For either  $A_1$  or  $A_2$ , if  $(x, y) \in A_i$ ,  $v = |y| > 0$  and for a fixed value of  $v = |y|$ ,  $u = x/y$  can be any real number since  $x$  can be any real number. Thus,  $\mathcal{B} = \{(u, v) : v > 0\}$  is the image of both  $A_1$  and  $A_2$  under the transformation. Furthermore, the inverse transformations from  $\mathcal{B}$  to  $A_1$  and  $\mathcal{B}$  to  $A_2$  are given by  $x = h_{11}(u, v) = uv$ ,  $y = h_{21}(u, v) = v$ , and  $x = h_{12}(u, v) = -uv$ ,  $y = h_{22}(u, v) = -v$ . Note that the first inverse gives positive values of  $y$  and the second gives negative values of  $y$ . The Jacobians from the two inverses are  $J_1 = J_2 = v$ . Using

$$f_{X,Y}(x, y) = \frac{1}{2\pi} e^{-x^2/2} e^{-y^2/2},$$

from (4.3.6) we obtain

$$\begin{aligned} f_{U,V}(u, v) &= \frac{1}{2\pi} e^{-(uv)^2/2} e^{-v^2/2} |v| + \frac{1}{2\pi} e^{-(-uv)^2/2} e^{-(v)^2/2} |v| \\ &= \frac{v}{\pi} e^{-(u^2+1)v^2/2}, \quad -\infty < u < \infty, \quad 0 < v < \infty. \end{aligned}$$

From this the marginal pdf of  $U$  can be computed to be

$$\begin{aligned} f_U(u) &= \int_0^\infty \frac{v}{\pi} e^{-(u^2+1)v^2/2} dv \\ &= \frac{1}{2\pi} \int_0^\infty e^{-(u^2+1)z/2} dz \quad (\text{change of variable } z = v^2) \\ &= \frac{1}{2\pi} \frac{2}{(u^2+1)} \quad \left( \begin{array}{l} \text{integrand is kernel of} \\ \text{exponential } (\beta = 2/(u^2+1)) \text{ pdf} \end{array} \right) \\ &= \frac{1}{\pi(u^2+1)}, \quad -\infty < u < \infty. \end{aligned}$$

So we see that the ratio of two independent standard normal random variables is a Cauchy random variable. (See Exercise 4.28 for more relationships between normal and Cauchy random variables.) ||

## 4.4 Hierarchical Models and Mixture Distributions

In the cases we have seen thus far, a random variable has a single distribution, possibly depending on parameters. While, in general, a random variable can have only one distribution, it is often easier to model a situation by thinking of things in a hierarchy.

**Example 4.4.1:** Perhaps the most classic hierarchical model is the following. An insect lays a large number of eggs, each surviving with probability  $p$ . On the average, how many eggs will survive?

The “large number” of eggs laid is a random variable, often taken to be  $\text{Poisson}(\lambda)$ . Furthermore, if we assume that each egg’s survival is independent, then we have Bernoulli trials. Therefore, if we let  $X = \text{number of survivors}$  and  $Y = \text{number of eggs laid}$ , we have

$$X|Y \sim \text{binomial}(Y, p)$$

$$Y \sim \text{Poisson}(\lambda),$$

a hierarchical model. (Recall that we use notation such as  $X|Y \sim \text{binomial}(Y, p)$  to mean that the conditional distribution of  $X$  given  $Y = y$  is  $\text{binomial}(y, p)$ .) ||

The advantage of the hierarchy is that complicated processes may be modeled by a sequence of relatively simple models placed in a hierarchy. Also, dealing with the hierarchy is no more difficult than dealing with conditional and marginal distributions.

**Example 4.4.1 (Continued):** The random variable of interest,  $X = \text{number of survivors}$ , has the distribution given by

$$\begin{aligned} P(X = x) &= \sum_{y=0}^{\infty} P(X = x, Y = y) \\ &= \sum_{y=0}^{\infty} P(X = x|Y = y)P(Y = y) && \left( \begin{array}{l} \text{definition of} \\ \text{conditional probability} \end{array} \right) \\ &= \sum_{y=x}^{\infty} \left[ \binom{y}{x} p^x (1-p)^{y-x} \right] \left[ \frac{e^{-\lambda} \lambda^y}{y!} \right], && \left( \begin{array}{l} \text{conditional probability} \\ \text{is zero if } y < x \end{array} \right) \end{aligned}$$

since  $X|Y = y$  is  $\text{binomial}(y, p)$  and  $Y$  is  $\text{Poisson}(\lambda)$ . If we now simplify this last expression, cancelling what we can, and multiplying by  $\lambda^x/\lambda^x$ , we get

$$\begin{aligned} P(X = x) &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{y-x}}{(y-x)!} \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{((1-p)\lambda)^t}{t!} && (t = y - x) \end{aligned}$$

$$\begin{aligned}
 &= \frac{(\lambda p)^x e^{-\lambda}}{\lambda!} e^{(1-p)\lambda} \quad \left( \begin{array}{l} \text{sum is a kernel for} \\ \text{a Poisson distribution} \end{array} \right) \\
 &= \frac{(\lambda p)^x}{x!} e^{-\lambda p},
 \end{aligned}$$

so  $X \sim \text{Poisson}(\lambda p)$ . Thus, any marginal inference on  $X$  is with respect to a  $\text{Poisson}(\lambda p)$  distribution, with  $Y$  playing no part at all. Introducing  $Y$  in the hierarchy was mainly to aid our understanding of the model. There was an added bonus in that the parameter of the distribution of  $X$  is the product of two parameters, each relatively simple to understand.

The answer to the original question is now easy to compute:

$$EX = \lambda p,$$

so, on the average,  $\lambda p$  eggs will survive. If we were interested only in this mean, and did not need the distribution, we could have used properties of conditional expectations. ||

Sometimes, calculations can be greatly simplified by using the following theorem. Recall from Section 4.2 that  $E(X|y)$  is a function of  $y$  and  $E(X|Y)$  is a random variable whose value depends on the value of  $Y$ .

**THEOREM 4.4.1:** If  $X$  and  $Y$  are any two random variables, then

$$(4.4.1) \quad EX = E(E(X|Y)),$$

provided that the expectations exist.

*Proof:* Let  $f(x, y)$  denote the joint pdf of  $X$  and  $Y$ . By definition, we have

$$(4.4.2) \quad EX = \int \int x f(x, y) dx dy = \int \left[ \int x f(x|y) dx \right] f_Y(y) dy,$$

where  $f(x|y)$  and  $f_Y(y)$  are the conditional pdf of  $X$  given  $Y = y$  and the marginal pdf of  $Y$ , respectively. But now notice that the inner integral in (4.4.2) is the conditional expectation  $E(X|y)$ , and we have

$$EX = \int E(X|y) f_Y(y) dy = E(E(X|Y)),$$

as desired. Replace integrals by sums to prove the discrete case. □

Note that equation (4.4.1) contains an abuse of notation, since we have used the “E” to stand for different expectations in the same equation. The “E” in the left-hand side of (4.4.1) is expectation with respect to the marginal distribution of  $X$ . The first “E” in the right-hand side of (4.4.1) is expectation with respect to the marginal

distribution of  $Y$ , while the second one stands for expectation with respect to the conditional distribution of  $X|Y$ . However, there is really no cause for confusion because these interpretations are the only ones that the symbol “E” can take!

We can now easily compute the expected number of survivors in Example 4.4.1. From Theorem 4.4.1 we have

$$\begin{aligned} EX &= E(E(X|Y)) \\ &= E(pY) \quad (\text{since } X|Y \sim \text{binomial}(Y, p)) \\ &= p\lambda. \quad (\text{since } Y \sim \text{Poisson}(\lambda)) \end{aligned}$$

The term *mixture distribution* in the title of this section refers to a distribution arising from a hierarchical structure. Although there is no standardized definition for this term, we will use the following definition, which seems to be a popular one.

**DEFINITION 4.4.1:** A random variable  $X$  is said to have a *mixture distribution* if the distribution of  $X$  depends on a quantity which also has a distribution.

Thus, in Example 4.4.1 the  $\text{Poisson}(\lambda p)$  distribution is a mixture distribution since it is the result of combining a  $\text{binomial}(Y, p)$  with  $Y \sim \text{Poisson}(\lambda)$ . In general, we can say that hierarchical models lead to mixture distributions.

There is nothing to stop the hierarchy at two stages, but it should be easy to see that any more complicated hierarchy can be treated as a two-stage hierarchy theoretically. There may be advantages, however, in modeling a phenomenon as a multistage hierarchy. It may be easier to understand.

**Example 4.4.2:** Consider a generalization of Example 4.4.1, where instead of one mother insect there are a large number of mothers and one mother is chosen at random. We are still interested in knowing the average number of survivors, but it is no longer clear that the number of eggs laid follows the same Poisson distribution for each mother. The following three-stage hierarchy may be more appropriate. Let  $X$  = number of survivors in a litter; then

$$\begin{aligned} X|Y &\sim \text{binomial}(Y, p), \\ Y|\Lambda &\sim \text{Poisson}(\Lambda), \\ \Lambda &\sim \text{exponential}(\beta), \end{aligned}$$

where the last stage of the hierarchy accounts for the variability across different mothers.

The mean of  $X$  can easily be calculated as

$$\begin{aligned} EX &= E(E(X|Y)) \\ &= E(pY) \quad (\text{as before}) \\ &= E(E(pY|\Lambda)) \end{aligned}$$

$$\begin{aligned}
 &= E(p\Lambda) \\
 &= p\beta, \quad (\text{exponential expectation})
 \end{aligned}$$

completing the calculation. ||

In this example we have used a slightly different type of model than before in that two of the random variables are discrete and one is continuous. Using these models should present no problems. We can define a joint density,  $f(x, y, \lambda)$ ; conditional densities,  $f(x|y)$ ,  $f(x|y, \lambda)$ , etc.; and marginal densities,  $f(x)$ ,  $f(x, y)$ , etc.; as before. Simply understand that, when probabilities or expectations are calculated, discrete variables are summed and continuous variables are integrated.

Note that this three-stage model can also be thought of as a two-stage hierarchy by combining the last two stages. If  $Y|\Lambda \sim \text{Poisson}(\Lambda)$  and  $\Lambda \sim \text{exponential}(\beta)$ , then

$$\begin{aligned}
 P(Y = y) &= P(Y = y, 0 < \Lambda < \infty) \\
 &= \int_0^\infty f(y, \lambda) d\lambda \\
 &= \int_0^\infty f(y|\lambda)f(\lambda) d\lambda \\
 &= \int_0^\infty \left[ \frac{e^{-\lambda}\lambda^y}{y!} \right] \frac{1}{\beta} e^{-\lambda/\beta} d\lambda \\
 &= \frac{1}{\beta y!} \int_0^\infty \lambda^y e^{-\lambda(1+\beta^{-1})} d\lambda \quad \begin{pmatrix} \text{gamma} \\ \text{pdf kernel} \end{pmatrix} \\
 &= \frac{1}{\beta y!} \Gamma(y+1) \left( \frac{1}{1+\beta^{-1}} \right)^{y+1} \\
 &= \frac{1}{(1+\beta)} \left( \frac{1}{1+\beta^{-1}} \right)^y.
 \end{aligned}$$

This expression for the pmf of  $Y$  is the form (3.1.10) of the negative binomial pmf. Therefore, our three-stage hierarchy in Example 4.4.2 is equivalent to the two-stage hierarchy

$$X|Y \sim \text{binomial}(Y, p),$$

$$Y \sim \text{negative binomial}(p = \frac{1}{1+\beta}, r = 1).$$

However, in terms of understanding the model, the three-stage model is much easier to understand!

A useful generalization is a Poisson–gamma mixture, which is a generalization of a part of the previous model. If we have the hierarchy

$$Y|\Lambda \sim \text{Poisson}(\Lambda),$$

$$\Lambda \sim \text{gamma}(\alpha, \beta),$$

then the marginal distribution of  $Y$  is negative binomial (see Exercise 4.34). This model for the negative binomial distribution shows that it can be considered to be a “more variable” Poisson. Solomon (1983) explains these and other biological and mathematical models that lead to the negative binomial distribution. (See Exercise 4.35.)

Aside from the advantage in aiding understanding, hierarchical models can often make calculations easier. For example, a distribution that often occurs in statistics is the *noncentral chi squared distribution*. With  $p$  degrees of freedom and noncentrality parameter  $\lambda$ , the pdf is given by

$$(4.4.3) \quad f(x|\lambda, p) = \sum_{k=0}^{\infty} \frac{x^{p/2+k-1} e^{-x/2}}{\Gamma(p/2+k) 2^{p/2+k}} \frac{\lambda^k e^{-\lambda}}{k!},$$

an extremely messy expression. Calculating  $EX$ , for example, looks like quite a chore. However, if we examine the pdf closely, we see that this is a mixture distribution, made up of central chi squared densities (like those given in (3.2.10)) and Poisson distributions. That is, if we set up the hierarchy,

$$X|K \sim \chi^2_{p+2K},$$

$$K \sim \text{Poisson}(\lambda),$$

then the marginal distribution of  $X$  is given by (4.4.3). Hence

$$\begin{aligned} EX &= E(E(X|K)) \\ &= E(p + 2K) \\ &= p + 2\lambda, \end{aligned}$$

a relatively simple calculation.  $\text{Var } X$  can also be calculated in this way.

We close this section with one more hierarchical model, and illustrate one more conditional expectation calculation.

**Example 4.4.3:** One generalization of Bernoulli trials is to allow the success probability to vary from trial to trial, keeping the trials independent. A standard model for this situation is

$$X_i|P_i \sim \text{Bernoulli}(P_i), \quad i = 1, \dots, n,$$

$$P_i \sim \text{beta}(\alpha, \beta).$$

This model might be appropriate, for example, if we are measuring the success of a drug on  $n$  patients and, because the patients are different, we are reluctant to assume that the success probabilities are constant.

A random variable of interest is  $Y = \sum_{i=1}^n X_i$ , the total number of successes, and we might want to know its mean and variance. First,

$$\begin{aligned}\mathbb{E}Y &= \sum_{i=1}^n \mathbb{E}X_i \\ &= \sum_{i=1}^n \mathbb{E}(\mathbb{E}(X_i|P_i)) \\ &= \sum_{i=1}^n EP_i && (X_i|P_i \sim \text{Bernoulli}(P_i)) \\ &= \sum_{i=1}^n \frac{\alpha}{\alpha + \beta} && (P_i \sim \text{beta}(\alpha, \beta)) \\ &= \frac{n\alpha}{\alpha + \beta}. && \parallel\end{aligned}$$

Calculating the variance of  $Y$  is only slightly more involved. We can make use of a formula for conditional variances, similar in spirit to the expected value identity of Theorem 4.4.1.

**THEOREM 4.4.2:** For any two random variables  $X$  and  $Y$ ,

$$(4.4.4) \quad \text{Var } X = \mathbb{E}(\text{Var}(X|Y)) + \text{Var}(\mathbb{E}(X|Y)),$$

provided that the expectations exist.

*Proof:* By definition, we have

$$\text{Var } X = \mathbb{E}([X - EX]^2) = \mathbb{E}([X - \mathbb{E}(X|Y) + \mathbb{E}(X|Y) - EX]^2),$$

where in the last step we have added and subtracted  $\mathbb{E}(X|Y)$ . Expanding the square in this last expectation now gives

$$\begin{aligned}(4.4.5) \quad \text{Var } X &= \mathbb{E}([X - \mathbb{E}(X|Y)]^2) + \mathbb{E}([\mathbb{E}(X|Y) - EX]^2) \\ &\quad + 2\mathbb{E}([X - \mathbb{E}(X|Y)][\mathbb{E}(X|Y) - EX]).\end{aligned}$$

The last term in this expression is equal to 0, however, which can easily be seen by iterating the expectation:

$$(4.4.6) \quad \mathbb{E}([X - \mathbb{E}(X|Y)][\mathbb{E}(X|Y) - EX]) = \mathbb{E}(\mathbb{E}\{[X - \mathbb{E}(X|Y)][\mathbb{E}(X|Y) - EX]|Y\}).$$

In the conditional distribution  $X|Y$ ,  $X$  is the random variable. So in the expression

$$E\{[X - E(X|Y)][E(X|Y) - EX]|Y\},$$

$E(X|Y)$  and  $EX$  are constants. Thus,

$$\begin{aligned} E\{[X - E(X|Y)][E(X|Y) - EX]|Y\} &= (E(X|Y) - EX)(E\{[X - E(X|Y)]|Y\}) \\ &= (E(X|Y) - EX)(E(X|Y) - E(X|Y)) \\ &= (E(X|Y) - EX)(0) \\ &= 0. \end{aligned}$$

Thus, from (4.4.6), we have that  $E((X - E(X|Y))(E(X|Y) - EX)) = E(0) = 0$ . Referring back to equation (4.4.5), we see that

$$\begin{aligned} E([X - E(X|Y)]^2) &= E(E\{[X - E(X|Y)]^2|Y\}) \\ &= E(\text{Var}(X|Y)), \end{aligned}$$

and

$$E([E(X|Y) - EX]^2) = \text{Var}(E(X|Y)),$$

establishing (4.4.4).  $\square$

**Example 4.4.3 (Continued):** To calculate the variance of  $Y$ , we first note that

$$(4.4.7) \quad \text{Var } Y = \text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var } X_i,$$

since the  $X_i$ s are independent. Using (4.4.4),

$$\text{Var } X_i = \text{Var}(E(X_i|P_i)) + E(\text{Var}(X_i|P_i)).$$

Now  $E(X_i|P_i) = P_i$ , and since  $P_i \sim \text{beta}(\alpha, \beta)$ ,

$$\begin{aligned} \text{Var}(E(X_i|P_i)) &= \text{Var}(P_i) \\ &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \end{aligned}$$

Also, since  $X_i|P_i$  is Bernoulli( $P_i$ ),  $\text{Var}(X_i|P_i) = P_i(1 - P_i)$ . We then have

$$\begin{aligned} E(\text{Var}(X_i|P_i)) &= E(P_i(1 - P_i)) \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p_i(1 - p_i)p_i^{\alpha-1}(1 - p_i)^{\beta-1} dp_i. \end{aligned}$$

Notice that the integrand is the kernel of another beta pdf (with parameters  $\alpha + 1$  and  $\beta + 1$ ) so

$$\begin{aligned} E(\text{Var}(X_i|P_i)) &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \left[ \frac{\Gamma(\alpha + 1)\Gamma(\beta + 1)}{\Gamma(\alpha + \beta + 2)} \right] \\ &= \frac{\alpha\beta}{(\alpha + \beta)(\alpha + \beta + 1)}. \end{aligned}$$

Adding together the two pieces and simplifying, we get

$$\text{Var } X_i = \frac{\alpha\beta}{(\alpha + \beta)^2}.$$

Thus,  $\text{Var } X_i$  does not depend on  $i$  and, from (4.4.7),  $\text{Var } Y = n\alpha\beta/(\alpha + \beta)^2$ . ||

## 4.5 Covariance and Correlation

In earlier sections, we have discussed the absence or presence of a relationship between two random variables, independence or nonindependence. But if there is a relationship, the relationship may be strong or weak. In this section we discuss two numerical measures of the strength of a relationship between two random variables, the covariance and correlation.

To illustrate what we mean by the strength of a relationship between two random variables, consider two different experiments. In the first, random variables  $X$  and  $Y$  are measured where  $X$  is the weight of a sample of water and  $Y$  is the volume of the same sample of water. Clearly there is a strong relationship between  $X$  and  $Y$ . If  $(X, Y)$  pairs are measured on several samples and the observed data pairs are plotted, the data points should fall on a straight line because of the physical relationship between  $X$  and  $Y$ . This will not be exactly the case because of measurement errors, impurities in the water, etc. But with careful laboratory technique, the data points will fall very nearly on a straight line. Now consider another experiment in which  $X$  and  $Y$  are measured where  $X$  is the body weight of a human and  $Y$  is the same human's height. Clearly there is also a relationship between  $X$  and  $Y$  here but the relationship is not nearly as strong. We would not expect a plot of  $(X, Y)$  pairs measured on different people to form a straight line although we might expect to see an upward trend in the plot. The covariance and correlation are two measures that quantify this difference in the strength of a relationship between two random variables.

Throughout this section we will frequently be referring to the mean and variance of  $X$  and the mean and variance of  $Y$ . For these we will use the notation  $EX = \mu_X$ ,  $EY = \mu_Y$ ,  $\text{Var } X = \sigma_X^2$ , and  $\text{Var } Y = \sigma_Y^2$ . We will assume throughout that  $0 < \sigma_X^2 < \infty$  and  $0 < \sigma_Y^2 < \infty$ .

**DEFINITION 4.5.1:** The *covariance of  $X$  and  $Y$*  is the number defined by

$$\text{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y)).$$

**DEFINITION 4.5.2:** The *correlation of X and Y* is the number defined by

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

The value  $\rho_{XY}$  is also called the *correlation coefficient*.

If large values of  $X$  tend to be observed with large values of  $Y$  and small values of  $X$  with small values of  $Y$ , then  $\text{Cov}(X, Y)$  will be positive. If  $X > \mu_X$ , then  $Y > \mu_Y$  is likely to be true and the product  $(X - \mu_X)(Y - \mu_Y)$  will be positive. If  $X < \mu_X$ , then  $Y < \mu_Y$  is likely to be true and the product  $(X - \mu_X)(Y - \mu_Y)$  will again be positive. Thus  $\text{Cov}(X, Y) = E(X - \mu_X)(Y - \mu_Y) > 0$ . If large values of  $X$  tend to be observed with small values of  $Y$  and small values of  $X$  with large values of  $Y$ , then  $\text{Cov}(X, Y)$  will be negative because when  $X > \mu_X$ ,  $Y$  will tend to be less than  $\mu_Y$  and vice versa, and hence  $(X - \mu_X)(Y - \mu_Y)$  will tend to be negative. Thus the sign of  $\text{Cov}(X, Y)$  gives information regarding the relationship between  $X$  and  $Y$ .

But  $\text{Cov}(X, Y)$  can be any number and a given value of  $\text{Cov}(X, Y)$ , say  $\text{Cov}(X, Y) = 3$ , does not in itself give information about the strength of the relationship between  $X$  and  $Y$ . On the other hand, the correlation is always between  $-1$  and  $1$  with the values  $-1$  and  $1$  indicating a perfect *linear* relationship between  $X$  and  $Y$ . This is proved in Theorem 4.5.4.

Before investigating these properties of covariance and correlation, we will first calculate these measures in a given example. This calculation will be simplified by the following result.

**THEOREM 4.5.1:** For any random variables  $X$  and  $Y$ ,

$$\text{Cov}(X, Y) = EXY - \mu_X \mu_Y.$$

*Proof:*

$$\begin{aligned}\text{Cov}(X, Y) &= E((X - \mu_X)(Y - \mu_Y)) \\ &= E((XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y)) \quad (\text{expanding the product}) \\ &= EXY - \mu_X EY - \mu_Y EX + \mu_X \mu_Y \quad (\mu_X \text{ and } \mu_Y \text{ are constants}) \\ &= EXY - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y \\ &= EXY - \mu_X \mu_Y.\end{aligned}\quad \square$$

**Example 4.5.1:** Let the joint pdf of  $(X, Y)$  be  $f(x, y) = 1$ ,  $0 < x < 1, x < y < x + 1$ . The marginal distribution of  $X$  is uniform( $0, 1$ ) so  $\mu_X = \frac{1}{2}$  and  $\sigma_X^2 = \frac{1}{12}$ . The marginal pdf of  $Y$  is  $f_Y(y) = y$ ,  $0 < y < 1$ , and  $f_Y(y) = 2 - y$ ,  $1 \leq y < 2$ , with  $\mu_Y = 1$  and  $\sigma_Y^2 = \frac{1}{6}$ . We also have

$$\begin{aligned} EXY &= \int_0^1 \int_x^{x+1} xy \, dy \, dx = \int_0^1 \frac{1}{2} xy^2 \Big|_x^{x+1} \, dx \\ &= \int_0^1 \left( x^2 + \frac{1}{2}x \right) \, dx = \frac{7}{12}. \end{aligned}$$

Using Theorem 4.5.1, we have  $\text{Cov}(X, Y) = \frac{7}{12} - \left(\frac{1}{2}\right)(1) = \frac{1}{12}$ . The correlation is

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{1/12}{\sqrt{1/12} \sqrt{1/6}} = \frac{1}{\sqrt{2}}. \quad \square$$

In the next three theorems we describe some of the fundamental properties of covariance and correlation.

**THEOREM 4.5.2:** If  $X$  and  $Y$  are independent random variables then  $\text{Cov}(X, Y) = 0$  and  $\rho_{XY} = 0$ .

*Proof:* Since  $X$  and  $Y$  are independent, from Theorem 4.2.1 we have  $EXY = (EX)(EY)$ . Thus

$$\text{Cov}(X, Y) = EXY - (EX)(EY) = (EX)(EY) - (EX)(EY) = 0$$

and

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{0}{\sigma_X \sigma_Y} = 0. \quad \square$$

Thus, the values  $\text{Cov}(X, Y) = \rho_{XY} = 0$  in some sense indicate that there is no relationship between  $X$  and  $Y$ . It is important to note, however, that Theorem 4.5.2 does *not* say that if  $\text{Cov}(X, Y) = 0$  then  $X$  and  $Y$  are independent. In fact, covariance and correlation measure only a particular kind of *linear* relationship that will be described further in Theorem 4.5.4. Example 4.5.3 discusses two random variables that have a strong relationship but whose covariance and correlation are zero because the relationship is not linear.

Covariance also plays an important role in understanding the variation in sums of random variables, as the next theorem, a generalization of Theorem 2.3.1, indicates.

**THEOREM 4.5.3:** If  $X$  and  $Y$  are any two random variables, and  $a$  and  $b$  are any two constants, then

$$\text{Var}(aX + bY) = a^2 \text{Var } X + b^2 \text{Var } Y + 2ab \text{Cov}(X, Y).$$

If  $X$  and  $Y$  are independent random variables, then

$$\text{Var}(aX + bY) = a^2 \text{Var } X + b^2 \text{Var } Y.$$

*Proof:* The mean of  $aX + bY$  is  $E(aX + bY) = aEX + bEY = a\mu_X + b\mu_Y$ . Thus

$$\begin{aligned}\text{Var}(aX + bY) &= E((aX + bY) - (a\mu_X + b\mu_Y))^2 \\ &= E(a(X - \mu_X) + b(Y - \mu_Y))^2 \\ &= E(a^2(X - \mu_X)^2 + b^2(Y - \mu_Y)^2 + 2ab(X - \mu_X)(Y - \mu_Y)) \\ &= a^2E(X - \mu_X)^2 + b^2E(Y - \mu_Y)^2 + 2abE(X - \mu_X)(Y - \mu_Y) \\ &= a^2\text{Var } X + b^2\text{Var } Y + 2ab\text{Cov}(X, Y).\end{aligned}$$

If  $X$  and  $Y$  are independent then, from Theorem 4.5.2,  $\text{Cov}(X, Y) = 0$  and the second equality is immediate from the first.  $\square$

From Theorem 4.5.3 we see that if  $X$  and  $Y$  are positively correlated ( $\text{Cov}(X, Y) > 0$ ), then the variation in  $X + Y$  is greater than the sum of the variations in  $X$  and  $Y$ . But if they are negatively correlated then the variation in  $X + Y$  is less than the sum. For negatively correlated random variables, large values of one tend to be observed with small values of the other and in the sum these two extremes cancel. The result,  $X + Y$ , tends not to have as many extreme values and hence has smaller variance. By choosing  $a = 1$  and  $b = -1$  we get an expression for the variance of the difference of two random variables, and similar arguments apply.

The nature of the linear relationship measured by covariance and correlation is somewhat explained by the following theorem.

**THEOREM 4.5.4:** For any random variables  $X$  and  $Y$ ,

- a.  $-1 \leq \rho_{XY} \leq 1$ .
- b.  $|\rho_{XY}| = 1$  if and only if there exist numbers  $a \neq 0$  and  $b$  such that  $P(Y = aX + b) = 1$ . If  $\rho_{XY} = 1$  then  $a > 0$ , and if  $\rho_{XY} = -1$  then  $a < 0$ .

*Proof:* Consider the function  $h(t)$  defined by

$$h(t) = E((X - \mu_X)t + (Y - \mu_Y))^2.$$

Expanding this expression, we obtain

$$\begin{aligned}h(t) &= t^2E(X - \mu_X)^2 + 2tE(X - \mu_X)(Y - \mu_Y) + E(Y - \mu_Y)^2 \\ &= t^2\sigma_X^2 + 2t\text{Cov}(X, Y) + \sigma_Y^2.\end{aligned}$$

This quadratic function of  $t$  is greater than or equal to 0 for all values of  $t$  since it is the expected value of a nonnegative random variable. Thus, this quadratic function can have at most one real root and thus must have a nonpositive discriminant. That is,

$$(2\text{Cov}(X, Y))^2 - 4\sigma_X^2\sigma_Y^2 \leq 0.$$

This is equivalent to

$$-\sigma_X\sigma_Y \leq \text{Cov}(X, Y) \leq \sigma_X\sigma_Y.$$

Dividing by  $\sigma_X\sigma_Y$  yields

$$-1 \leq \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} = \rho_{XY} \leq 1.$$

Also,  $|\rho_{XY}| = 1$  if and only if the discriminant is equal to 0. That is,  $|\rho_{XY}| = 1$  if and only if  $h(t)$  has a single root. But since  $((X - \mu_X)t + (Y - \mu_Y))^2 \geq 0$ , the expected value  $h(t) = E((X - \mu_X)t + (Y - \mu_Y))^2 = 0$  if and only if

$$P([(X - \mu_X)t + (Y - \mu_Y)]^2 = 0) = 1.$$

This is equivalent to

$$P((X - \mu_X)t + (Y - \mu_Y) = 0) = 1.$$

This is  $P(Y = aX + b) = 1$  with  $a = -t$  and  $b = \mu_X t + \mu_Y$  where  $t$  is the root of  $h(t)$ . Using the quadratic formula, we see that this root is  $t = -\text{Cov}(X, Y)/\sigma_X^2$ . Thus  $a = -t$  has the same sign as  $\rho_{XY}$ , proving the final assertion.  $\square$

In Section 4.7 we will prove a theorem called the Cauchy–Schwarz Inequality. This theorem has as a direct consequence that  $\rho_{XY}$  is bounded between  $-1$  and  $1$  and we will see that, with this inequality, the preceding proof can be shortened.

If there is a line  $y = ax + b$ , with  $a \neq 0$ , such that the values of  $(X, Y)$  have a high probability of being near this line, then the correlation between  $X$  and  $Y$  will be near 1 or  $-1$ . But if no such line exists, the correlation will be near zero. This is an intuitive notion of the linear relationship that is being measured by correlation. This idea will be illustrated further in the next two examples.

**Example 4.5.2:** This example is similar to Example 4.5.1 but we develop it differently to illustrate other model building and computational techniques. Let  $X$  have a uniform(0, 1) distribution and  $Z$  have a uniform(0,  $\frac{1}{10}$ ) distribution. Suppose  $X$  and  $Z$  are independent. Let  $Y = X + Z$  and consider the random vector  $(X, Y)$ . The joint distribution of  $(X, Y)$  can be derived from the joint distribution of  $(X, Z)$  using the techniques of Section 4.3. The joint pdf of  $(X, Y)$  is

$$f(x, y) = 10, \quad 0 < x < 1, \quad x < y < x + \frac{1}{10}.$$

Rather than using the formal techniques of Section 4.3, we can justify this as follows. Given  $X = x$ ,  $Y = x + Z$ . The conditional distribution of  $Z$  given  $X = x$  is just uniform(0,  $\frac{1}{10}$ ) since  $X$  and  $Z$  are independent. Thus  $x$  serves as a location parameter in the conditional distribution of  $Y$  given  $X = x$ , and this conditional distribution is

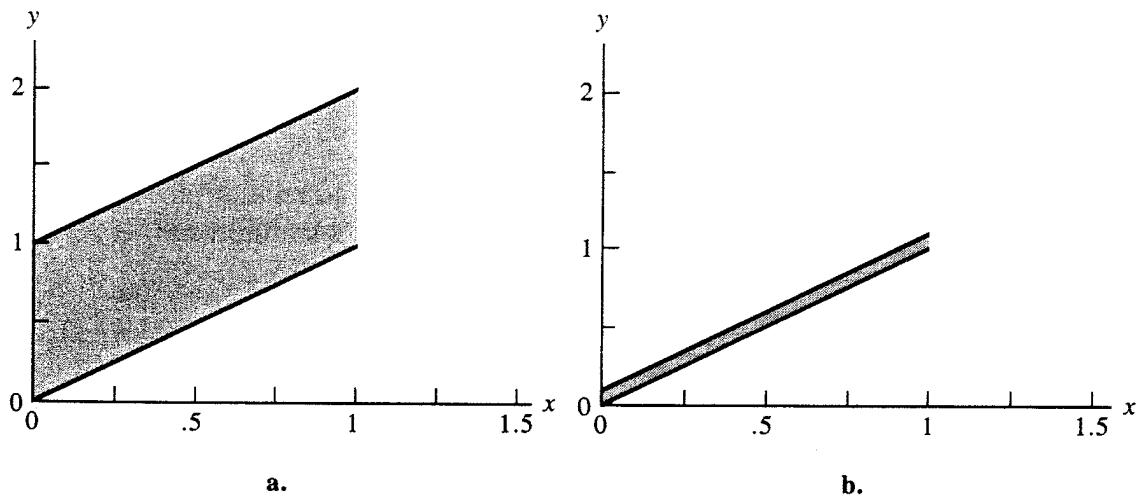
just  $\text{uniform}(x, x + \frac{1}{10})$ . Multiplying this conditional pdf by the marginal pdf of  $X$  ( $\text{uniform}(0, 1)$ ) yields the joint pdf above. This representation of  $Y = X + Z$  makes the computation of the covariance and correlation easy. The expected values of  $X$  and  $Y$  are  $EY = \frac{1}{2}$  and  $EY = E(X + Z) = EX + EZ = \frac{1}{2} + \frac{1}{20} = \frac{11}{20}$ , giving

$$\begin{aligned}\text{Cov}(X, Y) &= EXY - (EX)(EY) \\ &= EX(X + Z) - (EX)(E(X + Z)) \\ &= EX^2 + EXZ - (EX)^2 - (EX)(EZ) \\ &= EX^2 - (EX)^2 + (EX)(EZ) - (EX)(EZ) \quad \left( \begin{array}{l} \text{independence of} \\ X \text{ and } Z \end{array} \right) \\ &= \sigma_X^2 = \frac{1}{12}.\end{aligned}$$

From Theorem 4.5.3, the variance of  $Y$  is  $\sigma_Y^2 = \text{Var}(X + Z) = \text{Var } X + \text{Var } Z = \frac{1}{12} + \frac{1}{1200}$ . Thus

$$\rho_{XY} = \frac{\frac{1}{12}}{\sqrt{\frac{1}{12}} \sqrt{\frac{1}{12} + \frac{1}{1200}}} = \sqrt{\frac{100}{101}}.$$

This is much larger than the value of  $\rho_{XY} = 1/\sqrt{2}$  obtained in Example 4.5.1. The sets on which  $f(x, y)$  is positive for Example 4.5.1 and this example are illustrated in Figure 4.5.1. (Recall that this set is called the support of a distribution.) In each case,  $(X, Y)$  is a random point from the set. In both cases there is a linearly increasing relationship between  $X$  and  $Y$ . But the relationship is much stronger in Figure 4.5.1b. Another way to see this is by noting that in this example, the conditional distribution of  $Y$  given  $X = x$  is  $\text{uniform}(x, x + \frac{1}{10})$ . In Example 4.5.1,



**FIGURE 4.5.1** a. Region where  $f(x, y) > 0$  for Example 4.5.1; b. Region where  $f(x, y) > 0$  for Example 4.5.2

the conditional distribution of  $Y$  given  $X = x$  is uniform( $x, x + 1$ ). The knowledge that  $X = x$  gives us much more information about the value of  $Y$  in this model than in the one in Example 4.5.1. Hence the correlation is nearer to 1 in this example. ||

The next example illustrates that there may be a strong relationship between  $X$  and  $Y$  but, if the relationship is not linear, the correlation may be small.

**Example 4.5.3:** In this example, let  $X$  have a uniform( $-1, 1$ ) distribution and let  $Z$  have a uniform( $0, \frac{1}{10}$ ) distribution. Let  $X$  and  $Z$  be independent. Let  $Y = X^2 + Z$  and consider the random vector  $(X, Y)$ . As in Example 4.5.2, given  $X = x$ ,  $Y = x^2 + Z$  and the conditional distribution of  $Y$  given  $X = x$  is uniform( $x^2, x^2 + \frac{1}{10}$ ). The joint pdf of  $X$  and  $Y$ , the product of this conditional pdf and the marginal pdf of  $X$ , is thus

$$f(x, y) = 5, \quad -1 < x < 1, \quad x^2 < y < x^2 + \frac{1}{10}.$$

The set on which  $f(x, y) > 0$  is illustrated in Figure 4.5.2. There is a strong relationship between  $X$  and  $Y$ , as indicated by the conditional distribution of  $Y$  given  $X = x$ . But the relationship is not linear. The possible values of  $(X, Y)$  cluster

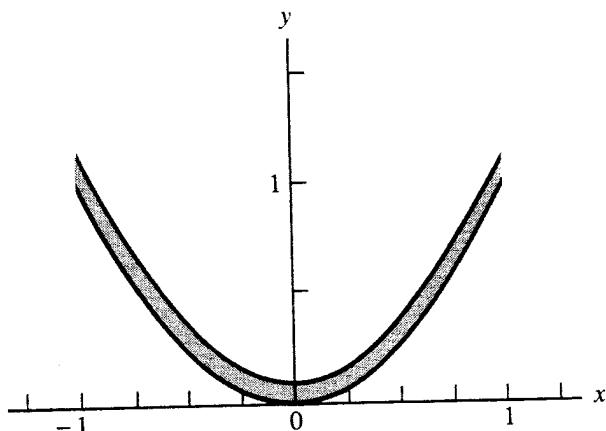


FIGURE 4.5.2 Region where  $f(x, y) > 0$  for Example 4.5.3

around a parabola rather than a straight line. The correlation does not measure this nonlinear relationship. In fact,  $\rho_{XY} = 0$ . Since  $X$  has a uniform( $-1, 1$ ) distribution,  $EX = EX^3 = 0$  and since  $X$  and  $Z$  are independent,  $EXZ = (EX)(EZ)$ . Thus,

$$\begin{aligned} \text{Cov}(X, Y) &= E(X(X^2 + Z)) - (EX)(E(X^2 + Z)) \\ &= EX^3 + EXZ - 0E(X^2 + Z) \\ &= 0 + (EX)(EZ) = 0(EZ) = 0 \end{aligned}$$

and  $\rho_{XY} = \text{Cov}(X, Y)/(\sigma_X \sigma_Y) = 0$ . ||

We close this section by introducing a very important bivariate distribution in which the correlation coefficient arises naturally as a parameter.

**DEFINITION 4.5.3:** Let  $-\infty < \mu_X < \infty$ ,  $-\infty < \mu_Y < \infty$ ,  $0 < \sigma_X$ ,  $0 < \sigma_Y$ , and  $-1 < \rho < 1$  be five real numbers. The *bivariate normal pdf with means  $\mu_X$  and  $\mu_Y$ , variances  $\sigma_X^2$  and  $\sigma_Y^2$ , and correlation  $\rho$*  is the bivariate pdf given by

$$f(x, y) = \left( 2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2} \right)^{-1} \times \exp \left( -\frac{1}{2(1-\rho^2)} \left( \left( \frac{x-\mu_X}{\sigma_X} \right)^2 - 2\rho \left( \frac{x-\mu_X}{\sigma_X} \right) \left( \frac{y-\mu_Y}{\sigma_Y} \right) + \left( \frac{y-\mu_Y}{\sigma_Y} \right)^2 \right) \right)$$

for  $-\infty < x < \infty$  and  $-\infty < y < \infty$ .

Although the formula for the bivariate normal pdf looks formidable, this bivariate distribution is one of the most frequently used. (In fact, the derivation of the formula need not be formidable at all. See Exercise 4.42.)

The many nice properties of this distribution include these:

- a. The marginal distribution of  $X$  is  $n(\mu_X, \sigma_X^2)$ .
- b. The marginal distribution of  $Y$  is  $n(\mu_Y, \sigma_Y^2)$ .
- c. The correlation between  $X$  and  $Y$  is  $\rho_{XY} = \rho$ .
- d. For any constants  $a$  and  $b$ , the distribution of  $aX + bY$  is  $n(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$ .

We will leave the verification of properties (a), (b), and (d) as exercises (Exercise 4.41). Assuming (a) and (b) are true, we will prove (c). We have by definition

$$\begin{aligned} \rho_{XY} &= \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} \\ &= \frac{\mathbb{E}(X - \mu_X)(Y - \mu_Y)}{\sigma_X\sigma_Y} \\ &= \mathbb{E} \left( \frac{X - \mu_X}{\sigma_X} \right) \left( \frac{Y - \mu_Y}{\sigma_Y} \right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left( \frac{x - \mu_X}{\sigma_X} \right) \left( \frac{y - \mu_Y}{\sigma_Y} \right) f(x, y) dx dy. \end{aligned}$$

Make the change of variable

$$s = \left( \frac{x - \mu_X}{\sigma_X} \right) \left( \frac{y - \mu_Y}{\sigma_Y} \right) \quad \text{and} \quad t = \left( \frac{x - \mu_X}{\sigma_X} \right).$$

Then  $x = \sigma_X t + \mu_X$ ,  $y = (\sigma_Y s/t) + \mu_Y$ , and the Jacobian of the transformation is  $J = \sigma_X\sigma_Y/t$ . With this change of variable, we obtain

$$\begin{aligned}\rho_{XY} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} sf\left(\sigma_X t + \mu_X, \frac{\sigma_Y s}{t} + \mu_Y\right) \left| \frac{\sigma_X \sigma_Y}{t} \right| ds dt \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s \left(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\right)^{-1} \\ &\quad \exp\left(-\frac{1}{2(1-\rho^2)}\left(t^2 - 2\rho s + \left(\frac{s}{t}\right)^2\right)\right) \frac{\sigma_X \sigma_Y}{|t|} ds dt.\end{aligned}$$

Noting that  $|t| = \sqrt{t^2}$  and  $t^2 - 2\rho s + \left(\frac{s}{t}\right)^2 = \left(\frac{s-\rho t^2}{t}\right)^2 + (1-\rho^2)t^2$ ,

we can rewrite this as

$$\rho_{XY} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \left[ \int_{-\infty}^{\infty} \frac{s}{\sqrt{2\pi}\sqrt{(1-\rho^2)t^2}} \exp\left(-\frac{(s-\rho t^2)^2}{2(1-\rho^2)t^2}\right) ds \right] dt.$$

The inner integral is  $ES$ , where  $S$  is a normal random variable with  $ES = \rho t^2$  and  $\text{Var } S = (1-\rho^2)t^2$ . Thus the inner integral is  $\rho t^2$ . Hence we have

$$\rho_{XY} = \int_{-\infty}^{\infty} \frac{\rho t^2}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt.$$

But this integral is  $\rho ET^2$ , where  $T$  is a  $n(0, 1)$  random variable. Hence  $ET^2 = 1$  and  $\rho_{XY} = \rho$ .

All the conditional distributions of  $Y$  given  $X = x$  and of  $X$  given  $Y = y$  are also normal distributions. Using the joint and marginal pdfs given above, it is straightforward to verify that

the conditional distribution of  $Y$  given  $X = x$  is

$$n(\mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X), \sigma_Y^2(1 - \rho^2)).$$

As  $\rho$  converges to 1 or  $-1$ , the conditional variance  $\sigma_Y^2(1 - \rho^2)$  converges to 0. Thus, the conditional distribution of  $Y$  given  $X = x$  becomes more concentrated about the point  $\mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X)$ , and the joint probability distribution of  $(X, Y)$  becomes more concentrated about the line  $y = \mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X)$ . This illustrates again the point made earlier that a correlation near 1 or  $-1$  means that there is a line  $y = ax + b$  about which the values of  $(X, Y)$  cluster with high probability.

Note one important fact: All of the normal marginal and conditional pdfs are derived from the starting point of bivariate normality. The derivation does not go in the opposite direction. That is, marginal normality does not imply joint normality. See Exercise 4.43 for an illustration of this.

## 4.6 Multivariate Distributions

At the beginning of this chapter, we discussed observing more than two random variables in an experiment. In the previous sections our discussions have concentrated

on a bivariate random vector  $(X, Y)$ . In this section we discuss a multivariate random vector  $(X_1, \dots, X_n)$ . In the example at the beginning of this chapter, temperature, height, weight, and blood pressure were observed on an individual. In this example,  $n = 4$  and the observed random vector is  $(X_1, X_2, X_3, X_4)$  where  $X_1$  is temperature,  $X_2$  is height, etc. The concepts from the earlier sections, including marginal and conditional distributions, generalize from the bivariate to the multivariate setting. We introduce some of these generalizations in this section.

*A note on notation:* We will use boldface letters to denote multiple variates. Thus, we write  $\mathbf{X}$  to denote the random variables  $X_1, \dots, X_n$ , and  $\mathbf{x}$  to denote the sample  $x_1, \dots, x_n$ .

The random vector  $\mathbf{X} = (X_1, \dots, X_n)$  has a sample space that is a subset of  $\mathbb{R}^n$ . If  $(X_1, \dots, X_n)$  is a discrete random vector (the sample space is countable), then the *joint pmf of*  $(X_1, \dots, X_n)$  is the function defined by  $f(\mathbf{x}) = f(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n)$  for each  $(x_1, \dots, x_n) \in \mathbb{R}^n$ . Then for any  $A \subset \mathbb{R}^n$ ,

$$(4.6.1) \quad P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} f(\mathbf{x}).$$

If  $(X_1, \dots, X_n)$  is a continuous random vector, the *joint pdf of*  $(X_1, \dots, X_n)$  is a function  $f(x_1, \dots, x_n)$  that satisfies

$$(4.6.2) \quad P(\mathbf{X} \in A) = \int_A \cdots \int f(\mathbf{x}) d\mathbf{x} = \int_A \cdots \int f(x_1, \dots, x_n) dx_1 \cdots dx_n.$$

These integrals are  $n$ -fold integrals with limits of integration set so that the integration is over all points  $\mathbf{x} \in A$ .

Let  $g(\mathbf{x}) = g(x_1, \dots, x_n)$  be a real-valued function defined on the sample space of  $\mathbf{X}$ . Then  $g(\mathbf{X})$  is a random variable and the *expected value of*  $g(\mathbf{X})$  is

$$(4.6.3) \quad \text{E}g(\mathbf{X}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \quad \text{and} \quad \text{E}g(\mathbf{X}) = \sum_{\mathbf{x} \in \mathbb{R}^n} g(\mathbf{x}) f(\mathbf{x})$$

in the continuous and discrete cases, respectively. These and other definitions are analogous to the bivariate definitions except that now the integrals or sums are over the appropriate subset of  $\mathbb{R}^n$  rather than  $\mathbb{R}^2$ .

The *marginal pdf or pmf* of any subset of the coordinates of  $(X_1, \dots, X_n)$  can be computed by integrating or summing the joint pdf or pmf over all possible values of the other coordinates. Thus, for example, the marginal distribution of  $(X_1, \dots, X_k)$ , the first  $k$  coordinates of  $(X_1, \dots, X_n)$ , is given by the pdf or pmf

$$(4.6.4) \quad f(x_1, \dots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) dx_{k+1} \cdots dx_n$$

or

$$(4.6.5) \quad f(x_1, \dots, x_k) = \sum_{(x_{k+1}, \dots, x_n) \in \mathbb{R}^{n-k}} f(x_1, \dots, x_n)$$

for every  $(x_1, \dots, x_k) \in \mathbb{R}^k$ . The *conditional pdf or pmf* of a subset of the coordinates of  $(X_1, \dots, X_n)$  given the values of the remaining coordinates is obtained by dividing the joint pdf or pmf by the marginal pdf or pmf of the remaining coordinates. Thus, for example, if  $f(x_1, \dots, x_k) > 0$ , the conditional pdf or pmf of  $(X_{k+1}, \dots, X_n)$  given  $X_1 = x_1, \dots, X_k = x_k$  is the function of  $(x_{k+1}, \dots, x_n)$  defined by

$$(4.6.6) \quad f(x_{k+1}, \dots, x_n | x_1, \dots, x_k) = \frac{f(x_1, \dots, x_n)}{f(x_1, \dots, x_k)}.$$

These ideas are illustrated in the following example.

**Example 4.6.1:** Let  $n = 4$  and

$$f(x_1, x_2, x_3, x_4) = \begin{cases} \frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2), & 0 < x_i < 1, i = 1, 2, 3, 4 \\ 0 & \text{otherwise} \end{cases}$$

This nonnegative function is the joint pdf of a random vector  $(X_1, X_2, X_3, X_4)$  and it can be verified that

$$\begin{aligned} & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4) dx_1 dx_2 dx_3 dx_4 \\ &= \int_0^1 \int_0^1 \int_0^1 \int_0^1 \frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2) dx_1 dx_2 dx_3 dx_4 = 1. \end{aligned}$$

This joint pdf can be used to compute probabilities such as

$$\begin{aligned} & P\left(X_1 < \frac{1}{2}, X_2 < \frac{3}{4}, X_4 > \frac{1}{2}\right) \\ &= \int_{\frac{1}{2}}^1 \int_0^1 \int_0^{\frac{3}{4}} \int_0^{\frac{1}{2}} \frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2) dx_1 dx_2 dx_3 dx_4. \end{aligned}$$

Note how the limits of integration restrict the integration to those values of  $(x_1, x_2, x_3, x_4)$  which are in the event in question and for which  $f(x_1, x_2, x_3, x_4) > 0$ . Each of the four terms,  $\frac{3}{4}x_1^2$ ,  $\frac{3}{4}x_2^2$ , etc., can be integrated separately and the results summed. For example,

$$\int_{\frac{1}{2}}^1 \int_0^1 \int_0^{\frac{3}{4}} \int_0^{\frac{1}{2}} \frac{3}{4}x_1^2 dx_1 dx_2 dx_3 dx_4 = \frac{3}{256}.$$

The other three integrals are  $\frac{7}{1024}$ ,  $\frac{3}{64}$ , and  $\frac{21}{256}$ . Thus

$$P(X_1 < \frac{1}{2}, X_2 < \frac{3}{4}, X_4 > \frac{1}{2}) = \frac{3}{256} + \frac{7}{1024} + \frac{3}{64} + \frac{21}{256} = \frac{151}{1024}.$$

Using (4.6.4), we can obtain the marginal pdf of  $(X_1, X_2)$  by integrating out the variables  $x_3$  and  $x_4$  to obtain

$$\begin{aligned} f(x_1, x_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4) dx_3 dx_4 \\ &= \int_0^1 \int_0^1 \frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2) dx_3 dx_4 = \frac{3}{4}(x_1^2 + x_2^2) + \frac{1}{2} \end{aligned}$$

for  $0 < x_1 < 1$  and  $0 < x_2 < 1$ . Any probability or expected value which involves only  $X_1$  and  $X_2$  can be computed using this marginal pdf. For example,

$$\begin{aligned} EX_1 X_2 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 f(x_1, x_2) dx_1 dx_2 \\ &= \int_0^1 \int_0^1 x_1 x_2 \left( \frac{3}{4}(x_1^2 + x_2^2) + \frac{1}{2} \right) dx_1 dx_2 \\ &= \int_0^1 \int_0^1 \left( \frac{3}{4}x_1^3 x_2 + \frac{3}{4}x_1 x_2^3 + \frac{1}{2}x_1 x_2 \right) dx_1 dx_2 \\ &= \int_0^1 \left( \frac{3}{16}x_2 + \frac{3}{8}x_2^3 + \frac{1}{4}x_2 \right) dx_2 = \frac{3}{32} + \frac{3}{32} + \frac{1}{8} = \frac{5}{16}. \end{aligned}$$

For any  $(x_1, x_2)$  with  $0 < x_1 < 1$  and  $0 < x_2 < 1$ ,  $f(x_1, x_2) > 0$  and the conditional pdf of  $(X_3, X_4)$  given  $X_1 = x_1$  and  $X_2 = x_2$  can be found using (4.6.6). For any such  $(x_1, x_2)$ ,  $f(x_1, x_2, x_3, x_4) > 0$  if  $0 < x_3 < 1$  and  $0 < x_4 < 1$ , and for these values of  $(x_3, x_4)$ , the conditional pdf is

$$\begin{aligned} f(x_3, x_4 | x_1, x_2) &= \frac{f(x_1, x_2, x_3, x_4)}{f(x_1, x_2)} \\ &= \frac{\frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2)}{\frac{3}{4}(x_1^2 + x_2^2) + \frac{1}{2}} \\ &= \frac{x_1^2 + x_2^2 + x_3^2 + x_4^2}{x_1^2 + x_2^2 + \frac{2}{3}}. \end{aligned}$$

For example, the conditional pdf of  $(X_3, X_4)$  given  $X_1 = \frac{1}{3}$  and  $X_2 = \frac{2}{3}$  is

$$f(x_3, x_4 | x_1 = \frac{1}{3}, x_2 = \frac{2}{3}) = \frac{\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + x_3^2 + x_4^2}{\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \frac{2}{3}} = \frac{5}{11} + \frac{9}{11}x_3^2 + \frac{9}{11}x_4^2.$$

This can be used to compute

$$\begin{aligned}
 P\left(X_3 > \frac{3}{4}, X_4 < \frac{1}{2} \mid X_1 = \frac{1}{3}, X_2 = \frac{2}{3}\right) &= \int_0^{\frac{1}{2}} \int_{\frac{3}{4}}^1 \left(\frac{5}{11} + \frac{9}{11}x_3^2 + \frac{9}{11}x_4^2\right) dx_3 dx_4 \\
 &= \int_0^{\frac{1}{2}} \left(\frac{5}{44} + \frac{111}{704} + \frac{9}{44}x_4^2\right) dx_4 \\
 &= \frac{5}{88} + \frac{111}{1408} + \frac{3}{352} = \frac{203}{1408}. \quad \parallel
 \end{aligned}$$

Before giving examples of computations with conditional and marginal distributions for a discrete multivariate random vector, we will introduce an important family of discrete multivariate distributions. This family generalizes the binomial family to the situation in which each trial has  $n$  (rather than two) distinct possible outcomes.

**DEFINITION 4.6.1:** Let  $n$  and  $m$  be positive integers and let  $p_1, \dots, p_n$  be numbers satisfying  $0 \leq p_i \leq 1$ ,  $i = 1, \dots, n$ , and  $\sum_{i=1}^n p_i = 1$ . Then the random vector  $(X_1, \dots, X_n)$  has a *multinomial distribution with  $m$  trials and cell probabilities  $p_1, \dots, p_n$*  if the joint pmf of  $(X_1, \dots, X_n)$  is

$$f(x_1, \dots, x_n) = \frac{m!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n} = m! \prod_{i=1}^n \frac{p_i^{x_i}}{x_i!}$$

on the set of  $(x_1, \dots, x_n)$  such that each  $x_i$  is a nonnegative integer and  $\sum_{i=1}^n x_i = m$ .

The multinomial distribution is a model for the following kind of experiment. The experiment consists of  $m$  independent trials. Each trial results in one of  $n$  distinct possible outcomes. The probability of the  $i$ th outcome is  $p_i$  on every trial. And  $X_i$  is the count of the number of times the  $i$ th outcome occurred in the  $m$  trials. For  $n = 2$ , this is just a binomial experiment in which each trial has  $n = 2$  possible outcomes and  $X_1$  counts the number of “successes” and  $X_2 = m - X_1$  counts the number of “failures” in  $m$  trials. In a general multinomial experiment, there are  $n$  different possible outcomes to count.

**Example 4.6.2:** Consider tossing a six-sided die ten times. Suppose the die is unbalanced so that the probability of observing a 1 is  $\frac{1}{21}$ , the probability of observing a 2 is  $\frac{2}{21}$ , and, in general, the probability of observing an  $i$  is  $\frac{i}{21}$ . Now consider the random vector  $(X_1, \dots, X_6)$  where  $X_i$  counts the number of times  $i$  comes up in the ten tosses. Then  $(X_1, \dots, X_6)$  has a multinomial distribution with  $m = 10$  trials,  $n = 6$  possible outcomes, and cell probabilities  $p_1 = \frac{1}{21}, p_2 = \frac{2}{21}, \dots, p_6 = \frac{6}{21}$ . The formula in Definition 4.6.1 may be used to calculate the probability of rolling four 6s, three 5s, two 4s, and one 3 to be

$$\begin{aligned}
 f(0, 0, 1, 2, 3, 4) &= \frac{10!}{0!0!1!2!3!4!} \left(\frac{1}{21}\right)^0 \left(\frac{2}{21}\right)^0 \left(\frac{3}{21}\right)^1 \left(\frac{4}{21}\right)^2 \left(\frac{5}{21}\right)^3 \left(\frac{6}{21}\right)^4 = .0059. \quad \parallel
 \end{aligned}$$

The factor  $m!/(x_1! \cdots x_n!)$  is called a *multinomial coefficient*. It is the number of ways that  $m$  objects can be divided into  $n$  groups with  $x_1$  in the first group,  $x_2$  in the second group, ..., and  $x_n$  in the  $n$ th group. A generalization of the Binomial Theorem 3.1.1 is the Multinomial Theorem.

**THEOREM 4.6.1 (Multinomial Theorem):** Let  $m$  and  $n$  be positive integers. Let  $\mathcal{A}$  be the set of vectors  $\mathbf{x} = (x_1, \dots, x_n)$  such that each  $x_i$  is a nonnegative integer and  $\sum_{i=1}^n x_i = m$ . Then, for any real numbers  $p_1, \dots, p_n$ ,

$$(p_1 + \cdots + p_n)^m = \sum_{\mathbf{x} \in \mathcal{A}} \frac{m!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n}. \quad \square$$

Theorem 4.6.1 shows that a multinomial pmf sums to 1. The set  $\mathcal{A}$  is the set of points with positive probability in Definition 4.6.1. The sum of the pmf over all those points is, by Theorem 4.6.1,  $(p_1 + \cdots + p_n)^m = 1^m = 1$ .

Now we consider some marginal and conditional distributions for the multinomial model. Consider a single coordinate  $X_i$ . If the occurrence of the  $i$ th outcome is labelled a “success” and anything else is labelled a “failure,” then  $X_i$  is the count of the number of successes in  $m$  independent trials where the probability of a success is  $p_i$  on each trial. Thus  $X_i$  should have a binomial( $m, p_i$ ) distribution. To verify this the marginal distribution of  $X_i$  should be computed using (4.6.5). For example, consider the marginal pmf of  $X_n$ . For a fixed value of  $x_n \in \{0, 1, \dots, n\}$ , to compute the marginal pmf  $f(x_n)$ , we must sum over all possible values of  $(x_1, \dots, x_{n-1})$ . That is, we must sum over all  $(x_1, \dots, x_{n-1})$  such that the  $x_i$ s are all nonnegative integers and  $\sum_{i=1}^{n-1} x_i = m - x_n$ . Denote this set by  $\mathcal{B}$ . Then

$$\begin{aligned} f(x_n) &= \sum_{(x_1, \dots, x_{n-1}) \in \mathcal{B}} \frac{m!}{x_1! \cdots x_{n-1}!} (p_1)^{x_1} \cdots (p_{n-1})^{x_{n-1}} (p_n)^{x_n} \\ &= \sum_{(x_1, \dots, x_{n-1}) \in \mathcal{B}} \frac{m!}{x_1! \cdots x_{n-1}!} p_1^{x_1} \cdots p_{n-1}^{x_{n-1}} p_n^{x_n} \frac{(m-x_n)!}{(m-x_n)!} \frac{(1-p_n)^{m-x_n}}{(1-p_n)^{m-x_n}} \\ &= \frac{m!}{x_n!(m-x_n)!} p_n^{x_n} (1-p_n)^{m-x_n} \\ &\quad \times \sum_{(x_1, \dots, x_{n-1}) \in \mathcal{B}} \frac{(m-x_n)!}{x_1! \cdots x_{n-1}!} \left(\frac{p_1}{1-p_n}\right)^{x_1} \cdots \left(\frac{p_{n-1}}{1-p_n}\right)^{x_{n-1}}. \end{aligned}$$

But using the facts that  $x_1 + \cdots + x_{n-1} = m - x_n$  and  $p_1 + \cdots + p_{n-1} = 1 - p_n$ , and Theorem 4.6.1, we see that the last summation is 1. Hence the marginal distribution of  $X_n$  is binomial( $m, p_n$ ). Similar arguments show that each of the other coordinates is marginally binomially distributed.

Given that  $X_n = x_n$ , there must have been  $m - x_n$  trials that resulted in one of the first  $n - 1$  outcomes. The vector  $(X_1, \dots, X_{n-1})$  counts the number of these  $m - x_n$  trials that are of each type. Thus it seems that given  $X_n = x_n$ ,  $(X_1, \dots, X_{n-1})$

might have a multinomial distribution. This is true. Using (4.6.6), the conditional pmf of  $(X_1, \dots, X_{n-1})$  given  $X_n = x_n$  is

$$\begin{aligned} f(x_1, \dots, x_{n-1} | x_n) &= \frac{f(x_1, \dots, x_n)}{f(x_n)} \\ &= \frac{\frac{m!}{x_1! \cdots x_n!} (p_1)^{x_1} \cdots (p_n)^{x_n}}{\frac{m!}{x_n!(m-x_n)!} (p_n)^{x_n} (1-p_n)^{m-x_n}} \\ &= \frac{(m-x_n)!}{x_1! \cdots x_{n-1}!} \left( \frac{p_1}{1-p_n} \right)^{x_1} \cdots \left( \frac{p_{n-1}}{1-p_n} \right)^{x_{n-1}}. \end{aligned}$$

This is the pmf of a multinomial distribution with  $m - x_n$  trials and cell probabilities  $p_1/(1-p_n), \dots, p_{n-1}/(1-p_n)$ . In fact, the conditional distribution of any subset of the coordinates of  $(X_1, \dots, X_n)$  given the values of the rest of the coordinates is a multinomial distribution.

As in the bivariate case, relationships between coordinates of a random vector are important. The important situation in which there are no relationships of any kind is called mutual independence.

**DEFINITION 4.6.2:** Let  $\mathbf{X}_1, \dots, \mathbf{X}_n$  be random vectors with joint pdf or pmf  $f(\mathbf{x}_1, \dots, \mathbf{x}_n)$ . Let  $f_{\mathbf{X}_i}(\mathbf{x}_i)$  denote the marginal pdf or pmf of  $\mathbf{X}_i$ . Then  $\mathbf{X}_1, \dots, \mathbf{X}_n$  are called *mutually independent random vectors* if, for every  $(\mathbf{x}_1, \dots, \mathbf{x}_n)$

$$f(\mathbf{x}_1, \dots, \mathbf{x}_n) = f_{\mathbf{X}_1}(\mathbf{x}_1) \cdots f_{\mathbf{X}_n}(\mathbf{x}_n) = \prod_{i=1}^n f_{\mathbf{X}_i}(\mathbf{x}_i).$$

If the  $\mathbf{X}_i$ s are all one-dimensional, then  $X_1, \dots, X_n$  are called *mutually independent random variables*.

If  $X_1, \dots, X_n$  are mutually independent, then knowledge about the values of some coordinates gives us no information about the values of the other coordinates. Using Definition 4.6.2, one can show that the conditional distribution of any subset of the coordinates, given the values of the rest of the coordinates, is the same as the marginal distribution of the subset. Mutual independence implies that any pair, say  $X_i$  and  $X_j$ , are pairwise independent. That is, their bivariate marginal pdf or pmf,  $f(x_i, x_j)$ , satisfies Definition 4.2.3. But mutual independence implies more than pairwise independence. As in Example 1.3.6, it is possible to specify a probability distribution for  $(X_1, \dots, X_n)$  with the property that each pair,  $(X_i, X_j)$ , is pairwise independent but  $X_1, \dots, X_n$  are not mutually independent.

Mutually independent random variables have many nice properties. The proofs of the following theorems are analogous to the proofs of their counterparts in Sections 4.2 and 4.3.

**THEOREM 4.6.2 (Generalization of Theorem 4.2.1):** Let  $X_1, \dots, X_n$  be mutually independent random variables. Let  $g_1, \dots, g_n$  be real-valued functions such that  $g_i(x_i)$  is a function only of  $x_i, i = 1, \dots, n$ . Then

$$\mathbf{E}(g_1(X_1) \cdots g_n(X_n)) = (\mathbf{E}g_1(X_1)) \cdots (\mathbf{E}g_n(X_n)). \quad \square$$

**THEOREM 4.6.3 (Generalization of Theorem 4.2.2):** Let  $X_1, \dots, X_n$  be mutually independent random variables with mgfs  $M_{X_1}(t), \dots, M_{X_n}(t)$ . Let  $Z = X_1 + \cdots + X_n$ . Then the mgf of  $Z$  is

$$M_Z(t) = M_{X_1}(t) \cdots M_{X_n}(t).$$

In particular, if  $X_1, \dots, X_n$  all have the same distribution with mgf  $M_X(t)$ , then

$$M_Z(t) = (M_X(t))^n. \quad \square$$

**Example 4.6.3:** Suppose  $X_1, \dots, X_n$  are mutually independent random variables, and the distribution of  $X_i$  is  $\text{gamma}(\alpha_i, \beta)$ . From Example 2.3.3, the mgf of a  $\text{gamma}(\alpha, \beta)$  distribution is  $M(t) = (1 - \beta t)^{-\alpha}$ . Thus, if  $Z = X_1 + \cdots + X_n$ , the mgf of  $Z$  is

$$M_Z(t) = M_{X_1}(t) \cdots M_{X_n}(t) = (1 - \beta t)^{-\alpha_1} \cdots (1 - \beta t)^{-\alpha_n} = (1 - \beta t)^{-(\alpha_1 + \cdots + \alpha_n)}.$$

This is the mgf of a  $\text{gamma}(\alpha_1 + \cdots + \alpha_n, \beta)$  distribution. Thus, the sum of independent gamma random variables that have a common scale parameter  $\beta$  also has a gamma distribution.  $\square$

A generalization of Theorem 4.6.3 is obtained if we consider a sum of linear functions of independent random variables.

**COROLLARY 4.6.1:** Let  $X_1, \dots, X_n$  be mutually independent random variables with mgfs  $M_{X_1}(t), \dots, M_{X_n}(t)$ . Let  $a_1, \dots, a_n$  and  $b_1, \dots, b_n$  be fixed constants. Let  $Z = (a_1 X_1 + b_1) + \cdots + (a_n X_n + b_n)$ . Then the mgf of  $Z$  is

$$M_Z(t) = (e^{t(\sum b_i)}) M_{X_1}(a_1 t) \cdots M_{X_n}(a_n t).$$

*Proof:* From the definition, the mgf of  $Z$  is

$$\begin{aligned} M_Z(t) &= \mathbf{E} e^{tZ} \\ &= \mathbf{E} e^{t \sum (a_i X_i + b_i)} \\ &= (e^{t(\sum b_i)}) \mathbf{E}(e^{ta_1 X_1} \cdots e^{ta_n X_n}) \quad \left( \begin{array}{l} \text{properties of exponentials} \\ \text{and expectations} \end{array} \right) \\ &= (e^{t(\sum b_i)}) M_{X_1}(a_1 t) \cdots M_{X_n}(a_n t), \quad (\text{Theorem 4.6.2}) \end{aligned}$$

as was to be shown.  $\square$

Undoubtedly, the most important application of Corollary 4.6.1 is to the case of normal random variables. A *linear combination of independent normal random variables is normally distributed*.

**COROLLARY 4.6.2:** Let  $X_1, \dots, X_n$  be mutually independent random variables with  $X_i \sim n(\mu_i, \sigma_i^2)$ . Let  $a_1, \dots, a_n$  and  $b_1, \dots, b_n$  be fixed constants. Then

$$Z = \sum_{i=1}^n (a_i X_i + b_i) \sim n\left(\sum_{i=1}^n (a_i \mu_i + b_i), \sum_{i=1}^n a_i^2 \sigma_i^2\right).$$

*Proof:* Recall that the mgf of a  $n(\mu, \sigma^2)$  random variable is  $M(t) = e^{\mu t + \sigma^2 t^2/2}$ . Substituting into the expression in Corollary 4.6.1 yields

$$\begin{aligned} M_Z(t) &= (e^{t(\sum b_i)}) e^{\mu_1 a_1 t + \sigma_1^2 a_1^2 t^2/2} \cdots e^{\mu_n a_n t + \sigma_n^2 a_n^2 t^2/2} \\ &= e^{((\sum (a_i \mu_i + b_i))t + (\sum a_i^2 \sigma_i^2)t^2/2)}, \end{aligned}$$

the mgf of the indicated normal distribution.  $\square$

**THEOREM 4.6.4 (Generalization of Lemma 4.2.1):** Let  $X_1, \dots, X_n$  be random vectors. Then  $X_1, \dots, X_n$  are mutually independent random vectors if and only if there exist functions  $g_i(x_i)$ ,  $i = 1, \dots, n$ , such that the joint pdf or pmf of  $(X_1, \dots, X_n)$  can be written as

$$f(x_1, \dots, x_n) = g_1(x_1) \cdots g_n(x_n). \quad \square$$

**THEOREM 4.6.5 (Generalization of Theorem 4.3.2):** Let  $X_1, \dots, X_n$  be independent random vectors. Let  $g_i(x_i)$  be a function only of  $x_i$ ,  $i = 1, \dots, n$ . Then the random variables  $U_i = g_i(X_i)$ ,  $i = 1, \dots, n$ , are mutually independent.  $\square$

We close this section by describing the generalization of a technique for finding the distribution of a transformation of a random vector. We will present the generalization of formula (4.3.6) that gives the pdf of the new random vector in terms of the pdf of the original random vector. Note that to fully understand the remainder of this section, some knowledge of matrix algebra is required. (See, for example, Searle (1982).) In particular, we will need to compute the determinant of a matrix. This is the only place in the book where such knowledge is required.

Let  $(X_1, \dots, X_n)$  be a random vector with pdf  $f_X(x_1, \dots, x_n)$ . Let  $\mathcal{A} = \{x : f_X(x) > 0\}$ . Consider a new random vector,  $(U_1, \dots, U_n)$ , defined by  $U_1 = g_1(X_1, \dots, X_n)$ ,  $U_2 = g_2(X_1, \dots, X_n)$ ,  $\dots$ ,  $U_n = g_n(X_1, \dots, X_n)$ . Suppose that  $A_0, A_1, \dots, A_k$  form a partition of  $\mathcal{A}$  with these properties. The set  $A_0$ , which may be empty, satisfies  $P((X_1, \dots, X_n) \in A_0) = 0$ . The transformation  $(U_1, \dots, U_n) = (g_1(X), \dots, g_n(X))$  is a one-to-one transformation from  $A_i$  onto  $\mathcal{B}$  for each  $i = 1, 2, \dots, k$ . Then for each  $i$ , the inverse functions from  $\mathcal{B}$  to  $A_i$  can be found. Denote the  $i$ th inverse by  $x_1 = h_{1i}(u_1, \dots, u_n)$ ,  $x_2 = h_{2i}(u_1, \dots, u_n)$ ,  $\dots$ ,  $x_n = h_{ni}(u_1, \dots, u_n)$ . This  $i$ th inverse gives, for  $(u_1, \dots, u_n) \in \mathcal{B}$ , the unique  $(x_1, \dots, x_n)$

$\in A_i$  such that  $(u_1, \dots, u_n) = (g_1(x_1, \dots, x_n), \dots, g_n(x_1, \dots, x_n))$ . Let  $J_i$  denote the Jacobian computed from the  $i$ th inverse. That is,

$$J_i = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} & \dots & \frac{\partial x_1}{\partial u_n} \\ \frac{\partial x_2}{\partial u_1} & \frac{\partial x_2}{\partial u_2} & \dots & \frac{\partial x_2}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \frac{\partial x_n}{\partial u_2} & \dots & \frac{\partial x_n}{\partial u_n} \end{vmatrix} = \begin{vmatrix} \frac{\partial h_{1i}(\mathbf{u})}{\partial u_1} & \frac{\partial h_{1i}(\mathbf{u})}{\partial u_2} & \dots & \frac{\partial h_{1i}(\mathbf{u})}{\partial u_n} \\ \frac{\partial h_{2i}(\mathbf{u})}{\partial u_1} & \frac{\partial h_{2i}(\mathbf{u})}{\partial u_2} & \dots & \frac{\partial h_{2i}(\mathbf{u})}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_{ni}(\mathbf{u})}{\partial u_1} & \frac{\partial h_{ni}(\mathbf{u})}{\partial u_2} & \dots & \frac{\partial h_{ni}(\mathbf{u})}{\partial u_n} \end{vmatrix},$$

the determinant of an  $n \times n$  matrix. Assuming that these Jacobians do not vanish identically on  $\mathcal{B}$ , we have the following representation of the joint pdf,  $f_U(u_1, \dots, u_n)$ , for  $\mathbf{u} \in \mathcal{B}$ :

$$(4.6.7) \quad f_U(u_1, \dots, u_n) = \sum_{i=1}^k f_X(h_{1i}(u_1, \dots, u_n), \dots, h_{ni}(u_1, \dots, u_n)) |J_i|.$$

**Example 4.6.4:** Let  $(X_1, X_2, X_3, X_4)$  have joint pdf

$$f_X(x_1, x_2, x_3, x_4) = 24e^{-x_1-x_2-x_3-x_4}, \quad 0 < x_1 < x_2 < x_3 < x_4 < \infty.$$

Consider the transformation

$$U_1 = X_1, \quad U_2 = X_2 - X_1, \quad U_3 = X_3 - X_2, \quad U_4 = X_4 - X_3.$$

This transformation maps the set  $\mathcal{A}$  onto the set  $\mathcal{B} = \{\mathbf{u}: 0 < u_i < \infty, i = 1, 2, 3, 4\}$ . The transformation is one-to-one, so  $k = 1$ , and the inverse is

$$X_1 = U_1, \quad X_2 = U_1 + U_2, \quad X_3 = U_1 + U_2 + U_3, \quad X_4 = U_1 + U_2 + U_3 + U_4.$$

The Jacobian of the inverse is

$$J = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{vmatrix} = 1.$$

Since the matrix is triangular, the determinant is equal to the product of the diagonal elements. Thus, from (4.6.7) we obtain

$$\begin{aligned}f_U(u_1, \dots, u_4) &= 24e^{-u_1-(u_1+u_2)-(u_1+u_2+u_3)-(u_1+u_2+u_3+u_4)} \\&= 24e^{-4u_1-3u_2-2u_3-u_4} \quad \text{on } \mathcal{B}.\end{aligned}$$

From this the marginal pdfs of  $U_1, U_2, U_3$ , and  $U_4$  can be calculated. It turns out that  $f_U(u_i) = (5-i)e^{-(5-i)u_i}, 0 < u_i$ . That is,  $U_i \sim \text{exponential}(1/(5-i))$ . From Theorem 4.6.4 we see that  $U_1, U_2, U_3$ , and  $U_4$  are mutually independent random variables. ||

The model in Example 4.6.4 can arise in the following way. Suppose  $Y_1, Y_2, Y_3$ , and  $Y_4$  are mutually independent random variables, each with an exponential(1) distribution. Define  $X_1 = \min(Y_1, Y_2, Y_3, Y_4)$ ,  $X_2 = \text{second smallest value of } (Y_1, Y_2, Y_3, Y_4)$ ,  $X_3 = \text{second largest value of } (Y_1, Y_2, Y_3, Y_4)$ , and  $X_4 = \max(Y_1, Y_2, Y_3, Y_4)$ . These variables will be called *order statistics* in Section 5.5. There we will see that the joint pdf of  $(X_1, X_2, X_3, X_4)$  is the pdf given in Example 4.6.4. Now the variables  $U_2, U_3$ , and  $U_4$  defined in the example are called the *spacings* between the order statistics. The example showed that, for these exponential random variables  $(Y_1, \dots, Y_n)$ , the spacings between the order statistics are mutually independent and also have exponential distributions.

## 4.7 Inequalities and Identities

Statistical theory is literally brimming with inequalities and identities—so many that entire books are devoted to the topic. The major work by Marshall and Olkin (1979) contains many of the newer inequalities, many using the concept of majorization. The older work by Hardy, Littlewood, and Polya (1952) is a compendium of classic inequalities. In this section we will mix some old and some new, giving some idea of the types of results that exist.

We find that inequalities can be broadly classified into certain categories, and we use these categories as subsections. We have not categorized identities and treat them in their own subsection.

### 4.7.1 Numerical Inequalities

The inequalities in this subsection, although often stated in terms of expectations, rely mainly on properties of numbers. In fact, they are all based on the following simple lemma.

**LEMMA 4.7.1:** Let  $a$  and  $b$  be any positive numbers, and let  $p$  and  $q$  be any positive numbers (necessarily greater than 1) satisfying

$$(4.7.1) \quad \frac{1}{p} + \frac{1}{q} = 1.$$

Then

$$(4.7.2) \quad \frac{1}{p}a^p + \frac{1}{q}b^q \geq ab,$$

with equality if and only if  $a^p = b^q$ .

*Proof:* Fix  $b$ , and consider the function

$$g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab.$$

To minimize  $g(a)$ , differentiate and set equal to 0:

$$\frac{d}{da}g(a) = 0 \Rightarrow a^{p-1} - b = 0 \Rightarrow b = a^{p-1}.$$

A check of the second derivative will establish that this is indeed a minimum. The value of the function at the minimum is

$$\begin{aligned} \frac{1}{p}a^p + \frac{1}{q}(a^{p-1})^q - aa^{p-1} &= \frac{1}{p}a^p + \frac{1}{q}a^p - a^p && \left( \begin{array}{l} (p-1)q = p \text{ follows} \\ \text{from (4.7.1)} \end{array} \right) \\ &= 0. && \text{(again from (4.7.1))} \end{aligned}$$

Hence the minimum is 0 and (4.7.2) is established. Since the minimum is unique (why?), equality holds only if  $a^{p-1} = b$ , which is equivalent to  $a^p = b^q$ , again from (4.7.1).  $\square$

The first of our expectation inequalities, one of the most used and most important, follows easily from the lemma.

**HÖLDER'S INEQUALITY:** Let  $X$  and  $Y$  be any two random variables, and let  $p$  and  $q$  satisfy (4.7.1). Then

$$(4.7.3) \quad |\mathbb{E}XY| \leq \mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q}.$$

*Proof:* The first inequality follows from  $-|XY| \leq XY \leq |XY|$  and Theorem 2.2.1. To prove the second inequality, define

$$a = \frac{|X|}{(\mathbb{E}|X|^p)^{1/p}} \quad \text{and} \quad b = \frac{|Y|}{(\mathbb{E}|Y|^q)^{1/q}}.$$

Applying Lemma 4.7.1, we get

$$\frac{1}{p} \frac{|X|^p}{\mathbb{E}|X|^p} + \frac{1}{q} \frac{|Y|^q}{\mathbb{E}|Y|^q} \geq \frac{|XY|}{(\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q}}.$$

Now take expectations of both sides. The expectation of the left-hand side is 1, and rearrangement gives (4.7.3).  $\square$

Perhaps the most famous special case of Hölder's Inequality is that for which  $p = q = 2$ . This is called the Cauchy–Schwarz Inequality.

**CAUCHY–SCHWARZ INEQUALITY:** For any two random variables  $X$  and  $Y$ ,

$$(4.7.4) \quad |\mathbb{E}XY| \leq \mathbb{E}|XY| \leq (\mathbb{E}|X|^2)^{1/2}(\mathbb{E}|Y|^2)^{1/2}. \quad \square$$

**Example 4.7.1:** If  $X$  and  $Y$  have means  $\mu_X, \mu_Y$  and variances  $\sigma_X^2$  and  $\sigma_Y^2$ , respectively, we can apply the Cauchy–Schwarz Inequality to get

$$\mathbb{E}|(X - \mu_X)(Y - \mu_Y)| \leq \{\mathbb{E}(X - \mu_X)^2\}^{1/2} \{\mathbb{E}(Y - \mu_Y)^2\}^{1/2}.$$

Squaring both sides, and using statistical notation, we have

$$(\text{Cov}(X, Y))^2 \leq \sigma_X^2 \sigma_Y^2.$$

Recalling the definition of the correlation coefficient,  $\rho$ , we have proved that  $0 \leq \rho^2 \leq 1$ . Furthermore, the condition for equality in Lemma 4.7.1 still carries over, and equality is attained here only if  $X - \mu_X = c(Y - \mu_Y)$ , for some constant  $c$ . That is, the correlation is  $\pm 1$  if and only if  $X$  and  $Y$  are linearly related. Compare the ease of this proof to the one used in Theorem 4.5.4, before we had the Cauchy–Schwarz Inequality.  $\parallel$

Some other special cases of Hölder's Inequality are often useful. If we set  $Y \equiv 1$  in (4.7.3) we get

$$(4.7.5) \quad \mathbb{E}|X| \leq \{\mathbb{E}(|X|^p)\}^{1/p}, \quad 1 < p < \infty.$$

For  $1 < r < p$ , if we replace  $|X|$  by  $|X|^r$  in (4.7.5) we obtain

$$\mathbb{E}|X|^r \leq \{\mathbb{E}(|X|^{pr})\}^{1/p}.$$

Now write  $s = pr$  (note that  $s > r$ ) and rearrange terms to get

$$(4.7.6) \quad \{\mathbb{E}|X|^r\}^{1/r} \leq \{\mathbb{E}(|X|^s)\}^{1/s}, \quad 1 < r < s < \infty,$$

which is known as *Liapounov's Inequality*.

Our next named inequality is similar in spirit to Hölder's Inequality and, in fact, follows from it.

**MINKOWSKI'S INEQUALITY:** Let  $X$  and  $Y$  be any two random variables. Then for  $1 \leq p < \infty$ ,

$$(4.7.7) \quad [\mathbb{E}|X+Y|^p]^{1/p} \leq [\mathbb{E}|X|^p]^{1/p} + [\mathbb{E}|Y|^p]^{1/p}.$$

*Proof:* Write

$$(4.7.8) \quad \begin{aligned} \mathbb{E}|X+Y|^p &= \mathbb{E}(|X+Y| |X+Y|^{p-1}) \\ &\leq \mathbb{E}(|X| |X+Y|^{p-1}) + \mathbb{E}(|Y| |X+Y|^{p-1}), \end{aligned}$$

where we have used the fact that  $|X+Y| \leq |X| + |Y|$  (the *triangle inequality*; see Exercise 4.53). Now apply Hölder's Inequality to each expectation on the right-hand side of (4.7.8) to get

$$\begin{aligned} \mathbb{E}(|X+Y|^p) &\leq [\mathbb{E}(|X|^p)]^{1/p} [\mathbb{E}|X+Y|^{q(p-1)}]^{1/q} \\ &\quad + [\mathbb{E}(|Y|^p)]^{1/p} [\mathbb{E}|X+Y|^{q(p-1)}]^{1/q}, \end{aligned}$$

where  $q$  satisfies  $1/p + 1/q = 1$ . Now divide through by  $[\mathbb{E}(|X+Y|^{q(p-1)})]^{1/q}$ . Noting that  $q(p-1) = p$  and  $1 - 1/q = 1/p$ , we obtain (4.7.7).  $\square$

The preceding theorems also apply to numerical sums where there is no explicit reference to an expectation. For example, for numbers  $a_i, b_i, i = 1, \dots, n$ , the inequality

$$(4.7.9) \quad \sum_{i=1}^n |a_i b_i| \leq \left( \sum_{i=1}^n a_i^p \right)^{1/p} \left( \sum_{i=1}^n b_i^q \right)^{1/q}, \quad \frac{1}{p} + \frac{1}{q} = 1,$$

is a version of Hölder's Inequality. To establish (4.7.9) we can formally set up an expectation with respect to random variables taking values  $a_1, \dots, a_n$  and  $b_1, \dots, b_n$ . (This is done in Example 4.7.2.)

An important special case of (4.7.9) occurs when  $b_i \equiv 1, p = q = 2$ . We then have

$$\frac{1}{n} \left( \sum_{i=1}^n |a_i| \right)^2 \leq \sum_{i=1}^n a_i^2.$$

## 4.7.2 Functional Inequalities

The inequalities in this section rely on properties of real-valued functions, rather than on any statistical properties. In many cases, however, they prove to be very useful. One of the most useful is Jensen's Inequality, which applies to convex functions.

DEFINITION 4.7.1: A function  $g(x)$  is *convex* if  $g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y)$  for all  $x$  and  $y$ , and  $0 < \lambda < 1$ . The function  $g(x)$  is *concave* if  $-g(x)$  is convex.

Informally, we can think of convex functions as functions that “hold water”—that is, they are bowl-shaped ( $g(x) = x^2$  is convex), while concave functions “spill water” ( $g(x) = \log x$  is concave). More formally, convex functions lie below lines connecting any two points (see Figure 4.7.1). As  $\lambda$  goes from 0 to 1,  $\lambda g(x_1) + (1 - \lambda)g(x_2)$  defines a line connecting  $g(x_1)$  and  $g(x_2)$ . This line lies above  $g(x)$  if  $g(x)$  is convex. Furthermore, a convex function lies above all of its tangent lines (also shown in Figure 4.7.1), and that fact is the basis of Jensen’s Inequality.

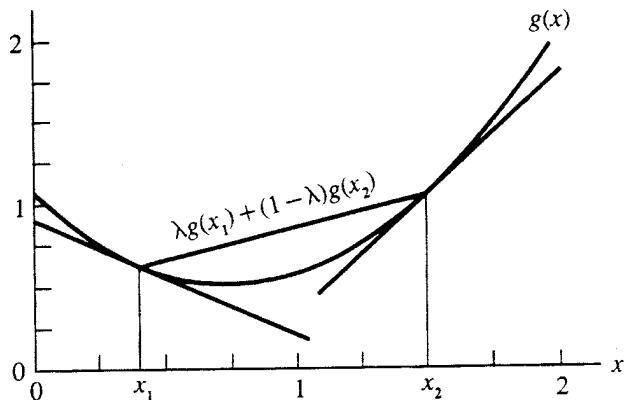


FIGURE 4.7.1 Convex function and tangent lines at  $x_1$  and  $x_2$

*JENSEN’S INEQUALITY:* For any random variable  $X$ , if  $g(x)$  is a convex function then

$$\mathbb{E}g(X) \geq g(\mathbb{E}X).$$

*Proof:* Let  $l(x)$  be a tangent line to  $g(x)$  at the point  $g(\mathbb{E}X)$ . (Recall that  $\mathbb{E}X$  is a constant.) Write  $l(x) = a + bx$  for some  $a$  and  $b$ . The situation is illustrated in Figure 4.7.2.

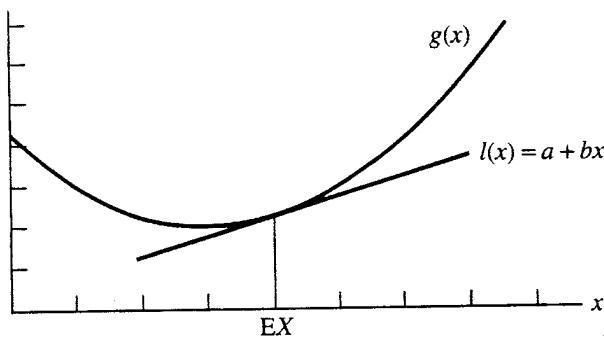


FIGURE 4.7.2 Graphical illustration of Jensen’s Inequality

Now, by the convexity of  $g$  we have  $g(x) \geq a + bx$ . Since expectations preserve inequalities,

$$\begin{aligned}
 \mathbb{E}g(X) &\geq \mathbb{E}(a + bX) \\
 &= a + b\mathbb{E}X && \left( \begin{array}{l} \text{linearity of expectation,} \\ \text{Theorem 2.2.1} \end{array} \right) \\
 &= l(\mathbb{E}X) && (\text{definition of } l(x)) \\
 &= g(\mathbb{E}X) && (l \text{ is tangent at } \mathbb{E}X)
 \end{aligned}$$

as was to be shown.  $\square$

One immediate application of Jensen's Inequality shows that  $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$ , since  $g(x) = x^2$  is convex. Also, if  $x$  is positive, then  $1/x$  is convex, hence  $\mathbb{E}(1/X) \geq 1/\mathbb{E}X$ , another useful application.

To check convexity of a twice differentiable function is quite easy. The function  $g(x)$  is convex if  $g''(x) \geq 0$ , for all  $x$ , and  $g(x)$  is concave if  $g''(x) \leq 0$ , for all  $x$ . Jensen's Inequality applies to concave functions as well. If  $g$  is concave then  $\mathbb{E}g(X) \leq g(\mathbb{E}X)$ .

**Example 4.7.2:** Jensen's Inequality can be used to prove an inequality between three different kinds of means. If  $a_1, \dots, a_n$  are positive numbers, define

$$\begin{aligned}
 a_A &= \frac{1}{n}(a_1 + a_2 + \dots + a_n), && (\text{Arithmetic mean}) \\
 a_G &= [a_1 a_2 \cdots a_n]^{1/n}, && (\text{Geometric mean}) \\
 a_H &= \frac{1}{\frac{1}{a_1} + \frac{1}{a_2} + \dots + \frac{1}{a_n}}. && (\text{Harmonic mean})
 \end{aligned}$$

An inequality relating these means is

$$a_H \leq a_G \leq a_A.$$

To apply Jensen's Inequality, let  $X$  be a random variable with range  $a_1, \dots, a_n$  and  $P(X = a_i) = 1/n$ ,  $i = 1, \dots, n$ . Since  $\log x$  is a concave function, Jensen's Inequality shows that  $\mathbb{E}(\log X) \leq \log(\mathbb{E}X)$ , hence

$$\log a_G = \frac{1}{n} \sum_{i=1}^n \log a_i = \mathbb{E}(\log X) \leq \log(\mathbb{E}X) = \log \left( \frac{1}{n} \sum_{i=1}^n a_i \right) = \log a_A,$$

so  $a_G \leq a_A$ . Now again use the fact that  $\log x$  is concave to get

$$\log \frac{1}{a_H} = \log \left( \frac{1}{n} \sum_{i=1}^n \frac{1}{a_i} \right) = \log \mathbb{E} \frac{1}{X} \geq \mathbb{E} \left( \log \frac{1}{X} \right) = -\mathbb{E}(\log X).$$

Since  $\mathbb{E}(\log X) = \log a_G$ , it then follows that  $\log(1/a_H) \geq \log(1/a_G)$ , or  $a_G \geq a_H$ .  $\parallel$

The next inequality merely exploits the definition of covariance, but sometimes proves to be useful. If  $X$  is a random variable with finite mean  $\mu$  and  $g(x)$  is a nondecreasing function, then

since

$$\begin{aligned}
 & E(g(X)(X - \mu)) \\
 &= E(g(X)(X - \mu)I_{(-\infty, 0)}(X - \mu)) + E(g(X)(X - \mu)I_{[0, \infty)}(X - \mu)) \\
 &\geq E(g(\mu)(X - \mu)I_{(-\infty, 0)}(X - \mu)) \\
 &\quad + E(g(\mu)(X - \mu)I_{[0, \infty)}(X - \mu)) \quad (\text{since } g \text{ is nondecreasing}) \\
 &= g(\mu)E(X - \mu) \\
 &= 0.
 \end{aligned}$$

A generalization of this argument can be used to establish the following inequality (see Exercise 4.58).

**COVARIANCE INEQUALITY:** Let  $X$  be any random variable, and  $g(x)$  and  $h(x)$  any functions such that  $Eg(X)$ ,  $Eh(X)$ , and  $E(g(X)h(X))$  exist.

- a. If  $g(x)$  is a nondecreasing function and  $h(x)$  is a nonincreasing function, then

$$E(g(X)h(X)) \leq (Eg(X))(Eh(X)).$$

- b. If  $g(x)$  and  $h(x)$  are either both nondecreasing or both nonincreasing, then

$$E(g(X)h(X)) \geq (Eg(X))(Eh(X)). \quad \square$$

The intuition behind the inequality is easy. In case (a) there is negative correlation between  $g$  and  $h$ , while in case (b) there is positive correlation. The inequalities merely reflect this fact. The usefulness of the Covariance Inequality is that it allows us to bound an expectation without using higher-order moments.

### 4.7.3 Probability Inequalities

The most famous, and perhaps most useful, probability inequality is Chebychev's Inequality. Its usefulness comes from its wide applicability. As with many important results, its proof is almost trivial.

**CHEBYCHEV'S INEQUALITY:** Let  $X$  be a random variable and let  $g(x)$  be a non-negative function. Then, for any  $r > 0$ ,

$$P(g(X) \geq r) \leq \frac{Eg(X)}{r}.$$

*Proof:*

$$\begin{aligned}
 \mathbb{E}g(X) &= \int_{-\infty}^{\infty} g(x)f_X(x)dx \\
 &\geq \int_{\{x:g(x)\geq r\}} g(x)f_X(x)dx \quad (g \text{ is nonnegative}) \\
 &\geq r \int_{\{x:g(x)\geq r\}} f_X(x)dx \\
 &= rP(g(X) \geq r). \quad (\text{definition})
 \end{aligned}$$

Rearranging now produces the desired inequality.  $\square$

**Example 4.7.3:** The most widespread use of Chebychev's Inequality involves means and variances. Let  $g(x) = (x - \mu)^2/\sigma^2$ , where  $\mu = \mathbb{E}X$  and  $\sigma^2 = \text{Var } X$ . For convenience write  $r = t^2$ . Then

$$P\left(\frac{(X - \mu)^2}{\sigma^2} \geq t^2\right) \leq \frac{1}{t^2} \mathbb{E}\frac{(X - \mu)^2}{\sigma^2} = \frac{1}{t^2}.$$

Doing some obvious algebra, we get the inequality

$$P(|X - \mu| \geq t\sigma) \leq \frac{1}{t^2},$$

and its companion

$$P(|X - \mu| < t\sigma) \geq 1 - \frac{1}{t^2},$$

which gives a universal bound on the deviation  $|X - \mu|$  in terms of  $\sigma$ . For example, taking  $t = 2$ , we get

$$P(|X - \mu| \geq 2\sigma) \leq \frac{1}{2^2} = .25,$$

so there is at least a 75% chance that a random variable will be within  $2\sigma$  of its mean (no matter what the distribution of  $X$ ).  $\parallel$

While Chebychev's Inequality is widely applicable, it is necessarily conservative. (See, for example, Exercise 4.56 and the Miscellanea section.) In particular, we can often get tighter bounds for some specific distributions.

**Example 4.7.4:** If  $Z$  is standard normal, then

$$(4.7.10) \quad P(|Z| \geq t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}, \quad \text{for all } t > 0.$$

Compare this with Chebychev's Inequality. For  $t = 2$ , Chebychev gives  $P(|Z| \geq t) \leq .25$ , but  $\sqrt{(2/\pi)}e^{-2}/2 = .054$ , a vast improvement.

To prove (4.7.10), write

$$\begin{aligned} P(Z \geq t) &= \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2} dx \\ &\leq \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t} e^{-x^2/2} dx \quad \left( \begin{array}{l} \text{since } x/t > 1 \\ \text{for } x > t \end{array} \right) \\ &= \frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t}, \end{aligned}$$

and use the fact that  $P(|Z| \geq t) = 2P(Z \geq t)$ . A lower bound on  $P(|Z| \geq t)$  can be established in a similar way (see Exercise 4.57). ||

Many other probability inequalities exist, and almost all of them are similar in spirit to Chebychev's. For example, we have already seen (Exercise 2.33) that

$$P(X \geq a) \leq e^{-at} M_X(t),$$

but, of course, this inequality requires the existence of the mgf. Other inequalities, tighter than Chebychev but requiring more assumptions, exist. Almost all of them involve bounding a probability by a moment.

#### 4.7.4 Identities

In this section we present a sampling of various identities that can be useful not only in establishing theorems, but also in easing numerical calculations. An entire class of identities can be thought of as "recursion relations," a few of which we have already seen. Recall that if  $X$  is Poisson( $\lambda$ ), then

$$(4.7.11) \quad P(X = x + 1) = \frac{\lambda}{x + 1} P(X = x),$$

allowing us to calculate Poisson probabilities recursively starting from  $P(X = 0) = e^{-\lambda}$ . Relations like (4.7.11) exist for almost all discrete distributions (see Exercise 4.59). Sometimes they exist in a slightly different form for continuous distributions.

**THEOREM 4.7.1:** Let  $X_{\alpha,\beta}$  denote a gamma( $\alpha, \beta$ ) random variable, with pdf  $f(x|\alpha, \beta)$ , where  $\alpha > 1$ . Then for any constants  $a$  and  $b$ ,

$$(4.7.12) \quad P(a < X_{\alpha,\beta} < b) = \beta (f(a|\alpha, \beta) - f(b|\alpha, \beta)) + P(a < X_{\alpha-1,\beta} < b).$$

*Proof:* By definition,

$$\begin{aligned} P(a < X_{\alpha,\beta} < b) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_a^b x^{\alpha-1} e^{-x/\beta} dx \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \left[ -x^{\alpha-1} \beta e^{-x/\beta} \Big|_a^b + \int_a^b (\alpha-1)x^{\alpha-2} \beta e^{-x/\beta} dx \right], \end{aligned}$$

where we have done an integration by parts with  $u = x^{\alpha-1}$  and  $dv = e^{-x/\beta} dx$ . Continuing, we have

$$P(a < X_{\alpha,\beta} < b) = \beta (f(a|\alpha, \beta) - f(b|\alpha, \beta)) + \frac{(\alpha-1)}{\Gamma(\alpha)\beta^{\alpha-1}} \int_a^b x^{\alpha-2} e^{-x/\beta} dx.$$

Using the fact that  $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ , we see that the last term is  $P(a < X_{\alpha-1,\beta} < b)$ .  $\square$

If  $\alpha$  is an integer, repeated use of (4.7.12) will eventually lead to an integral that can be evaluated analytically (when  $\alpha = 1$ , the exponential distribution). Thus, we can easily compute these gamma probabilities.

There is a whole class of identities that rely on integration by parts. The first of these is attributed to Charles Stein, who used it in his work on estimation of multivariate normal means (Stein, 1973, 1981).

**LEMMA 4.7.2 (Stein's Lemma):** Let  $X \sim n(\theta, \sigma^2)$ , and let  $g$  be a differentiable function satisfying  $E|g'(X)| < \infty$ . Then

$$E[g(X)(X - \theta)] = \sigma^2 E g'(X).$$

*Proof:* The left-hand side is

$$E[g(X)(X - \theta)] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} g(x)(x - \theta) e^{-(x-\theta)^2/(2\sigma^2)} dx.$$

Use integration by parts with  $u = g(x)$  and  $dv = (x - \theta)e^{-(x-\theta)^2/(2\sigma^2)} dx$  to get

$$\begin{aligned} E[g(X)(X - \theta)] &= \frac{1}{\sqrt{2\pi}\sigma} \left[ -\sigma^2 g(x) e^{-(x-\theta)^2/(2\sigma^2)} \Big|_{-\infty}^{\infty} + \sigma^2 \int_{-\infty}^{\infty} g'(x) e^{-(x-\theta)^2/(2\sigma^2)} dx \right]. \end{aligned}$$

The condition on  $g'$  is enough to ensure that the first term is 0 and what remains on the right-hand side is  $\sigma^2 E g'(X)$ .  $\square$

**Example 4.7.5:** Stein's Lemma makes calculation of higher-order moments quite easy. For example, if  $X \sim n(\theta, \sigma^2)$ , then

$$\begin{aligned}
EX^3 &= EX^2(X - \theta + \theta) \\
&= EX^2(X - \theta) + \theta EX^2 \\
&= 2\sigma^2 EX + \theta EX^2 \quad (g(x) = x^2, g'(x) = 2x) \\
&= 2\sigma^2\theta + \theta(\sigma^2 + \theta^2) \\
&= 3\theta\sigma^2 + \theta^3. \quad \parallel
\end{aligned}$$

Similar integration-by-parts identities exist for many distributions (see Exercise 4.60 and Hudson (1978)). One can also get useful identities by exploiting properties of a particular distribution, as the next theorem shows.

**THEOREM 4.7.2:** Let  $\chi_p^2$  denote a chi squared random variable with  $p$  degrees of freedom. For any function  $h(x)$ ,

$$(4.7.13) \quad Eh(\chi_p^2) = pE\left(\frac{h(\chi_{p+2}^2)}{\chi_{p+2}^2}\right)$$

provided the expectations exist.

*Proof:* The phrase “provided the expectations exist” is a lazy way of avoiding specification of conditions on  $h$ . In general, reasonable functions will satisfy (4.7.13). We have

$$\begin{aligned}
Eh(\chi_p^2) &= \frac{1}{\Gamma(p/2)2^{p/2}} \int_0^\infty h(x)x^{(p/2)-1}e^{-x/2} dx \\
&= \frac{1}{\Gamma(p/2)2^{p/2}} \int_0^\infty \left(\frac{h(x)}{x}\right) x^{((p+2)/2)-1}e^{-x/2} dx,
\end{aligned}$$

where we have multiplied the integrand by  $x/x$ . Now write

$$\Gamma\left(\frac{p}{2}\right)2^{p/2} = \frac{\Gamma((p+2)/2)2^{(p+2)/2}}{p},$$

so we have

$$\begin{aligned}
Eh(\chi_p^2) &= \frac{p}{\Gamma((p+2)/2)2^{(p+2)/2}} \int_0^\infty \left(\frac{h(x)}{x}\right) x^{((p+2)/2)-1}e^{-x/2} dx, \\
&= pE\left(\frac{h(\chi_{p+2}^2)}{\chi_{p+2}^2}\right). \quad \square
\end{aligned}$$

Some moment calculations are very easy with (4.7.13). For example, the mean of a  $\chi_p^2$  is

$$E\chi_p^2 = pE\left(\frac{\chi_{p+2}^2}{\chi_{p+2}^2}\right) = pE(1) = p,$$

and the second moment is

$$\begin{aligned} E(\chi_p^2)^2 &= pE\left(\frac{(\chi_{p+2}^2)^2}{\chi_{p+2}^2}\right) = pE(\chi_{p+2}^2) \\ &= p(p+2). \end{aligned}$$

So  $\text{Var } \chi_p^2 = p(p+2) - p^2 = 2p$ .

We close our section on identities with some discrete analogs of the previous identities. A general version of the two identities in Theorem 4.7.3 is due to Hwang (1982).

**THEOREM 4.7.3 (Hwang):** Let  $g(x)$  be a function with  $-\infty < Eg(X) < \infty$  and  $-\infty < g(-1) < \infty$ . Then

- a. If  $X \sim \text{Poisson}(\lambda)$ ,

$$(4.7.14) \quad E(\lambda g(X)) = E(Xg(X-1)).$$

- b. If  $X \sim \text{negative binomial}(r, p)$ ,

$$(4.7.15) \quad E((1-p)g(X)) = E\left(\frac{X}{r+X-1}g(X-1)\right).$$

*Proof:* We will prove part (a), saving part (b) for Exercise 4.61. We have

$$\begin{aligned} E(\lambda g(X)) &= \sum_{x=0}^{\infty} \lambda g(x) \frac{e^{-\lambda} \lambda^x}{x!} \\ &= \sum_{x=0}^{\infty} g(x) \frac{e^{-\lambda} \lambda^{x+1}}{x!} \frac{(x+1)}{(x+1)} \\ &= \sum_{x=0}^{\infty} (x+1)g(x) \frac{e^{-\lambda} \lambda^{x+1}}{(x+1)!}. \end{aligned}$$

Now transform the summation index, writing  $y = x + 1$ . As  $x$  goes from 0 to  $\infty$ ,  $y$  goes from 1 to  $\infty$ . Thus

$$E(\lambda g(X)) = \sum_{y=1}^{\infty} yg(y-1) \frac{e^{-\lambda} \lambda^y}{y!}$$

$$\begin{aligned}
 &= \sum_{y=0}^{\infty} yg(y-1) \frac{e^{-\lambda} \lambda^y}{y!} \quad (\text{added term is } 0) \\
 &= E(Xg(X-1)),
 \end{aligned}$$

since this last sum is a Poisson( $\lambda$ ) expectation.  $\square$

Hwang (1982) used his identity in a manner similar to Stein, proving results about multivariate estimators. The identity has other applications, in particular in moment calculations.

**Example 4.7.6:** For  $X \sim \text{Poisson}(\lambda)$ , take  $g(x) = x^2$  and use (4.7.14):

$$\begin{aligned}
 E(\lambda X^2) &= E(X(X-1)^2) \\
 &= E(X^3 - 2X^2 + X).
 \end{aligned}$$

Therefore, the third moment of a Poisson( $\lambda$ ) is

$$\begin{aligned}
 EX^3 &= \lambda EX^2 + 2EX^2 - EX \\
 &= \lambda(\lambda + \lambda^2) + 2(\lambda + \lambda^2) - \lambda \\
 &= \lambda^3 + 3\lambda^2 + \lambda.
 \end{aligned}$$

For the negative binomial, the mean can be calculated by taking  $g(x) = r + x$  in (4.7.15):

$$E((1-p)(r+X)) = E\left(\frac{X}{r+X-1}(r+X-1)\right) = EX,$$

so, rearranging, we get

$$(EX)((1-p)-1) = -r(1-p),$$

or

$$EX = \frac{r(1-p)}{p}.$$

Other moments can be calculated similarly.  $\parallel$

## EXERCISES

- 4.1** A random point  $(X, Y)$  is distributed uniformly on the square with vertices  $(1, 1), (1, -1), (-1, 1), (-1, -1)$ . That is, the joint pdf is  $f(x, y) = \frac{1}{4}$  on the square. Determine the probabilities of the following events.

- a.  $X^2 + Y^2 < 1$    b.  $2X - Y > 0$   
c.  $|X + Y| < 2$

**4.2** Prove the following properties of bivariate expectations (the bivariate analogue to Theorem 2.2.1). For random variables  $X$  and  $Y$ , functions  $g_1(x, y)$  and  $g_2(x, y)$ , and constants  $a$ ,  $b$ , and  $c$ ,

- a.  $E(ag_1(X, Y) + bg_2(X, Y) + c) = aE(g_1(X, Y)) + bE(g_2(X, Y)) + c$   
b. If  $g_1(x, y) \geq 0$  then  $E(g_1(X, Y)) \geq 0$ .  
c. If  $g_1(x, y) \geq g_2(x, y)$  then  $E(g_1(X, Y)) \geq E(g_2(X, Y))$ .  
d. If  $a \leq g_1(x, y) \leq b$  then  $a \leq E(g_1(X, Y)) \leq b$ .

**4.3** Using Definition 4.1.1, show that the random vector  $(X, Y)$  defined at the end of Example 4.1.2 has the pmf given in that example.

**4.4** A pdf is defined by

$$f(x, y) = \begin{cases} C(x + 2y) & \text{if } 0 < y < 1 \text{ and } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

- a. Find the value of  $C$ .  
b. Find the marginal distribution of  $X$ .  
c. Find the joint cdf of  $X$  and  $Y$ .  
d. Find the pdf of the random variable  $Z = 9/(X + 1)^2$ .
- 4.5** Find  $P(X > \sqrt{Y})$  if  $X$  and  $Y$  are jointly distributed with pdf

$$f(x, y) = x + y, \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1.$$

- 4.6** Find  $P(X^2 < Y < X)$  if  $X$  and  $Y$  are jointly distributed with pdf

$$f(x, y) = 2x, \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1.$$

- 4.7** A and B agree to meet at a certain place between 1 P.M. and 2 P.M. Suppose they arrive at the meeting place independently and randomly during the hour. Find the distribution of the length of time that A waits for B. (If B arrives before A, define A's waiting time as 0.)

- 4.8** A woman leaves for work between 8 A.M. and 8:30 A.M., and takes between 40 and 50 minutes to get there. Let the random variable  $X$  denote her time of departure, and the random variable  $Y$  the travel time. Assuming that these variables are independent and uniformly distributed, find the probability that the woman arrives at work before 9 A.M.

- 4.9** Prove that if the joint cdf of  $X$  and  $Y$  satisfies

$$F_{X,Y}(x, y) = F_X(x)F_Y(y),$$

then for any pair of intervals  $(a, b)$ , and  $(c, d)$ ,

$$P(a \leq X \leq b, c \leq Y \leq d) = P(a \leq X \leq b)P(c \leq Y \leq d).$$

- 4.10** The random pair  $(X, Y)$  has the distribution

		X		
		1	2	3
2		$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$
Y	3	$\frac{1}{6}$	0	$\frac{1}{6}$
	4	0	$\frac{1}{3}$	0

- a. Show that  $X$  and  $Y$  are dependent.  
b. Give a probability table for random variables  $U$  and  $V$  that have the same marginals as  $X$  and  $Y$  but are independent.
- 4.11 Let  $U$  = the number of trials needed to get the first head and  $V$  = the number of trials needed to get two heads, in repeated tosses of a fair coin. Are  $U$  and  $V$  independent random variables?
- 4.12 If a stick is broken at random into three pieces, what is the probability that the pieces can be put together in a triangle? (See Gardner (1961) for a complete discussion of this problem.)
- 4.13 Let  $X$  and  $Y$  be random variables with finite means.  
a. Show that

$$\min_{g(x)} E(Y - g(X))^2 = E(Y - E(Y|X))^2,$$

- where  $g(x)$  ranges over all functions. ( $E(Y|X)$  is sometimes called the *regression of  $Y$  on  $X$* , the “best” predictor of  $Y$  conditional on  $X$ .)  
b. Show that equation (2.2.4) can be derived as a special case of part (a).
- 4.14 Suppose  $X$  and  $Y$  are independent  $n(0, 1)$  random variables.  
a. Find  $P(X^2 + Y^2 < 1)$ .  
b. Find  $P(X^2 < 1)$ , after verifying that  $X^2$  is distributed  $\chi_1^2$ .
- 4.15 Let  $X \sim \text{Poisson}(\theta)$ ,  $Y \sim \text{Poisson}(\lambda)$ , independent. It was shown in Theorem 4.3.1 that the distribution of  $X + Y$  is Poisson( $\theta + \lambda$ ). Show that the distribution of  $X|X + Y$  is binomial with success probability  $\theta/(\theta + \lambda)$ . What is the distribution of  $Y|X + Y$ ?
- 4.16 Let  $X$  and  $Y$  be independent random variables with the same geometric distribution.  
a. Show that  $U$  and  $V$  are independent, where  $U$  and  $V$  are defined by

$$U = \min(X, Y) \quad \text{and} \quad V = X - Y.$$

- b. Find the distribution of  $Z = X/(X + Y)$ , where we define  $Z = 0$  if  $X + Y = 0$ .  
c. Find the joint mgf of  $X$  and  $X + Y$ .
- 4.17 Let  $X$  be an exponential(1) random variable, and define  $Y$  to be the integer part of  $X + 1$ , that is

$$Y = i + 1 \quad \text{if and only if} \quad i \leq X < i + 1, \quad i = 0, 1, 2, \dots$$

- a. Find the distribution of  $Y$ . What well-known distribution does  $Y$  have?  
b. Find the conditional distribution of  $X - 4$  given  $Y \geq 5$ .
- 4.18 Given that  $g(x) \geq 0$  has the property that

$$\int_0^\infty g(x) dx = 1,$$

show that

$$f(x, y) = \frac{2g(\sqrt{x^2 + y^2})}{\pi\sqrt{x^2 + y^2}}, \quad x, y > 0,$$

is a pdf.

- 4.19** a. Let  $X_1$  and  $X_2$  be independent  $n(0, 1)$  random variables. Find the pdf of  $(X_1 - X_2)^2/2$ .  
 b. If  $X_i, i = 1, 2$ , are independent  $\text{gamma}(\alpha_i, 1)$  random variables, find the marginal distributions of  $X_1/(X_1 + X_2)$  and  $X_2/(X_1 + X_2)$ .
- 4.20**  $X_1$  and  $X_2$  are independent  $n(0, \sigma^2)$  random variables.  
 a. Find the joint distribution of  $Y_1$  and  $Y_2$ , where

$$Y_1 = X_1^2 + X_2^2 \quad \text{and} \quad Y_2 = \frac{X_1}{\sqrt{Y_1}}.$$

b. Show that  $Y_1$  and  $Y_2$  are independent, and interpret this result geometrically.

- 4.21** A point is generated at random in the plane according to the following polar scheme. A radius  $R$  is chosen, where the distribution of  $R^2$  is  $\chi^2$  with 2 degrees of freedom. Independently, an angle,  $\theta$ , is chosen, where  $\theta \sim \text{uniform}(0, 2\pi)$ . Find the joint distribution of  $X = R \cos \theta$  and  $Y = R \sin \theta$ .
- 4.22** Let  $(X, Y)$  be a bivariate random vector with joint pdf  $f(x, y)$ . Let  $U = aX + b$  and  $V = cY + d$ , where  $a, b, c$ , and  $d$  are fixed constants with  $a > 0$  and  $c > 0$ . Show that the joint pdf of  $(U, V)$  is

$$f_{U,V}(u, v) = \frac{1}{ac} f\left(\frac{u-b}{a}, \frac{v-d}{c}\right).$$

- 4.23** For  $X$  and  $Y$  as in Example 4.3.2, find the distribution of  $XY$  by making the transformations given in (a) and (b), and integrating out  $V$ .  
 a.  $U = XY, V = Y$       b.  $U = XY, V = X/Y$
- 4.24** Let  $X$  and  $Y$  be independent random variables with  $X \sim \text{gamma}(r, 1)$  and  $Y \sim \text{gamma}(s, 1)$ . Show that  $Z_1 = X + Y$  and  $Z_2 = X/(X + Y)$  are independent, and find the distribution of each. ( $Z_1$  is gamma and  $Z_2$  is beta.)
- 4.25** Use the techniques of Section 4.3 to derive the joint distribution of  $(X, Y)$  from the joint distribution of  $(X, Z)$  in Examples 4.5.2 and 4.5.3.
- 4.26**  $X$  and  $Y$  are independent random variables with  $X \sim \text{exponential}(\lambda)$  and  $Y \sim \text{exponential}(\mu)$ . It is impossible to obtain direct observations of  $X$  and  $Y$ . Instead, we observe the random variables  $Z$  and  $W$ , where

$$Z = \min\{X, Y\} \quad \text{and} \quad W = \begin{cases} 1 & \text{if } Z = X \\ 0 & \text{if } Z = Y \end{cases}.$$

(This is a situation that arises, in particular, in medical experiments. The  $X$  and  $Y$  variables are *censored*.)

- a. Find the joint distribution of  $Z$  and  $W$ .  
 b. Prove that  $Z$  and  $W$  are independent. (Hint: Show that  $P(Z \leq z | W = i) = P(Z \leq z)$  for  $i = 0$  or  $1$ .)

- 4.27 Let  $X \sim n(\mu, \sigma^2)$  and let  $Y \sim n(\gamma, \sigma^2)$ . Suppose  $X$  and  $Y$  are independent. Define  $U = X + Y$  and  $V = X - Y$ . Show that  $U$  and  $V$  are independent normal random variables. Find the distribution of each of them.
- 4.28 Let  $X$  and  $Y$  be independent standard normal random variables.
- Show that  $X/(X + Y)$  has a Cauchy distribution.
  - Find the distribution of  $X/|Y|$ .
  - Is the answer to part (b) surprising? Can you formulate a general theorem?
- 4.29 There are many different methods used to generate what are called *pseudo-random variables*, that is, random variables that are generated through the use of computer algorithms. We investigate two methods that generate normal variables from uniforms. (Uniforms are much easier to generate than normals.) Define

$$X_1 = \cos(2\pi U_1) \sqrt{-2 \log(U_2)},$$

$$X_2 = \sin(2\pi U_1) \sqrt{-2 \log(U_2)},$$

where  $U_1$  and  $U_2$  are independent uniform(0, 1) random variables. Prove that  $X_1$  and  $X_2$  are independent  $n(0, 1)$  random variables.

- 4.30 Another method to generate normals from uniforms uses the following algorithm:
- Step 1:* Generate  $U_1$  and  $U_2$ , independent uniform(0, 1).
- Step 2:* Define  $V_1 = 2U_1 - 1$ ,  $V_2 = 2U_2 - 1$ .
- Step 3:* If  $V_1^2 + V_2^2 > 1$ , go back to step 1 and start again. If  $V_1^2 + V_2^2 \leq 1$ , define

$$X_i = V_i \sqrt{\frac{-2 \log(V_1^2 + V_2^2)}{V_1^2 + V_2^2}}, \quad i = 1, 2.$$

- a. Prove that  $X_1$  and  $X_2$  are independent  $n(0, 1)$  random variables.  
b. What is the expected number of uniforms required to generate a normal?
- 4.31 A relatively simple method of generating random variables from an arbitrary bounded pdf is the following. Let  $f(x)$  be any pdf on  $[a, b]$ , and define  $c = \max_{a \leq x \leq b} f(x)$ . Let  $X$  and  $Y$  be independent, with  $X \sim \text{uniform}(a, b)$  and  $Y \sim \text{uniform}(0, c)$ . Let  $d$  be a number greater than  $b$ , and define a new random variable

$$W = \begin{cases} X & \text{if } Y < f(X) \\ d & \text{if } Y \geq f(X) \end{cases}$$

Show that  $P(W \leq w) = \int_a^w f(t)dt/[c(b-a)]$ , for  $a \leq w \leq b$ . Using this, explain how a random variable with pdf  $f(x)$  can be generated from uniform random variables. (*Hint:* Use a geometric argument; a picture will help.)

- 4.32 Suppose the distribution of  $Y$ , conditional on  $X = x$ , is  $n(x, x^2)$  and that the marginal distribution of  $X$  is uniform (0, 1).
- Find  $EY$ ,  $\text{Var}Y$ , and  $\text{Cov}(X, Y)$ .
  - Prove that  $Y/X$  and  $X$  are independent.
- 4.33 Suppose that the random variable  $Y$  has a binomial distribution with  $n$  trials and success probability  $X$ , where  $n$  is a given constant and  $X$  is a uniform(0, 1) random variable.
- Find  $EY$  and  $\text{Var}Y$ .
  - Find the joint distribution of  $X$  and  $Y$ .
  - Find the marginal distribution of  $Y$ .
- 4.34 a. For the hierarchical model

$$Y|\Lambda \sim \text{Poisson}(\Lambda) \quad \text{and} \quad \Lambda \sim \text{gamma}(\alpha, \beta),$$

find the marginal distribution, mean and variance of  $Y$ . Show that the marginal distribution of  $Y$  is negative binomial if  $\alpha$  is an integer.

- b. Show that the three-stage model

$$Y|N \sim \text{binomial}(N, p), \quad N|\Lambda \sim \text{Poisson}(\Lambda), \quad \text{and} \quad \Lambda \sim \text{gamma}(\alpha, \beta),$$

leads to the same marginal (unconditional) distribution of  $Y$ .

- 4.35** (*Alternative derivation of the negative binomial distribution*) Solomon (1983) details the following biological model. Suppose that each of a random number,  $N$ , of insects lays  $X_i$  eggs, where the  $X_i$ s are independent, identically distributed random variables. The total number of eggs laid is  $H = X_1 + \dots + X_N$ . What is the distribution of  $H$ ?

It is common to assume that  $N$  is Poisson( $\lambda$ ). Furthermore, if we assume that each  $X_i$  has the logarithmic series distribution with success probability  $p$  (Exercise 3.12), we have the hierarchical model

$$H|N = X_1 + \dots + X_N, \quad P(X_i = t) = \frac{-1}{\log(p)} \frac{(1-p)^t}{t},$$

$$N \sim \text{Poisson}(\lambda).$$

Show that the marginal distribution of  $H$  is negative binomial  $(r, p)$ , where  $r = -\lambda/\log(p)$ . (It is easiest to calculate and identify the mgf of  $H$  using Theorems 4.4.1 and 4.6.3. Stuart and Ord (1987) also mention this derivation of the logarithmic series distribution. They refer to  $H$  as a *randomly stopped sum*.)

- 4.36** Consider the hierarchy

$$Y|P \sim \text{binomial}(n, P) \quad \text{and} \quad P \sim \text{beta}(\alpha, \beta).$$

- a. Show that the marginal distribution of  $Y$  is given by

$$P(Y = y) = \binom{n}{y} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y + \alpha)\Gamma(n - y + \beta)}{\Gamma(\alpha + \beta + n)}.$$

- b. Find the unconditional mean and variance of  $Y$ .

(The marginal distribution of  $Y$  is the *beta-binomial*  $(n, \alpha, \beta)$  distribution. Example 4.4.3 is a special case.)

- 4.37** (*The gamma as a mixture of exponentials*) Gleser (1989) shows that, in certain cases, the gamma distribution can be written as a scale mixture of exponentials, an identity suggested by different analyses of the same data. Let  $f(x)$  be a gamma( $r, \lambda$ ) pdf.

- a. Show that if  $r \leq 1$  then  $f(x)$  can be written

$$f(x) = \int_0^\lambda \frac{1}{\nu} e^{-x/\nu} p_\lambda(\nu) d\nu,$$

where

$$p_\lambda(\nu) = \frac{1}{\Gamma(r)\Gamma(1-r)} \frac{\nu^{r-1}}{(\lambda - \nu)^r}, \quad 0 < \nu < \lambda.$$

(Hint: Make a change of variable from  $\nu$  to  $u$ , where  $u = x/\nu - x/\lambda$ .)

- b. Show that  $p_\lambda(\nu)$  is a pdf, for  $r \leq 1$ , by showing

$$\int_0^\lambda p_\lambda(\nu) d\nu = 1.$$

- c. Show that the restriction  $r \leq 1$  is necessary for the representation in part (a) to be valid, that is, there is no such representation if  $r > 1$ . (Hint: Suppose  $f(x)$  can be written  $f(x) = \int_0^\infty (e^{-x/\nu}/\nu) q_\lambda(\nu) d\nu$  for some pdf  $q_\lambda(\nu)$ . Show that  $\frac{\partial}{\partial x} \log(f(x)) > 0$  but  $\frac{\partial}{\partial x} \log\left(\int_0^\infty (e^{-x/\nu}/\nu) q_\lambda(\nu) d\nu\right) < 0$ , a contradiction.)

- 4.38 Show that any random variable is uncorrelated with a constant.  
 4.39 Let  $X$  and  $Y$  be independent random variables with means  $\mu_X, \mu_Y$  and variances  $\sigma_X^2, \sigma_Y^2$ . Find an expression for the correlation of  $XY$  and  $Y$  in terms of these means and variances.  
 4.40 Let  $X_1, X_2$ , and  $X_3$  be uncorrelated random variables, each with mean  $\mu$  and variance  $\sigma^2$ . Find, in terms of  $\mu$  and  $\sigma^2$ ,  $\text{Cov}(X_1 + X_2, X_2 + X_3)$  and  $\text{Cov}(X_1 + X_2, X_1 - X_2)$ .  
 4.41 Show that if  $(X, Y) \sim \text{bivariate normal}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ , then the following are true.  
 a. The marginal distribution of  $X$  is  $n(\mu_X, \sigma_X^2)$  and the marginal distribution of  $Y$  is  $n(\mu_Y, \sigma_Y^2)$ .  
 b. The conditional distribution of  $Y$  given  $X = x$  is

$$n(\mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X), \sigma_Y^2(1 - \rho^2)).$$

- c. For any constants  $a$  and  $b$ , the distribution of  $aX + bY$  is

$$n(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y).$$

- 4.42 (A derivation of the bivariate normal distribution) Let  $Z_1$  and  $Z_2$  be independent  $n(0, 1)$  random variables, and define new random variables  $X$  and  $Y$  by

$$X = a_X Z_1 + b_X Z_2 + c_X, \quad Y = a_Y Z_1 + b_Y Z_2 + c_Y,$$

where  $a_X, b_X, c_X, a_Y, b_Y$ , and  $c_Y$  are constants.

- a. Show that

$$\begin{aligned} EX &= c_X, & \text{Var } X &= a_X^2 + b_X^2, \\ EY &= c_Y, & \text{Var } Y &= a_Y^2 + b_Y^2, \end{aligned}$$

$$\text{Cov}(X, Y) = a_X a_Y + b_X b_Y.$$

- b. If we define the constants  $a_X, b_X, c_X, a_Y, b_Y$ , and  $c_Y$  by

$$\begin{aligned} a_X &= \sqrt{\frac{1+\rho}{2}}\sigma_X, & b_X &= \sqrt{\frac{1-\rho}{2}}\sigma_X, & c_X &= \mu_X, \\ a_Y &= \sqrt{\frac{1+\rho}{2}}\sigma_Y, & b_Y &= -\sqrt{\frac{1-\rho}{2}}\sigma_Y, & c_Y &= \mu_Y, \end{aligned}$$

where  $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ , and  $\rho$  are constants,  $-1 \leq \rho \leq 1$ , then show that

$$\begin{aligned} EX &= \mu_X, & \text{Var } X &= \sigma_X^2, \\ EY &= \mu_Y, & \text{Var } Y &= \sigma_Y^2, \\ \rho_{XY} &= \rho. \end{aligned}$$

c. Show that  $(X, Y)$  has the bivariate normal pdf with parameters  $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ , and  $\rho$ .

d. If we start with bivariate normal parameters  $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ , and  $\rho$ , we can define constants  $a_X, b_X, c_X, a_Y, b_Y$ , and  $c_Y$  as the solution to the equations

$$\begin{aligned}\mu_X &= c_X, & \sigma_X^2 &= a_X^2 + b_X^2, \\ \mu_Y &= c_Y, & \sigma_Y^2 &= a_Y^2 + b_Y^2, \\ \rho\sigma_X\sigma_Y &= a_X a_Y + b_X b_Y.\end{aligned}$$

Show that the solution given in part (b) is not unique by exhibiting another solution to these equations. How many solutions are there?

- 4.43** (*Marginal normality does not imply bivariate normality.*) Let  $X$  and  $Y$  be independent  $n(0, 1)$  random variables and define a new random variable  $Z$  by

$$Z = \begin{cases} X & \text{if } XY > 0 \\ -X & \text{if } XY < 0 \end{cases}.$$

a. Show that  $Z$  has a normal distribution.

b. Show that the joint distribution of  $Z$  and  $Y$  is not bivariate normal. (*Hint:* Show that  $Z$  and  $Y$  always have the same sign.)

- 4.44** If  $(X, Y)$  has the bivariate normal pdf

$$f(x, y) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left(\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right),$$

show that  $\rho_{XY} = \rho$  and the correlation of  $X^2$  and  $Y^2$  is  $\rho^2$ . (Conditional expectations will simplify calculations.)

- 4.45** Let  $X$ ,  $Y$ , and  $Z$  be independent uniform(0, 1) random variables.

- a. Find  $P(X/Y \leq t)$  and  $P(XY \leq t)$ . (Pictures will help.)  
b. Find  $P(XY/Z \leq t)$ .

- 4.46** Bullets are fired at the origin of an  $(x, y)$  coordinate system, and the point hit, say  $(X, Y)$ , is a random variable. The variables  $X$  and  $Y$  are taken to be independent  $n(0, 1)$  random variables. If two bullets are fired independently, what is the distribution of the distance between them?

- 4.47** Let  $A$ ,  $B$ , and  $C$  be independent random variables, uniformly distributed on (0, 1). What is the probability that  $Ax^2 + Bx + C$  has real roots? (*Hint:* If  $X \sim \text{uniform}(0, 1)$ ,  $-\log X \sim \text{exponential}$ . The sum of two independent exponentials is gamma.)

- 4.48** Find the pdf of  $\prod_{i=1}^n X_i$ , where the  $X_i$ s are independent uniform(0, 1) random variables. (*Hint:* Try to calculate the cdf, and remember the relationship between uniforms and exponentials.)

- 4.49** A *parallel system* is one that functions as long as at least one component of it functions. A particular parallel system is composed of three independent components, each of which has a lifelength with an exponential( $\lambda$ ) distribution. The lifetime of the system is the maximum of the individual lifelengths. What is the distribution of the lifetime of the system?

- 4.50** A large number,  $N = mk$ , of people are subject to a blood test. This can be administered in two ways.

- i. Each person can be tested separately. In this case  $N$  tests are required.

- ii. The blood samples of  $k$  people can be pooled and analyzed together. If the test is negative, this *one* test suffices for  $k$  people. If the test is positive, each of the  $k$  persons must be tested separately, and, in all,  $k + 1$  tests are required for the  $k$  people.

Assume that the probability,  $p$ , that the test is positive is the same for all people and that the test results for different people are statistically independent.

- What is the probability that the test for a pooled sample of  $k$  people will be positive?
  - Let  $X = \text{number of blood tests necessary under plan (ii).}$  Find  $EX.$
  - In terms of minimizing the expected number of blood tests to be performed on the  $N$  people, which plan [(i) or (ii)] would be preferred if it is known that  $p$  is close to 0? Justify your answer using the expression derived in part (b).
- 4.51** For any three random variables  $X$ ,  $Y$ , and  $Z$ , prove that (as long as the quantities exist)
- $X$  and  $Y - E(Y|X)$  are uncorrelated.
  - $\text{Var}(Y - E(Y|X)) = E(\text{Var}(Y|X))$
  - $\text{Cov}(X, Y) = 0 \Rightarrow E(\text{Cov}(X, Y|Z)) = -\text{Cov}(E(X|Z), E(Y|Z))$
  - $\text{Cov}(Z, E(Y|Z)) = \text{Cov}(Z, Y)$
- 4.52** A random variable  $X$  is defined by  $Z = \log X$ , where  $EZ = 0$ . Is  $EX$  greater than, less than, or equal to 1?
- 4.53** This exercise involves a well-known inequality known as the *triangle inequality* (a special case of Minkowski's Inequality).
- Prove (without using Minkowski's Inequality) that for any numbers  $a$  and  $b$

$$|a + b| \leq |a| + |b|.$$

- b. Use part (a) to establish that for any random variables  $X$  and  $Y$  with finite expectations,

$$E|X + Y| \leq E|X| + E|Y|.$$

- 4.54** For any random variable  $X$  for which  $EX^2$  and  $E|X|$  exist, show that  $P(|X| \geq b)$  does not exceed either  $EX^2/b^2$  or  $E|X|/b$ , where  $b$  is a positive constant. If  $f(x) = e^{-x}$  for  $x > 0$ , show that one bound is better when  $b = 3$  and the other when  $b = \sqrt{2}$ . (Notice Markov's Inequality in the *Miscellanea* section.)
- 4.55** a. If  $X$  is a random variable whose mgf exists, prove that  $P(X \geq 0) \leq Ee^{tX}$ , for all  $t \geq 0$  for which the mgf is defined. (A proof similar to that used for Chebychev's Inequality will work.)
- b. What are general conditions on a function  $h(t, x)$  such that  $P(X \geq 0) \leq Eh(t, X)$ , for all  $t \geq 0$  for which  $Eh(t, X)$  exists. (Note that part (a) is a special case with  $h(t, x) = e^{tx}$ .)
- 4.56** Calculate  $P(|X - \mu_X| \geq k\sigma_X)$  for  $X \sim \text{uniform}(0, 1)$  and  $X \sim \text{exponential}(\lambda)$ , and compare your answers to the bound from Chebychev's Inequality.
- 4.57** If  $Z$  is a standard normal random variable, prove this companion to the inequality in Example 4.7.4:

$$P(|Z| \geq t) \geq \sqrt{\frac{2}{\pi}} \frac{t}{1+t^2} e^{-t^2/2}.$$

- 4.58** Prove the Covariance Inequality by generalizing the argument given in the text immediately preceding the inequality.

- 4.59 Derive recursion relations, similar to the one given in (4.7.11), for the binomial, negative binomial and hypergeometric distributions.
- 4.60 Prove the following analogues to Stein's Lemma, assuming appropriate conditions on the function  $g$ .
- If  $X \sim \text{gamma}(\alpha, \beta)$ , then

$$E(g(X)(X - \alpha\beta)) = \beta E(Xg'(X)).$$

- If  $X \sim \text{beta}(\alpha, \beta)$ , then

$$E\left[g(X)\left(\beta - (\alpha - 1)\frac{(1 - X)}{X}\right)\right] = E((1 - X)g'(X)).$$

- 4.61 Prove the identity for the negative binomial distribution given in Theorem 4.7.3, part (b).

## Miscellanea

---

### *Forced Binary Choice Models*

D. G. Morrison (1978) describes a probability model for forced binary choices. A forced binary choice occurs when a person is forced to choose between two alternatives, as in a taste test. It may be that a person cannot actually discriminate between the two choices (can you tell Coke from Pepsi?), but the set-up of the experiment is such that a choice must be made. Therefore, there is a confounding between discriminating correctly and guessing correctly. Morrison modeled this by defining the following parameters:

$p$  = probability that a person can actually discriminate,

$c$  = probability that a person discriminates correctly.

Then

$$c = p + \frac{1}{2}(1 - p) = \frac{1}{2}(1 + p), \quad \frac{1}{2} < c < 1,$$

where  $\frac{1}{2}(1 - p)$  is the probability that a person guesses correctly. We now run the experiment and observe  $X_1, \dots, X_n \sim \text{Bernoulli}(c)$ , so

$$P(\sum X_i = k | c) = \binom{n}{k} c^k (1 - c)^{n-k}.$$

However, it is probably the case that  $p$  is not constant from person to person, so  $p$  is allowed to vary according to a beta distribution,

$$P \sim \text{beta}(a, b).$$

Morrison discusses how to estimate the parameter of interest here,

$$E(P | \sum X_i = k),$$

which is a difficult problem.

### Chebychev's Inequality

Ghosh and Meeden (1977) discuss the fact that Chebychev's Inequality is very conservative, and is almost never attained. If we write  $\bar{X}_n$  for the mean of the random variables  $X_1, X_2, \dots, X_n$ , then Chebychev's Inequality states

$$P(|\bar{X}_n - \mu| \geq k\sigma) \leq \frac{1}{nk^2}.$$

They prove the following theorem.

*Theorem:* If  $0 < \sigma < \infty$ , then

- a. If  $n = 1$ , the inequality is attainable for  $k \geq 1$  and unattainable for  $0 < k < 1$ .
- b. If  $n = 2$ , the inequality is attainable if and only if  $k = 1$ .
- c. If  $n \geq 3$ , the inequality is not attainable. □

Examples are given for the cases when the inequality is attained. Most of their technical arguments are based on the following inequality, known as Markov's Inequality.

*Lemma (Markov's Inequality):* If  $P(Y \geq 0) = 1$  and  $P(Y = 0) < 1$ , then, for any  $r > 0$ ,

$$P(Y \geq r) \leq \frac{EY}{r},$$

with equality if and only if  $P(Y = r) = p = 1 - P(Y = 0)$ ,  $0 < p \leq 1$ . □

Markov's Inequality can then be applied to the quantity

$$Y = \frac{(\bar{X}_n - \mu)^2}{\sigma^2},$$

to get the above results.

# 5 Properties of a Random Sample

*“Data! Data! Data!” he cried impatiently. “I can’t make bricks without clay.”*

**Sherlock Holmes**  
*The Adventure of the Copper Beeches*

## 5.1 Basic Concepts of Random Samples

Often, the data collected in an experiment consist of several observations on a variable of interest. We discussed examples of this at the beginning of Chapter 4. In this chapter, we present a model for data collection that is often used to describe this situation, a model referred to as random sampling. The following definition explains mathematically what is meant by the random sampling method of data collection.

**DEFINITION 5.1.1:** The random variables  $X_1, \dots, X_n$  are called a *random sample of size n from the population f(x)* if  $X_1, \dots, X_n$  are mutually independent random variables and the marginal pdf or pmf of each  $X_i$  is the same function  $f(x)$ . Alternatively,  $X_1, \dots, X_n$  are called *independent and identically distributed random variables with pdf or pmf f(x)*. This is commonly abbreviated to iid random variables.

The random sampling model describes a type of experimental situation in which the variable of interest has a probability distribution described by  $f(x)$ . If only one observation  $X$  is made on this variable, then probabilities regarding  $X$  can be calculated using  $f(x)$ . In most experiments there are  $n > 1$  (a fixed, positive integer) repeated observations made on the variable, the first observation is  $X_1$ , the second is  $X_2$ , and so on. Under the random sampling model each  $X_i$  is an observation on the same variable and each  $X_i$  has a marginal distribution given by  $f(x)$ . Furthermore, the observations are taken in such a way that the value of one observation has no effect on or relationship with any of the other observations; that is,  $X_1, \dots, X_n$  are *mutually independent*. (See Exercise 5.4 for a generalization of independence.)

From Definition 4.6.2, the joint pdf or pmf of  $X_1, \dots, X_n$  is given by

$$(5.1.1) \quad f(x_1, \dots, x_n) = f(x_1)f(x_2) \cdots f(x_n) = \prod_{i=1}^n f(x_i).$$

This joint pdf or pmf can be used to calculate probabilities involving the sample. Since  $X_1, \dots, X_n$  are identically distributed, all the marginal densities  $f(x)$  are the same function. In particular, if the population pdf or pmf is a member of a parametric

family, say one of those introduced in Chapter 3, with pdf or pmf given by  $f(x|\theta)$ , then the joint pdf or pmf is

$$(5.1.2) \quad f(x_1, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta),$$

where the same parameter value  $\theta$  is used in each of the terms in the product. If, in a statistical setting, we assume that the population we are observing is a member of a specified parametric family but the true parameter value is unknown, then a random sample from this population has a joint pdf or pmf of the above form with the value of  $\theta$  unknown. By considering different possible values of  $\theta$ , we can study how a random sample would behave for different populations.

**Example 5.1.1:** Let  $X_1, \dots, X_n$  be a random sample from an exponential( $\beta$ ) population. Specifically,  $X_1, \dots, X_n$  might correspond to the times until failure (measured in years) for  $n$  identical circuit boards that are put on test and used until they fail. The joint pdf of the sample is

$$f(x_1, \dots, x_n | \beta) = \prod_{i=1}^n f(x_i | \beta) = \prod_{i=1}^n \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-(x_1 + \dots + x_n)/\beta}.$$

This pdf can be used to answer questions about the sample. For example, what is the probability that all the boards last more than 2 years? We can compute

$$\begin{aligned} P(X_1 > 2, \dots, X_n > 2) &= \int_2^\infty \cdots \int_2^\infty \prod_{i=1}^n \frac{1}{\beta} e^{-x_i/\beta} dx_1 \cdots dx_n \\ &= e^{-2/\beta} \int_2^\infty \cdots \int_2^\infty \prod_{i=2}^n \frac{1}{\beta} e^{-x_i/\beta} dx_2 \cdots dx_n && \text{(integrate out } x_1) \\ &\vdots && \text{(integrate out the remaining } x_i\text{'s successively)} \\ &= (e^{-2/\beta})^n \\ &= e^{-2n/\beta}. \end{aligned}$$

If  $\beta$ , the average lifelength of a circuit board, is large, we see that this probability is near 1.

The previous calculation illustrates how the pdf of a random sample defined by (5.1.1) or, more specifically, by (5.1.2), can be used to calculate probabilities about the sample. Realize that the independent and identically distributed property of a random sample can also be used directly in such calculations. For example, the above calculation can be done like this:

$$P(X_1 > 2, \dots, X_n > 2) = P(X_1 > 2) \cdots P(X_n > 2) \quad (\text{independence})$$

$$\begin{aligned}
 &= [P(X_1 > 2)]^n && \text{(identical distributions)} \\
 &= (e^{-2/\beta})^n && \text{(exponential calculation)} \\
 &= e^{-2n/\beta}. && \parallel
 \end{aligned}$$

The random sampling model in Definition 5.1.1 is sometimes called sampling from an *infinite* population. Think of obtaining the values of  $X_1, \dots, X_n$  sequentially. First, the experiment is performed and  $X_1 = x_1$  is observed. Then, the experiment is repeated and  $X_2 = x_2$  is observed. The assumption of independence in random sampling implies that the probability distribution for  $X_2$  is unaffected by the fact that  $X_1 = x_1$  was observed first. "Removing"  $x_1$  from the infinite population does not change the population, so  $X_2 = x_2$  is still a random observation from the same population.

When sampling is from a *finite* population, Definition 5.1.1 may or may not be relevant depending on how the data collection is done. A finite population is a finite set of numbers,  $\{x_1, \dots, x_N\}$ . A sample  $X_1, \dots, X_n$  is to be drawn from this population. Four ways of drawing this sample are described in Section 1.2.3. We will discuss the first two.

Suppose a value is chosen from the population in such a way that each of the  $N$  values is equally likely (probability =  $1/N$ ) to be chosen. (Think of drawing numbers from a hat.) This value is recorded as  $X_1 = x_1$ . Then the process is repeated. Again, each of the  $N$  values is equally likely to be chosen. The second value chosen is recorded as  $X_2 = x_2$ . (If the same number is chosen, then  $x_1 = x_2$ .) This process of drawing from the  $N$  values is repeated  $n$  times, yielding the sample  $X_1, \dots, X_n$ . This kind of sampling is called *with replacement* because the value chosen at any stage is "replaced" in the population and is available for choice again at the next stage. For this kind of sampling, the conditions of Definition 5.1.1 are met. Each  $X_i$  is a discrete random variable that takes on each of the values  $x_1, \dots, x_N$  with equal probability. The random variables  $X_1, \dots, X_n$  are independent because the process of choosing any  $X_i$  is the same, regardless of the values that are chosen for any of the other variables.

A second method for drawing a random sample from a finite population is called sampling *without replacement*. Sampling without replacement is done as follows. A value is chosen from  $\{x_1, \dots, x_N\}$  in such a way that each of the  $N$  values has probability  $1/N$  of being chosen. This value is recorded as  $X_1 = x_1$ . Now a second value is chosen from the remaining  $N - 1$  values. Each of the  $N - 1$  values has probability  $1/(N - 1)$  of being chosen. The second chosen value is recorded as  $X_2 = x_2$ . Choice of the remaining values continues in this way, yielding the sample  $X_1, \dots, X_n$ . But once a value is chosen, it is unavailable for choice at any later stage.

A sample drawn from a finite population without replacement does not satisfy all the conditions of Definition 5.1.1. The random variables  $X_1, \dots, X_n$  are not mutually independent. To see this, let  $x$  and  $y$  be distinct elements of  $\{x_1, \dots, x_N\}$ . Then  $P(X_2 = y | X_1 = x) = 0$ , since the value  $y$  cannot be chosen at the second stage if it was already chosen at the first. However,  $P(X_2 = y | X_1 = x) = 1/(N - 1)$ . The probability distribution for  $X_2$  depends on the value of  $X_1$  that is observed and, hence,  $X_1$  and  $X_2$  are not independent. However, it is interesting to note that  $X_1, \dots, X_n$  are

identically distributed. That is, the marginal distribution of  $X_i$  is the same for each  $i = 1, \dots, n$ . For  $X_1$  it is clear that the marginal distribution is  $P(X_1 = x) = 1/N$  for each  $x \in \{x_1, \dots, x_N\}$ . To compute the marginal distribution for  $X_2$ , use Theorem 1.2.3(a) and the definition of conditional probability to write

$$P(X_2 = x) = \sum_{i=1}^N P(X_2 = x | X_1 = x_i) P(X_1 = x_i).$$

For one value of the index, say  $k$ ,  $x = x_k$  and  $P(X_2 = x | X_1 = x_k) = 0$ . For all other  $j \neq k$ ,  $P(X_2 = x | X_1 = x_j) = 1/(N - 1)$ . Thus,

$$(5.1.3) \quad P(X_2 = x) = (N - 1) \left( \frac{1}{N - 1} \frac{1}{N} \right) = \frac{1}{N}.$$

Similar arguments can be used to show that each of the  $X_i$ 's has the same marginal distribution.

Sampling without replacement from a finite population is sometimes called *simple random sampling*. It is important to realize that this is not the same sampling situation as that described in Definition 5.1.1. However, if the population size  $N$  is large compared to the sample size  $n$ ,  $X_1, \dots, X_n$  are nearly independent and some approximate probability calculations can be made assuming they are independent. By saying they are “nearly independent” we simply mean that the conditional distribution of  $X_i$  given  $X_1, \dots, X_{i-1}$  is not too different from the marginal distribution of  $X_i$ . For example, the conditional distribution of  $X_2$  given  $X_1$  is

$$P(X_2 = x_1 | X_1 = x_1) = 0 \quad \text{and} \quad P(X_2 = x | X_1 = x_1) = \frac{1}{N - 1} \quad \text{for } x \neq x_1.$$

This is not too different from the marginal distribution of  $X_2$  given in (5.1.3) if  $N$  is large. The nonzero probabilities in the conditional distribution of  $X_i$  given  $X_1, \dots, X_{i-1}$  are  $1/(N - i + 1)$ , which are close to  $1/N$  if  $i \leq n$  is small compared with  $N$ .

**Example 5.1.2:** As an example of an approximate calculation using independence, suppose  $\{1, \dots, 1000\}$  is the finite population, so  $N = 1000$ . A sample of size  $n = 10$  is drawn without replacement. What is the probability that all ten sample values are greater than 200? If  $X_1, \dots, X_{10}$  were mutually independent we would have

$$(5.1.4) \quad \begin{aligned} P(X_1 > 200, \dots, X_{10} > 200) &= P(X_1 > 200) \cdots P(X_{10} > 200) \\ &= \left( \frac{800}{1000} \right)^{10} = .107374. \end{aligned}$$

To calculate this probability exactly, let  $Y$  be a random variable that counts the number of items in the sample that are greater than 200. Then  $Y$  has a hypergeometric ( $N = 1,000, M = 800, K = 10$ ) distribution. So

$$P(X_1 > 200, \dots, X_{10} > 200) = P(Y = 10) = \frac{\binom{800}{10} \binom{200}{0}}{\binom{1000}{10}} = .106164.$$

Thus, (5.1.4) is a reasonable approximation to the true value. ||

Throughout the remainder of the book, we will use Definition 5.1.1 as our definition of a random sample from a population.

## 5.2 Sums of Random Variables from a Random Sample

When a sample  $X_1, \dots, X_n$  is drawn, some summary of the values is usually computed. Any well-defined summary may be expressed mathematically as a function  $T(x_1, \dots, x_n)$  whose domain includes the sample space of the random vector  $(X_1, \dots, X_n)$ . The function  $T$  may be real-valued or vector-valued; thus the summary is a random variable (or vector),  $Y = T(X_1, \dots, X_n)$ . This definition of a random variable as a function of others was treated in detail in Chapter 4, and the techniques in Chapter 4 can be used to describe the distribution of  $Y$  in terms of the distribution of the population from which the sample was obtained. Since the random sample  $X_1, \dots, X_n$  has a simple probabilistic structure (because the  $X_i$ 's are independent and identically distributed) the distribution of  $Y$  is particularly tractable. Because this distribution is usually derived from the distribution of the variables in the random sample, it is called the *sampling distribution* of  $Y$ . This distinguishes the probability distribution of  $Y$  from the distribution of the population, that is, the marginal distribution of each  $X_i$ . In this section, we will discuss some properties of sampling distributions, especially for functions  $T(x_1, \dots, x_n)$  defined by sums of random variables.

**DEFINITION 5.2.1:** Let  $X_1, \dots, X_n$  be a random sample of size  $n$  from a population and let  $T(x_1, \dots, x_n)$  be a real-valued or vector-valued function whose domain includes the sample space of  $(X_1, \dots, X_n)$ . Then the random variable or random vector  $Y = T(X_1, \dots, X_n)$  is called a *statistic*. The probability distribution of a statistic  $Y$  is called the *sampling distribution* of  $Y$ .

The definition of a statistic is very broad, with the only restriction being that a statistic cannot be a function of a parameter. The sample summary given by a statistic can include many types of information. For example, it may give the smallest or largest value in the sample, the average sample value, or a measure of the variability in the sample observations. Three statistics that are often used, and provide good summaries of the sample, are now defined.

**DEFINITION 5.2.2:** The *sample mean* is the arithmetic average of the values in a random sample. It is usually denoted by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i.$$

**DEFINITION 5.2.3:** The *sample variance* is the statistic defined by

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

The *sample standard deviation* is the statistic defined by  $S = \sqrt{S^2}$ .

As is commonly done, we have suppressed the functional notation in the above definitions of these statistics. That is, we have written  $S$  rather than  $S(X_1, \dots, X_n)$ . The dependence of the statistic on the sample is understood. As before, we will denote observed values of statistics with lowercase letters. So  $\bar{x}$ ,  $s^2$ , and  $s$  denote observed values of  $\bar{X}$ ,  $S^2$ , and  $S$ .

The sample mean is certainly familiar to all. The sample variance and standard deviation are measures of variability in the sample that are related to the population variance and standard deviation in ways that we shall see below. We begin by deriving some properties of the sample mean and variance. In particular, the relationship for the sample variance given in Theorem 5.2.1 is related to (2.3.1), a similar relationship for the population variance.

**THEOREM 5.2.1:** Let  $x_1, \dots, x_n$  be any numbers and  $\bar{x} = (x_1 + \cdots + x_n)/n$ . Then

- a.  $\min_a \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x})^2$
- b.  $(n-1)s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$

*Proof:* To prove part (a), add and subtract  $\bar{x}$  to get

$$\begin{aligned} \sum_{i=1}^n (x_i - a)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - a)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + 2 \sum_{i=1}^n (x_i - \bar{x})(\bar{x} - a) + \sum_{i=1}^n (\bar{x} - a)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (\bar{x} - a)^2. \quad (\text{cross term is zero}) \end{aligned}$$

It is now clear that the right-hand side is minimized at  $a = \bar{x}$ . (Notice the similarity to Example 2.2.4 and Exercise 4.13.)

To prove part (b), expand the left-hand side and get

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2$$

$$= \sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2.$$

The expression in Theorem 5.2.1(b) is useful both computationally and theoretically because it allows us to express  $s^2$  in terms of sums that are easy to handle.  $\square$

We will begin our study of sampling distributions by considering the expected values of some statistics. The following result is quite useful.

**LEMMA 5.2.1:** Let  $X_1, \dots, X_n$  be a random sample from a population and let  $g(x)$  be a function such that  $Eg(X_1)$  and  $\text{Var } g(X_1)$  exist. Then

$$(5.2.1) \quad E \left( \sum_{i=1}^n g(X_i) \right) = n(Eg(X_1))$$

and

$$(5.2.2) \quad \text{Var} \left( \sum_{i=1}^n g(X_i) \right) = n(\text{Var } g(X_1)).$$

*Proof:* To prove (5.2.1), note that

$$E \left( \sum_{i=1}^n g(X_i) \right) = \sum_{i=1}^n Eg(X_i) = n(Eg(X_1)).$$

Since the  $X_i$ 's are identically distributed, the second equality is true because  $Eg(X_i)$  is the same for all  $i$ . Note that the independence of  $X_1, \dots, X_n$  is not needed for (5.2.1) to hold. Indeed, (5.2.1) is true for any collection of  $n$  identically distributed random variables.

To prove (5.2.2), note that

$$\begin{aligned} \text{Var} \left( \sum_{i=1}^n g(X_i) \right) &= E \left[ \sum_{i=1}^n g(X_i) - E \left( \sum_{i=1}^n g(X_i) \right) \right]^2 && \text{(definition of variance)} \\ &= E \left[ \sum_{i=1}^n (g(X_i) - Eg(X_i))^2 \right]. && \left( \begin{array}{l} \text{expectation property and} \\ \text{rearrangement of terms} \end{array} \right) \end{aligned}$$

In this last expression there are  $n^2$  terms. First, there are  $n$  terms  $(g(X_i) - Eg(X_i))^2$ ,  $i = 1, \dots, n$ , and for each, we have

$$\begin{aligned} E(g(X_i) - Eg(X_i))^2 &= \text{Var } g(X_i) && \text{(definition of variance)} \\ &= \text{Var } g(X_1). && \text{(identically distributed)} \end{aligned}$$

The remaining  $n(n - 1)$  terms are all of the form  $(g(X_i) - Eg(X_i))(g(X_j) - Eg(X_j))$ , with  $i \neq j$ . For each term,

$$\begin{aligned} E[(g(X_i) - Eg(X_i))(g(X_j) - Eg(X_j))] &= \text{Cov}(g(X_i), g(X_j)) \quad \left( \begin{array}{l} \text{definition of} \\ \text{covariance} \end{array} \right) \\ &= 0. \quad (\text{independence, Theorem 4.5.2}) \end{aligned}$$

Thus, we obtain equation (5.2.2).  $\square$

**THEOREM 5.2.2:** Let  $X_1, \dots, X_n$  be a random sample from a population with mean  $\mu$  and variance  $\sigma^2 < \infty$ . Then

- a.  $E\bar{X} = \mu$ ,
- b.  $\text{Var } \bar{X} = \frac{\sigma^2}{n}$ ,
- c.  $ES^2 = \sigma^2$ .

*Proof:* To prove (a), let  $g(X_i) = X_i/n$ , so  $Eg(X_i) = \mu/n$ . Then, by Lemma 5.2.1,

$$E\bar{X} = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} nEX_1 = \mu.$$

Similarly for (b), we have

$$\text{Var } \bar{X} = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} n \text{Var } X_1 = \frac{\sigma^2}{n}.$$

For the sample variance, using Theorem 5.2.1, we have

$$\begin{aligned} ES^2 &= E\left(\frac{1}{n-1} \left[ \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right] \right) = \frac{1}{n-1} \left( nEX_1^2 - nE\bar{X}^2 \right) \\ &= \frac{1}{n-1} \left( n(\sigma^2 + \mu^2) - n \left( \frac{\sigma^2}{n} + \mu^2 \right) \right) = \sigma^2, \end{aligned}$$

establishing part (c) and proving the theorem.  $\square$

The relationships (a) and (c) in Theorem 5.2.2, relationships between a statistic and a population parameter, are examples of *unbiased* statistics. These are discussed in Chapter 7. The statistic  $\bar{X}$  is an *unbiased estimator* of  $\mu$  and  $S^2$  is an *unbiased estimator* of  $\sigma^2$ . The use of  $n - 1$  in the definition of  $S^2$  may have seemed unintuitive. Now we see that, with this definition,  $ES^2 = \sigma^2$ . If  $S^2$  were defined as the usual average of the squared deviations with  $n$  rather than  $n - 1$  in the denominator, then  $ES^2$  would be  $\frac{n-1}{n}\sigma^2$  and  $S^2$  would not be an unbiased estimator of  $\sigma^2$ .

We now discuss in more detail the sampling distribution of  $\bar{X}$ . The methods from Sections 4.3 and 4.6 can be used to derive this sampling distribution from the population distribution. But because of the special probabilistic structure of a random

sample (iid random variables), the resulting sampling distribution of  $\bar{X}$  is simply expressed.

First we note some simple relationships. Since  $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ , if  $f(y)$  is the pdf of  $Y = (X_1 + \dots + X_n)$ , then  $f_{\bar{X}}(x) = nf(nx)$  is the pdf of  $\bar{X}$  (see Exercise 5.5). Thus, a result about the pdf of  $Y$  is easily transformed into a result about the pdf of  $\bar{X}$ . A similar relationship holds for mgfs:

$$M_{\bar{X}}(t) = Ee^{t\bar{X}} = Ee^{t(X_1 + \dots + X_n)/n} = Ee^{(t/n)Y} = M_Y(t/n).$$

Since  $X_1, \dots, X_n$  are identically distributed,  $M_{X_i}(t)$  is the same function for each  $i$ . Thus, by Theorem 4.6.3, we have the following.

**THEOREM 5.2.3:** Let  $X_1, \dots, X_n$  be a random sample from a population with mgf  $M_X(t)$ . Then the mgf of the sample mean is

$$M_{\bar{X}}(t) = [M_X(t/n)]^n. \quad \square$$

Of course, Theorem 5.2.3 is useful only if the expression for  $M_{\bar{X}}(t)$  is a familiar mgf. Cases when this is true are somewhat limited, but the following example illustrates that, when this method works, it provides a very slick derivation of the sampling distribution of  $\bar{X}$ .

**Example 5.2.1:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population. Then the mgf of the sample mean is

$$\begin{aligned} M_{\bar{X}}(t) &= \left[ \exp \left( \mu \frac{t}{n} + \frac{\sigma^2(t/n)^2}{2} \right) \right]^n \\ &= \exp \left( n \left( \mu \frac{t}{n} + \frac{\sigma^2(t/n)^2}{2} \right) \right) = \exp \left( \mu t + \frac{(\sigma^2/n)t^2}{2} \right). \end{aligned}$$

Thus,  $\bar{X}$  has a  $n(\mu, \sigma^2/n)$  distribution.

Another simple example is given by a gamma( $\alpha, \beta$ ) random sample (see Example 4.6.3). Here, we can also easily derive the distribution of the sample mean. The mgf of the sample mean is

$$M_{\bar{X}}(t) = \left[ \left( \frac{1}{1 - \beta(t/n)} \right)^\alpha \right]^n = \left( \frac{1}{1 - (\beta/n)t} \right)^{n\alpha},$$

which we recognize as the mgf of a gamma( $n\alpha, \beta/n$ ), the distribution of  $\bar{X}$ . ||

If Theorem 5.2.3 is not applicable, either because the resulting mgf of  $\bar{X}$  is unrecognizable or the population mgf does not exist, then the transformation method of Sections 4.3 and 4.6 might be used to find the pdf of  $Y = (X_1 + \dots + X_n)$  and  $\bar{X}$ . In such cases, the following *convolution formula* is useful.

**THEOREM 5.2.4:** If  $X$  and  $Y$  are independent continuous random variables with pdfs  $f_X(x)$  and  $f_Y(y)$ , then the pdf of  $Z = X + Y$  is

$$(5.2.3) \quad f_Z(z) = \int_{-\infty}^{\infty} f_X(w)f_Y(z-w) dw.$$

*Proof:* Let  $W = X$ . The Jacobian of the transformation from  $(X, Y)$  to  $(Z, W)$  is 1. So using (4.3.2), we obtain the joint pdf of  $(Z, W)$  as

$$f_{Z,W}(z,w) = f_{X,Y}(w,z-w) = f_X(w)f_Y(z-w).$$

Integrating out  $w$ , we obtain the marginal pdf of  $Z$  as given in (5.2.3).  $\square$

The limits of integration in (5.2.3) might be modified if  $f_X$  or  $f_Y$  or both are positive only for some values. For example, if  $f_X$  and  $f_Y$  are positive only for positive values, then the limits of integration are 0 and  $z$  because the integrand is zero for values of  $w$  outside this range. Equations similar to the convolution formula of (5.2.3) can be derived for operations other than summing; for example, formulas for differences, products, and quotients are also obtainable (see Exercise 5.6).

**Example 5.2.2:** As an example of a situation where the mgf technique fails, consider sampling from a Cauchy distribution. We will eventually derive the distribution of  $\bar{Z}$ , the mean of  $Z_1, \dots, Z_n$ , iid Cauchy(0, 1) observations. We start, however, with the distribution of the sum of two independent Cauchy random variables and apply formula (5.2.3).

Let  $U$  and  $V$  be independent Cauchy random variables,  $U \sim \text{Cauchy}(0, \sigma)$  and  $V \sim \text{Cauchy}(0, \tau)$ , that is,

$$f_U(u) = \frac{1}{\pi\sigma} \frac{1}{1+(u/\sigma)^2}, \quad f_V(v) = \frac{1}{\pi\tau} \frac{1}{1+(v/\tau)^2}, \quad -\infty < u < \infty, \\ -\infty < v < \infty.$$

Based on formula (5.2.3), the pdf of  $Z = U + V$  is given by

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{\pi\sigma} \frac{1}{1+(w/\sigma)^2} \frac{1}{\pi\tau} \frac{1}{1+((z-w)/\tau)^2} dw, \quad -\infty < z < \infty.$$

This integral is somewhat involved, but can be solved by a partial fraction decomposition and some careful antidifferentiation (see Exercise 5.7). The partial fraction decomposition allows us to write the above integrand as

$$\frac{1}{1+(w/\sigma)^2} \frac{1}{1+((z-w)/\tau)^2} \\ = \frac{Aw}{1+(w/\sigma)^2} + \frac{B}{1+(w/\sigma)^2} - \frac{Cw}{1+((z-w)/\tau)^2} - \frac{D}{1+((z-w)/\tau)^2},$$

where  $A, B, C$ , and  $D$  are constants that may depend on  $z$  but not on  $w$ . Now, using the facts that

$$\int \frac{t}{1+t^2} dt = \frac{1}{2} \log(1+t^2) + \text{constant}$$

and

$$\int \frac{1}{1+t^2} dt = \arctan(t) + \text{constant},$$

we can explicitly evaluate the pdf of  $Z$ . The result is

$$f_Z(z) = \frac{1}{\pi(\sigma+\tau)} \frac{1}{1+(z/(\sigma+\tau))^2}, \quad -\infty < z < \infty.$$

(Note that this integration is quite delicate. Since the mean of a Cauchy does not exist, the integrals

$$\int_{-\infty}^{\infty} \frac{Aw}{1+(w/\sigma)^2} dw \quad \text{and} \quad \int_{-\infty}^{\infty} \frac{Cw}{1+((z-w)/\tau)^2} dw$$

do not exist. However, the integral of the difference,

$$\int_{-\infty}^{\infty} \left[ \frac{Aw}{1+(w/\sigma)^2} - \frac{Cw}{1+((z-w)/\tau)^2} \right] dw,$$

*does exist*, which is all that is needed.) Thus, the sum of two independent Cauchy random variables is again a Cauchy, with the scale parameters adding. It therefore follows that if  $Z_1, \dots, Z_n$  are iid Cauchy( $0, 1$ ) random variables, then  $\sum Z_i$  is Cauchy( $0, n$ ) and also  $\bar{Z}$  is Cauchy( $0, 1$ )! The sample mean has the same distribution as the individual observations. ||

If we are sampling from a location-scale family or if we are sampling from certain types of exponential families, the sampling distribution of sums of random variables, and in particular of  $\bar{X}$ , is easy to derive. We will close this section by discussing these two situations.

We first treat the location-scale case discussed in Section 3.4. Suppose  $X_1, \dots, X_n$  is a random sample from  $(1/\sigma)f((x-\mu)/\sigma)$ , a member of a location-scale family. Then the distribution of  $\bar{X}$  has a simple relationship to the distribution of  $\bar{Z}$ , the sample mean from a random sample from the standard pdf  $f(z)$ . To see the nature of this relationship, note that from Theorem 3.4.2 there exist random variables  $Z_1, \dots, Z_n$  such that  $X_i = \sigma Z_i + \mu$  and the pdf of each  $Z_i$  is  $f(z)$ . Furthermore, we see that  $Z_1, \dots, Z_n$  are mutually independent. Thus  $Z_1, \dots, Z_n$  is a random sample from  $f(z)$ . The sample means  $\bar{X}$  and  $\bar{Z}$  are related by

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} \sum_{i=1}^n (\sigma Z_i + \mu) = \frac{1}{n} \left( \sigma \sum_{i=1}^n Z_i + n\mu \right) = \sigma \bar{Z} + \mu.$$

Thus, again applying Theorem 3.4.2, we find that if  $g(z)$  is the pdf of  $\bar{Z}$ , then  $(1/\sigma)g((x - \mu)/\sigma)$  is the pdf of  $\bar{X}$ . It may be easier to work first with  $Z_1, \dots, Z_n$  and  $f(z)$  to find the pdf  $g(z)$  of  $\bar{Z}$ . If this is done, the parameters  $\mu$  and  $\sigma$  do not have to be dealt with, which may make the computations less messy. Then we immediately know that the pdf of  $\bar{X}$  is  $(1/\sigma)g((x - \mu)/\sigma)$ .

In Example 5.2.2, we found that if  $Z_1, \dots, Z_n$  is a random sample from a Cauchy(0, 1) distribution, then  $\bar{Z}$  also has a Cauchy(0, 1) distribution. Now we can conclude that if  $X_1, \dots, X_n$  is a random sample from a Cauchy( $\mu, \sigma$ ) distribution, then  $\bar{X}$  also has a Cauchy( $\mu, \sigma$ ) distribution. It is important to note that the dispersion in the distribution of  $\bar{X}$ , as measured by  $\sigma$ , is the same, regardless of the sample size  $n$ . This is in sharp contrast to the more common situation in Theorem 5.2.2 (the population has finite variance) where  $\text{Var } \bar{X} = \sigma^2/n$  decreases as the sample size increases.

When sampling is from an exponential family, some sums from a random sample have sampling distributions that are easy to derive. The statistics  $T_1, \dots, T_k$  in the next theorem are important summary statistics, as will be seen in Section 6.1.1.

**THEOREM 5.2.5:** Suppose  $X_1, \dots, X_n$  is a random sample from a pdf or pmf  $f(x|\theta)$ , where

$$f(x|\theta) = h(x)c(\theta) \exp\left(\sum_{i=1}^k w_i(\theta)t_i(x)\right)$$

is a member of an exponential family. Define statistics  $T_1, \dots, T_k$  by

$$T_i(X_1, \dots, X_n) = \sum_{j=1}^n t_i(X_j), \quad i = 1, \dots, k.$$

Suppose

$$\{(w_1(\theta), \dots, w_k(\theta)) : \theta \in \Theta\}$$

and

$$\{(T_1(x_1, \dots, x_n), \dots, T_k(x_1, \dots, x_n)) : x_j \in \mathcal{X}\}$$

both contain open subsets of  $\mathbb{R}^k$  where  $\Theta$  is the parameter space and  $\mathcal{X}$  is the sample space of  $X_j$ . Then the distribution of  $(T_1, \dots, T_k)$  is an exponential family of the form

$$(5.2.4) \quad f_T(u_1, \dots, u_k | \theta) = H(u_1, \dots, u_k)[c(\theta)]^n \exp\left(\sum_{i=1}^k w_i(\theta)u_i\right). \quad \square$$

Note that in the pdf or pmf of  $(T_1, \dots, T_k)$ , the functions  $c(\theta)$  and  $w_i(\theta)$  are the same as in the original family although the function  $H(u_1, \dots, u_k)$  is, of course, different from  $h(x)$ . The requirement that the sample space of  $(T_1, \dots, T_k)$  contain an open subset of  $\mathbb{R}^k$  usually is equivalent to the requirement that  $n \geq k$ . We will not prove this theorem but will only illustrate the result in a simple case.

**Example 5.2.3:** Suppose  $X_1, \dots, X_n$  is a random sample from a Bernoulli( $p$ ) distribution. From Example 3.3.1 (with  $n = 1$ ) we see that a Bernoulli( $p$ ) distribution is an exponential family with  $k = 1$ ,  $c(p) = (1 - p)$ ,  $w_1(p) = \log(p/(1 - p))$ , and  $t_1(x) = x$ . Thus, in the previous theorem,  $T_1 = T_1(X_1, \dots, X_n) = X_1 + \dots + X_n$ . From the definition of the binomial distribution in Section 3.1, we know that  $T_1$  has a binomial( $n, p$ ) distribution. From Example 3.3.1 we also see that a binomial( $n, p$ ) distribution is an exponential family with the same  $w_1(p)$  and  $c(p) = (1 - p)^n$ . Thus expression (5.2.4) is verified for this example. ||

## 5.3 Convergence Concepts

This section treats the somewhat fanciful idea of allowing the sample size to approach infinity and investigates the behavior of certain sample quantities as this happens. Although the notion of an infinite sample size is a theoretical artifact, it can often provide us with some useful approximations for the finite-sample case, since it usually happens that expressions become simplified in the limit.

We are mainly concerned with three types of convergence, and treat them in varying amounts of detail. (A full treatment of convergence is given in Feller (1968, 1971) and Chung (1974), for example.) In particular, we want to look at the behavior of  $\bar{X}_n$ , the mean of  $n$  observations, as  $n \rightarrow \infty$ .

### 5.3.1 Convergence in Probability

This type of convergence is one of the weaker types and, hence, is usually quite easy to verify.

**DEFINITION 5.3.1:** A sequence of random variables,  $X_1, X_2, \dots$ , converges in probability to a random variable  $X$  if, for every  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P(|X_n - X| \geq \epsilon) = 0, \quad \text{or, equivalently,} \quad \lim_{n \rightarrow \infty} P(|X_n - X| < \epsilon) = 1.$$

The  $X_1, X_2, \dots$  in Definition 5.3.1 (and the other definitions in this section) are typically not independent and identically distributed random variables, as in a random sample. The distribution of  $X_n$  changes as the subscript changes, and the convergence concepts discussed in this section describe different ways in which the distribution of  $X_n$  converges to some limiting distribution as the subscript becomes large.

Frequently, statisticians are concerned with situations in which the limiting random variable is a constant and the random variables in the sequence are sample means (of some sort). The most famous result of this type is the following.

**THEOREM 5.3.1 (Weak Law of Large Numbers):** Let  $X_1, X_2, \dots$  be iid random variables with  $\text{E}X_i = \mu$  and  $\text{Var } X_i = \sigma^2 < \infty$ . Define  $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$ . Then, for every  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < \epsilon) = 1,$$

that is,  $\bar{X}_n$  converges in probability to  $\mu$ .

*Proof:* The proof is quite simple, being a straightforward application of Chebychev's Inequality. We have, for every  $\epsilon > 0$ ,

$$P(|\bar{X}_n - \mu| \geq \epsilon) = P((\bar{X}_n - \mu)^2 \geq \epsilon^2) \leq \frac{\text{E}(\bar{X}_n - \mu)^2}{\epsilon^2} = \frac{\text{Var } \bar{X}}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

Hence,  $P(|\bar{X}_n - \mu| < \epsilon) = 1 - P(|\bar{X}_n - \mu| \geq \epsilon) \geq 1 - \sigma^2/(n\epsilon^2) \rightarrow 1$ , as  $n \rightarrow \infty$ .  $\square$

The Weak Law of Large Numbers (WLLN) quite elegantly states that, under general conditions, the sample mean approaches the population mean as  $n \rightarrow \infty$ . In fact, there are more general versions of the WLLN, where we need assume only that the mean is finite. However, the version stated in Theorem 5.3.1 is applicable in most practical situations. (See Exercise 5.13 for one way of weakening the hypotheses of the WLLN.)

The property summarized by the WLLN, that a sequence of the "same" sample quantity approaches a constant as  $n \rightarrow \infty$ , is known as *consistency*. We will examine this property more closely in Chapter 7.

**Example 5.3.1:** Suppose we have a sequence  $X_1, X_2, \dots$  of iid random variables with  $\text{E}X_i = \mu$  and  $\text{Var } X_i = \sigma^2 < \infty$ . If we define

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

can we prove a WLLN for  $S_n^2$ ? Using Chebychev's Inequality, we have

$$P(|S_n^2 - \sigma^2| \geq \epsilon) \leq \frac{\text{E}(S_n^2 - \sigma^2)^2}{\epsilon^2} = \frac{\text{Var } S_n^2}{\epsilon^2}$$

and thus, a sufficient condition that  $S_n^2$  converges in probability to  $\sigma^2$  is that  $\text{Var } S_n^2 \rightarrow 0$  as  $n \rightarrow \infty$ .  $\parallel$

### 5.3.2 Almost Sure Convergence

A type of convergence that is stronger than convergence in probability is almost sure convergence (sometimes confusingly known as *convergence with probability 1*). This type of convergence is similar to pointwise convergence of a sequence of functions,

except that the convergence need not occur on a set with probability 0 (hence the “almost” sure).

**DEFINITION 5.3.2:** A sequence of random variables,  $X_1, X_2, \dots$ , converges almost surely to a random variable  $X$  if, for every  $\epsilon > 0$ ,

$$P(\lim_{n \rightarrow \infty} |X_n - X| < \epsilon) = 1.$$

Notice the similarity in the statements of Definitions 5.3.1 and 5.3.2. Although they look similar, they are very different statements with Definition 5.3.2 much stronger. To understand almost sure convergence, we must recall the basic definition of a random variable as given in Definition 1.4.1. A random variable is a real-valued function defined on a sample space  $S$ . If a sample space  $S$  has elements denoted by  $s$ , then  $X_n(s)$  and  $X(s)$  are all functions defined on  $S$ . Definition 5.3.2 states that  $X_n$  converges to  $X$  almost surely if the functions  $X_n(s)$  converge to  $X(s)$  for all  $s \in S$  except perhaps for  $s \in N$  where  $N \subset S$  and  $P(N) = 0$ . Example 5.3.2 illustrates almost sure convergence. Example 5.3.3 illustrates the difference between convergence in probability and almost sure convergence.

**Example 5.3.2:** Let the sample space  $S$  be the closed interval  $[0, 1]$  with the uniform probability distribution. Define random variables  $X_n(s) = s + s^n$  and  $X(s) = s$ . For every  $s \in [0, 1]$ ,  $s^n \rightarrow 0$  as  $n \rightarrow \infty$  and  $X_n(s) \rightarrow s = X(s)$ . However,  $X_n(1) = 2$  for every  $n$  so  $X_n(1)$  does not converge to  $1 = X(1)$ . But since the convergence occurs on the set  $[0, 1]$  and  $P([0, 1]) = 1$ ,  $X_n$  converges to  $X$  almost surely. ||

**Example 5.3.3:** In this example we describe a sequence that converges in probability, but not almost surely. Again, let the sample space  $S$  be the closed interval  $[0, 1]$  with the uniform probability distribution. Define the sequence  $X_1, X_2, \dots$  as follows:

$$\begin{aligned} X_1(s) &= s + I_{[0,1]}(s), & X_2(s) &= s + I_{[0,\frac{1}{2}]}(s), & X_4(s) &= s + I_{[0,\frac{1}{3}]}(s), \\ X_3(s) &= s + I_{[\frac{1}{2},1]}(s), & X_5(s) &= s + I_{[\frac{1}{3},\frac{2}{3}]}(s), & \\ X_6(s) &= s + I_{[\frac{2}{3},1]}(s), & & & \end{aligned}$$

etc. Let  $X(s) = s$ . It is straightforward to see that  $X_n$  converges to  $X$  in probability. As  $n \rightarrow \infty$ ,  $P(|X_n - X| \geq \epsilon)$  is equal to the probability of an interval of  $s$  values whose length is going to 0. However,  $X_n$  does not converge to  $X$  almost surely. Indeed, there is no value of  $s \in S$  for which  $X_n(s) \rightarrow s = X(s)$ . For every  $s$ , the value  $X_n(s)$  alternates between the values  $s$  and  $s + 1$  infinitely often. For example, if  $s = \frac{3}{8}$ ,  $X_1(s) = 1\frac{3}{8}$ ,  $X_2(s) = 1\frac{3}{8}$ ,  $X_3(s) = \frac{3}{8}$ ,  $X_4(s) = \frac{3}{8}$ ,  $X_5(s) = 1\frac{3}{8}$ ,  $X_6(s) = \frac{3}{8}$ , etc. No pointwise convergence occurs for this sequence. ||

As might be guessed, convergence almost surely, being the stronger criterion, implies convergence in probability. The converse is, of course, false, as Example 5.3.3 shows. However, if a sequence converges in probability, it is possible to find a subsequence that converges almost surely (see Chung (1974)) for theorems or Exercise 5.14 for an example).

Again, statisticians are often concerned with convergence to a constant. We now state, without proof, the stronger analogue of the WLLN, the Strong Law of Large Numbers (SLLN).

**THEOREM 5.3.2 (Strong Law of Large Numbers):** Let  $X_1, X_2, \dots$  be iid random variables with  $EX_i = \mu$  and  $\text{Var } X_i = \sigma^2 < \infty$ , and define  $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$ . Then, for every  $\epsilon > 0$ ,

$$P(\lim_{n \rightarrow \infty} |\bar{X}_n - \mu| < \epsilon) = 1,$$

that is,  $\bar{X}_n$  converges almost surely to  $\mu$ . □

### 5.3.3 Convergence in Distribution

We have already encountered the idea of convergence in distribution in Chapter 2. Remember the properties of moment generating functions (mgfs) and how their convergence implies convergence in distribution (Theorem 2.3.4).

**DEFINITION 5.3.3:** A sequence of random variables,  $X_1, X_2, \dots$ , converges in distribution to a random variable  $X$  if

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x),$$

at all points  $x$  where  $F_X(x)$  is continuous.

Note that although we talk of a sequence of random variables converging in distribution, it is really the cdfs that converge, not the random variables. In this very fundamental way convergence in distribution is quite different from convergence in probability or convergence almost surely.

We again want to look at the large-sample behavior of the sample mean and, in particular, investigate its limiting distribution. We begin by proving one of the most startling theorems in statistics, the Central Limit Theorem (CLT).

**THEOREM 5.3.3 (Central Limit Theorem):** Let  $X_1, X_2, \dots$  be a sequence of iid random variables whose mgfs exist in a neighborhood of 0 (that is,  $M_{X_i}(t)$  exists for  $|t| < h$ , for some positive  $h$ ). Let  $EX_i = \mu$  and  $\text{Var } X_i = \sigma^2 > 0$ . (Both  $\mu$  and  $\sigma^2$  are finite since the mgf exists.) Define  $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$ . Let  $G_n(x)$  denote the cdf of  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ . Then, for any  $x$ ,  $-\infty < x < \infty$ ,

$$\lim_{n \rightarrow \infty} G_n(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy,$$

that is,  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$  has a limiting standard normal distribution. □

Before we prove this theorem (the proof is somewhat anticlimactic) we first look at its implications. Starting from virtually no assumptions (other than independence

and finite variances), we end up with normality! The point here is that normality comes from sums of “small” (finite variance), independent disturbances. The assumption of finite variances is essentially necessary for convergence to normality. Although it can be relaxed somewhat, it cannot be eliminated. (Recall Example 5.2.2, dealing with the Cauchy distribution, where there is no convergence to normality.)

While reveling in the wonder of the CLT, it is also useful to reflect on its limitations. Although it gives us a useful general approximation, we have no way of knowing how good this approximation is. In fact, the goodness of the approximation is a function of the original distribution, and so must be checked case by case. Furthermore, with the current availability of cheap, plentiful computing power, the importance of approximations like the Central Limit Theorem is somewhat lessened. However, despite its limitations, it is still a marvelous result.

*Proof of Theorem 5.3.3:* We will show that, for  $|t| < h$ , the mgf of  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$  converges to  $e^{t^2/2}$ , the mgf of a  $n(0, 1)$  random variable.

Define  $Y_i = (X_i - \mu)/\sigma$ , and let  $M_Y(t)$  denote the common mgf of the  $Y_i$ s, which exists for  $|t| < \sigma h$  and is given by Theorem 2.3.5. Since

$$(5.3.1) \quad \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i,$$

we have, from the properties of mgfs (see Theorems 2.3.5 and 4.6.3)

$$\begin{aligned} (5.3.2) \quad M_{\sqrt{n}(\bar{X}_n - \mu)/\sigma}(t) &= M_{\sum_{i=1}^n Y_i / \sqrt{n}}(t) \\ &= M_{\sum_{i=1}^n Y_i} \left( \frac{t}{\sqrt{n}} \right) \quad (\text{Theorem 2.3.5}) \\ &= \left( M_Y \left( \frac{t}{\sqrt{n}} \right) \right)^n. \quad (\text{Theorem 4.6.3}) \end{aligned}$$

We now expand  $M_Y(t/\sqrt{n})$  in a Taylor series (power series) around 0. (See Definition 7.4.2.) We have

$$(5.3.3) \quad M_Y \left( \frac{t}{\sqrt{n}} \right) = \sum_{k=0}^{\infty} M_Y^{(k)}(0) \frac{(t/\sqrt{n})^k}{k!},$$

where  $M_Y^{(k)}(0) = (d^k/dt^k) M_Y(t)|_{t=0}$ . Since the mgfs exist for  $|t| < h$ , the power series expansion is valid if  $t < \sqrt{n}\sigma h$ .

Using the facts that  $M_Y^{(0)} = 1$ ,  $M_Y^{(1)} = 0$ , and  $M_Y^{(2)} = 1$  (by construction, the mean and variance of  $Y$  are 0 and 1), we have

$$(5.3.4) \quad M_Y \left( \frac{t}{\sqrt{n}} \right) = 1 + \frac{(t/\sqrt{n})^2}{2!} + R_Y \left( \frac{t}{\sqrt{n}} \right),$$

where  $R_Y$  is the remainder term in the Taylor expansion,

$$R_Y\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=3}^{\infty} M_Y^{(k)}(0) \frac{(t/\sqrt{n})^k}{k!}.$$

An application of Taylor's Theorem (Theorem 7.4.1) shows that, for fixed  $t \neq 0$ , we have

$$\lim_{n \rightarrow \infty} \frac{R_Y(t/\sqrt{n})}{(t/\sqrt{n})^2} = 0.$$

Since  $t$  is fixed, we also have

$$(5.3.5) \quad \lim_{n \rightarrow \infty} \frac{R_Y(t/\sqrt{n})}{(1/\sqrt{n})^2} = \lim_{n \rightarrow \infty} n R_Y\left(\frac{t}{\sqrt{n}}\right) = 0,$$

and (5.3.5) is also true at  $t = 0$  since  $R_Y(0/\sqrt{n}) = 0$ . Thus, for any fixed  $t$ , we can write

$$(5.3.6) \quad \begin{aligned} \lim_{n \rightarrow \infty} \left( M_Y\left(\frac{t}{\sqrt{n}}\right) \right)^n &= \lim_{n \rightarrow \infty} \left[ 1 + \frac{(t/\sqrt{n})^2}{2!} + R_Y\left(\frac{t}{\sqrt{n}}\right) \right]^n \\ &= \lim_{n \rightarrow \infty} \left[ 1 + \frac{1}{n} \left( \frac{t^2}{2} + n R_Y\left(\frac{t}{\sqrt{n}}\right) \right) \right]^n \\ &= e^{t^2/2}, \end{aligned}$$

by an application of Lemma 2.3.1, where we set  $a_n = (t^2/2) + n R_Y(t/\sqrt{n})$ . (Note that (5.3.5) implies that  $a_n \rightarrow t^2/2$  as  $n \rightarrow \infty$ .) Since  $e^{t^2/2}$  is the mgf of the  $\text{n}(0, 1)$  distribution, the theorem is proved.  $\square$

The Central Limit Theorem is valid in much more generality than is stated in Theorem 5.3.3 (see the *Miscellanea* section for a discussion). In particular, all of the assumptions about mgfs are not needed—the use of characteristic functions (Chapter 2 *Miscellanea*) can replace them. We state the next theorem without proof. It is a version of the Central Limit Theorem that is general enough for almost all statistical purposes. Notice that the only assumption on the parent distribution is that it has finite variance.

**THEOREM 5.3.4 (Stronger Form of the Central Limit Theorem):** Let  $X_1, X_2, \dots$  be a sequence of iid random variables with  $EX_i = \mu$  and  $0 < \text{Var } X_i = \sigma^2 < \infty$ . Define  $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$ . Let  $G_n(x)$  denote the cdf of  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ . Then, for any  $x$ ,  $-\infty < x < \infty$ ,

$$\lim_{n \rightarrow \infty} G_n(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy,$$

that is,  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$  has a limiting standard normal distribution.  $\square$

The proof is almost identical to that of Theorem 5.3.3, except that characteristic functions are used instead of mgfs. Since the characteristic function of a distribution always exists, it is not necessary to mention them in the assumptions of the theorem. The proof is more delicate, however, since functions of *complex variables* must be dealt with. Details can be found in Chung (1974) or Feller (1971).

It is also possible to prove this theorem without recourse to characteristic functions, using only elementary arguments. By doing careful analysis and being clever with Taylor series expansions, it can be shown directly that probabilities involving  $\sqrt{n}(\bar{X}_n - \mu)/\sigma$  converge to normal probabilities (Brown, 1988). (Brown's proof is similar in spirit to, but much more involved than, the proof of the *Demoivre–Laplace Limit Theorem* given in Feller (1968). The Demoivre–Laplace Limit Theorem is a special case of the CLT, that binomials converge to normals as  $n \rightarrow \infty$ .)

The Central Limit Theorem provides us with an all-purpose approximation (but, remember the warning about the goodness of the approximation). In practice, it can always be used for a first, rough calculation.

**Example 5.3.4:** Suppose  $X_1, \dots, X_n$  are a random sample from a negative binomial( $r, p$ ) distribution. Recall that

$$EX = \frac{r(1-p)}{p}, \quad \text{Var } X = \frac{r(1-p)}{p^2},$$

and the Central Limit Theorem tells us that

$$\frac{\sqrt{n}(\bar{X} - r(1-p)/p)}{\sqrt{r(1-p)/p^2}}$$

is approximately  $n(0, 1)$ . The approximate probability calculations are much easier than the exact calculations. For example, if  $r = 10$ ,  $p = \frac{1}{2}$ , and  $n = 30$ , an exact calculation would be

$$\begin{aligned} P(\bar{X} \leq 11) &= P\left(\sum_{i=1}^{30} X_i \leq 330\right) \\ &= \sum_{x=0}^{330} \binom{300+x-1}{x} \left(\frac{1}{2}\right)^{300} \left(\frac{1}{2}\right)^x \quad \left(\begin{array}{l} \text{$\sum X$ is negative} \\ \text{binomial}(nr, p) \end{array}\right) \\ &= .8916, \end{aligned}$$

which is a very difficult calculation. (Note that this calculation is difficult even with the aid of a computer—the magnitudes of the factorials cause great difficulty. Try it if you don't believe it!) The CLT gives us the approximation

$$P(\bar{X} \leq 11) = P\left(\frac{\sqrt{30}(\bar{X} - 10)}{\sqrt{20}} \leq \frac{\sqrt{30}(11 - 10)}{\sqrt{20}}\right)$$

$$\approx P(Z \leq 1.2247) \\ = .8888.$$

||

An approximation tool that can be used in conjunction with the Central Limit Theorem is known as Slutsky's Theorem.

**THEOREM 5.3.5 (Slutsky's Theorem):** If  $X_n \rightarrow X$  in distribution and  $Y_n \rightarrow a$ , a constant, in probability, then

- a.  $Y_n X_n \rightarrow aX$  in distribution.
- b.  $X_n + Y_n \rightarrow X + a$  in distribution.

□

The proof of Slutsky's Theorem is omitted, since it relies on a characterization of convergence in distribution that we have not discussed. A typical application is illustrated by the following example.

**Example 5.3.5:** Suppose that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightarrow n(0, 1)$$

but the value of  $\sigma$  is unknown. We have seen in Example 5.3.1 that, if  $\lim_{n \rightarrow \infty} \text{Var } S_n^2 = 0$ , then  $S_n^2 \rightarrow \sigma^2$  in probability. By Exercise 5.15,  $\sigma/S_n \rightarrow 1$  in probability. Hence, Slutsky's Theorem tells us

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sigma}{S_n} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightarrow n(0, 1). \quad ||$$

## 5.4 Sampling from the Normal Distribution

This section deals with the properties of sample quantities drawn from a normal population—still one of the most widely used statistical models. Sampling from a normal population leads to many useful properties of sample statistics, and also to many well-known sampling distributions.

### 5.4.1 Properties of the Sample Mean and Variance

We have already seen how to calculate the means and variances of  $\bar{X}$  and  $S^2$  in general. Now, under the additional assumption of normality, we can derive their full distributions, and more. The properties of  $\bar{X}$  and  $S^2$  are summarized in the following theorem.

**THEOREM 5.4.1:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  distribution, and let  $\bar{X} = (1/n)\sum_{i=1}^n X_i$  and  $S^2 = [1/(n-1)]\sum_{i=1}^n (X_i - \bar{X})^2$ . Then

- a.  $\bar{X}$  and  $S^2$  are independent random variables,

- b.  $\bar{X}$  has a  $n(\mu, \sigma^2/n)$  distribution,
- c.  $(n-1)S^2/\sigma^2$  has a chi squared distribution with  $n-1$  degrees of freedom.

*Proof:* First note that, from Section 3.4 on location-scale families, we can assume, without loss of generality, that  $\mu = 0$  and  $\sigma = 1$ . (Also see the discussion preceding Theorem 5.2.5.) Furthermore, part (b) has already been established in Example 5.2.1, leaving us to prove parts (a) and (c).

To prove part (a) we will apply Theorem 4.6.5, and show that  $\bar{X}$  and  $S^2$  are functions of independent random vectors. Note that we can write  $S^2$  as a function of  $n-1$  deviations as follows:

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{n-1} \left( (X_1 - \bar{X})^2 + \sum_{i=2}^n (X_i - \bar{X})^2 \right) \\ &= \frac{1}{n-1} \left( \left[ \sum_{i=2}^n (X_i - \bar{X}) \right]^2 + \sum_{i=2}^n (X_i - \bar{X})^2 \right). \quad \left( \sum_{i=1}^n (X_i - \bar{X}) = 0 \text{ since } \sum_{i=1}^n (X_i - \bar{X}) = 0 \right) \end{aligned}$$

Thus,  $S^2$  can be written as a function only of  $(X_2 - \bar{X}, \dots, X_n - \bar{X})$ . We will now show that these random variables are independent of  $\bar{X}$ . The joint pdf of the sample  $X_1, \dots, X_n$  is given by

$$f(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-(1/2)\sum_{i=1}^n x_i^2}, \quad -\infty < x_i < \infty.$$

Make the transformation

$$y_1 = \bar{x},$$

$$y_2 = x_2 - \bar{x},$$

$$\vdots$$

$$y_n = x_n - \bar{x}.$$

This is a linear transformation with a Jacobian equal to  $n$ . We have

$$\begin{aligned} f(y_1, \dots, y_n) &= \frac{n}{(2\pi)^{n/2}} e^{-(1/2)(y_1 - \sum_{i=2}^n y_i)^2} e^{-(1/2)\sum_{i=2}^n (y_i + y_1)^2}, \quad -\infty < y_i < \infty \\ &= \left[ \left( \frac{n}{2\pi} \right)^{1/2} e^{(-ny_1^2)/2} \right] \left[ \frac{n^{1/2}}{(2\pi)^{(n-1)/2}} e^{-(1/2)[\sum_{i=2}^n y_i^2 + (\sum_{i=2}^n y_i)^2]} \right], \quad -\infty < y_i < \infty. \end{aligned}$$

Since the joint pdf of  $Y_1, \dots, Y_n$  factors, it follows from Theorem 4.6.4 that  $Y_1$  is independent of  $Y_2, \dots, Y_n$ , and hence, from Theorem 4.6.5, that  $\bar{X}$  is independent of  $S^2$ .

To finish the proof of the theorem we must now derive the distribution of  $S^2$ . Before doing so, however, we digress a little and discuss the chi squared distribution, whose properties play an important part in the derivation of the pdf of  $S^2$ . Recall from Section 3.2 that the chi squared pdf is a special case of the gamma pdf, and is given by

$$f(x) = \frac{1}{\Gamma(p/2) 2^{p/2}} x^{(p/2)-1} e^{-x/2}, \quad 0 < x < \infty,$$

where  $p$  is called the *degrees of freedom*. We now summarize some pertinent facts about the chi squared distribution.

**LEMMA 5.4.1 (Facts About Chi Squared Random Variables):** We use the notation  $\chi_p^2$  to denote a chi squared random variable with  $p$  degrees of freedom.

- a. If  $Z$  is a  $n(0, 1)$  random variable, then  $Z^2 \sim \chi_1^2$ , that is, the square of a standard normal random variable is a chi squared random variable.
- b. If  $X_1, \dots, X_n$  are independent, and  $X_i \sim \chi_{p_i}^2$ , then  $X_1 + \dots + X_n \sim \chi_{p_1+\dots+p_n}^2$ , that is, independent chi squared variables add to a chi squared variable, and the degrees of freedom also add.

*Proof of Lemma:* We have encountered these facts already. Part (a) was established in Example 2.1.5. Part (b) is a special case of Example 4.6.3, which has to do with sums of independent gamma random variables. Since a  $\chi_p^2$  random variable is a  $\text{gamma}(p/2, 2)$ , application of the example gives part (b).  $\square$

Returning now to the proof of part (c) of Theorem 5.4.1, we will employ an induction argument to establish the distribution of  $S^2$ . (This proof is a slight variation of Kruskal's proof, detailed in Stigler (1984).) We use the notation  $\bar{X}_k$  and  $S_k^2$  to denote the sample mean and variance based on the first  $k$  observations. (Note that the actual ordering of the observations is immaterial—we are just considering them to be ordered to facilitate the proof.) It is straightforward to establish (see Exercise 5.16) that

$$(5.4.1) \quad (n-1)S_n^2 = (n-2)S_{n-1}^2 + \left(\frac{n-1}{n}\right)(X_n - \bar{X}_{n-1})^2.$$

Now consider  $n = 2$ . Defining  $0 \times S_1^2 = 0$ , we have from (5.4.1) that

$$S_2^2 = \frac{1}{2}(X_2 - X_1)^2.$$

Since the distribution of  $(X_2 - X_1)/\sqrt{2}$  is  $n(0, 1)$ , part (a) of Lemma 5.4.1 shows that  $S_2^2 \sim \chi_1^2$ . Proceeding with the induction, assume that for  $n = k$ ,  $(k-1)S_k^2 \sim \chi_{k-1}^2$ . For  $n = k+1$  we have from (5.4.1)

$$(5.4.2) \quad kS_{k+1}^2 = (k-1)S_k^2 + \left(\frac{k}{k+1}\right)(X_{k+1} - \bar{X}_k)^2.$$

According to the induction hypothesis,  $(k-1)S_k^2 \sim \chi_{k-1}^2$ . If we can establish that  $(k/(k+1))(X_{k+1} - \bar{X}_k)^2 \sim \chi_1^2$ , independent of  $S_k^2$ , it will follow from part (b) of Lemma 5.4.1 that  $kS_{k+1}^2 \sim \chi_k^2$ , and the theorem will be proved.

The independence of  $(X_{k+1} - \bar{X}_k)^2$  and  $S_k^2$  again follows from Theorem 4.6.5. The vector  $(X_{k+1}, \bar{X}_k)$  is independent of  $S_k^2$  and so is any function of the vector. Furthermore,  $X_{k+1} - \bar{X}_k$  is a normal random variable with mean 0 and variance

$$\text{Var}(X_{k+1} - \bar{X}_k) = \frac{k+1}{k},$$

and therefore  $(k/(k+1))(X_{k+1} - \bar{X}_k)^2 \sim \chi_1^2$ , and the theorem is established.  $\square$

The independence of  $\bar{X}$  and  $S^2$  can be established in a manner different from that used in the proof of Theorem 5.4.1. Rather than show that the joint pdf factors, we can use the following lemma, which ties together independence and correlation for normal samples.

**LEMMA 5.4.2:** Let  $X_j \sim n(\mu_j, \sigma_j^2)$ ,  $j = 1, \dots, n$ , independent. For constants  $a_{ij}$ ,  $b_{rj}$  ( $j = 1, \dots, n$ ;  $i = 1, \dots, k$ ; and  $r = 1, \dots, m$ ), where  $k + m \leq n$ , define

$$U_i = \sum_{j=1}^n a_{ij} X_j, \quad i = 1, \dots, k,$$

$$V_r = \sum_{j=1}^n b_{rj} X_j, \quad r = 1, \dots, m.$$

- a. The random variables  $U_i$  and  $V_r$  are independent if and only if  $\text{Cov}(U_i, V_r) = 0$ . Furthermore,  $\text{Cov}(U_i, V_r) = \sum_{j=1}^n a_{ij} b_{rj} \sigma_j^2$ .
- b. The random vectors  $(U_1, \dots, U_k)$  and  $(V_1, \dots, V_m)$  are independent if and only if  $U_i$  is independent of  $V_r$  for all pairs  $i, r$  ( $i = 1, \dots, k$ ;  $r = 1, \dots, m$ ).

*Proof:* It is sufficient to prove the lemma for  $\mu_i = 0$  and  $\sigma_i^2 = 1$ , since the general statement of the lemma then follows quickly. Furthermore, the implication from independence to zero covariance is immediate (Theorem 4.5.2) and the expression for the covariance is easily verified (Exercise 5.28). Note also that Corollary 4.6.2 shows that  $U_i$  and  $V_r$  are normally distributed.

Thus, we are left with proving that if the constants satisfy the above restriction (equivalently, the covariance is zero), then we have independence under normality. We prove the lemma only for  $n = 2$ , since the proof for general  $n$  is similar, but necessitates a detailed  $n$ -variate transformation.

To prove part (a) start with the joint pdf of  $X_1$  and  $X_2$

$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{2\pi} e^{-(1/2)(x_1^2 + x_2^2)}, \quad -\infty < x_1, x_2 < \infty.$$

Make the transformation (we can suppress the double subscript in the  $n = 2$  case)

$$u = a_1x_1 + a_2x_2, \quad v = b_1x_1 + b_2x_2,$$

so

$$x_1 = \frac{b_2u - a_2v}{a_1b_2 - b_1a_2}, \quad x_2 = \frac{a_1v - b_1u}{a_1b_2 - b_1a_2},$$

with Jacobian

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial u} & \frac{\partial x_1}{\partial v} \\ \frac{\partial x_2}{\partial u} & \frac{\partial x_2}{\partial v} \end{vmatrix} = \frac{1}{a_1b_2 - b_1a_2}.$$

Thus, the pdf of  $U$  and  $V$  is

$$\begin{aligned} f_{U,V}(u, v) &= f_{X_1, X_2} \left( \frac{b_2u - a_2v}{a_1b_2 - b_1a_2}, \frac{a_1v - b_1u}{a_1b_2 - b_1a_2} \right) |J| \\ &= \frac{1}{2\pi} \exp \left\{ \frac{-1}{2(a_1b_2 - b_1a_2)^2} [(b_2u - a_2v)^2 + (a_1v - b_1u)^2] \right\} |J|, \end{aligned}$$

$-\infty < u, v < \infty$ . Expanding the squares in the exponent, we can write

$$(b_2u - a_2v)^2 + (a_1v - b_1u)^2 = (b_1^2 + b_2^2)u^2 + (a_1^2 + a_2^2)v^2 - 2(a_1b_1 + a_2b_2)uv.$$

The assumption on the constants shows that the cross-term is identically zero. Hence, the pdf factors so, by Lemma 4.2.1,  $U$  and  $V$  are independent and part (a) is established.

A similar type of argument will work for part (b), the details of which we will not go into. If the appropriate transformation is made, the joint pdf of the vectors  $(U_1, \dots, U_k)$  and  $(V_1, \dots, V_m)$  can be obtained. By an application of Theorem 4.6.4, the vectors are independent if the joint pdf factors. From the form of the normal pdf, this will happen if and only if  $U_i$  is independent of  $V_r$  for all pairs  $i, r$  ( $i = 1, \dots, k; r = 1, \dots, m$ ).  $\square$

This lemma shows that, if we start with independent normal random variables, covariance and independence are equivalent. Thus, we can check independence for normal variables by merely checking the covariance term, a much simpler calculation. There is nothing magic about this, it just follows from the form of the normal pdf. Furthermore, part (b) allows us to infer overall independence of normal vectors by just checking pairwise independence, a property that does not hold for general random variables.

We can use Lemma 5.4.2 to provide an alternate proof of the independence of  $\bar{X}$  and  $S^2$  in normal sampling. Since we can write  $S^2$  as a function of  $n - 1$  deviations  $(X_2 - \bar{X}, \dots, X_n - \bar{X})$ , we must show that these random variables are uncorrelated

with  $\bar{X}$ . The normality assumption, together with Lemma 5.4.2, will then allow us to conclude independence.

As an illustration of the application of Lemma 5.4.2, write

$$\bar{X} = \sum_{i=1}^n \left( \frac{1}{n} \right) X_i,$$

$$X_j - \bar{X} = \sum_{i=1}^n \left( \delta_{ij} - \frac{1}{n} \right) X_i,$$

where  $\delta_{ij} = 1$  if  $i = j$  and  $\delta_{ij} = 0$  otherwise. It is then easy to show that

$$\text{Cov}(\bar{X}, X_j - \bar{X}) = \sum_{i=1}^n \left( \frac{1}{n} \right) \left( \delta_{ij} - \frac{1}{n} \right) = 0,$$

showing that  $\bar{X}$  and  $X_j - \bar{X}$  are independent (as long as the  $X_i$ 's have the same variance).

### 5.4.2 The Derived Distributions: Student's $t$ and Snedecor's $F$

The distributions derived in Section 5.4.1 are, in a sense, the first step in a statistical analysis. In particular, in most practical cases the variance,  $\sigma^2$ , is unknown. Thus, to get any idea of the variability of  $\bar{X}$  (as an estimate of  $\mu$ ), it is necessary to estimate this variance. This topic was first addressed by W. S. Gosset (who published under the pseudonym of Student) in the early 1900s. The landmark work of Student resulted in what is known today as Student's  $t$  distribution or, more simply, the  $t$  distribution.

If  $X_1, \dots, X_n$  are a random sample from a  $n(\mu, \sigma^2)$ , we know that the quantity

$$(5.4.3) \quad \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

is distributed as a  $n(0, 1)$  random variable. If we knew the value of  $\sigma$ , and we measured  $\bar{X}$ , then we could use (5.4.3) as a basis for inference about  $\mu$ , since  $\mu$  would then be the only unknown quantity. Most of the time, however,  $\sigma$  is unknown. Student did the obvious thing—he looked at the distribution of

$$(5.4.4) \quad \frac{\bar{X} - \mu}{S / \sqrt{n}},$$

a quantity that could be used as a basis for inference about  $\mu$  when  $\sigma$  was unknown.

The distribution of (5.4.4) is easy to derive, provided that we first notice a few simplifying maneuvers. Multiply (5.4.4) by  $\sigma / \sigma$  and rearrange slightly to obtain

$$(5.4.5) \quad \frac{\bar{X} - \mu}{S / \sqrt{n}} = \frac{(\bar{X} - \mu) / (\sigma / \sqrt{n})}{\sqrt{S^2 / \sigma^2}}.$$

The numerator of (5.4.5) is a  $n(0, 1)$  random variable, and the denominator is  $\sqrt{\chi_{n-1}^2/(n-1)}$ , *independent* of the numerator. Thus, the distribution of (5.4.4) can be found by solving the simplified problem of finding the distribution of  $U/\sqrt{V/p}$ , where  $U$  is  $n(0, 1)$ ,  $V$  is  $\chi_p^2$ , and  $U$  and  $V$  are independent. This gives us Student's  $t$  distribution.

**DEFINITION 5.4.1:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  distribution. The quantity  $(\bar{X} - \mu)/(S/\sqrt{n})$  has *Student's t distribution with  $n-1$  degrees of freedom*. Equivalently, a random variable  $T$  has Student's  $t$  distribution with  $p$  degrees of freedom, and we write  $T \sim t_p$ , if it has pdf

$$(5.4.6) \quad f_T(t) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \frac{1}{(p\pi)^{1/2}} \frac{1}{(1+t^2/p)^{(p+1)/2}}, \quad -\infty < t < \infty.$$

Notice that if  $p = 1$  then (5.4.6) becomes the pdf of the Cauchy distribution, which occurs for samples of size 2. Once again the Cauchy distribution has appeared in an ordinary situation.

The derivation of the  $t$  pdf is straightforward. Starting with  $U$  and  $V$  defined above, it follows from (5.4.5) that the joint pdf of  $U$  and  $V$  is

$$f_{U,V}(u, v) = \frac{1}{(2\pi)^{1/2}} e^{-u^2/2} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2}} v^{(p/2)-1} e^{-v/2}, \quad -\infty < u < \infty, \quad 0 < v < \infty.$$

(Recall that  $U$  and  $V$  are independent.) Now make the transformation

$$t = \frac{u}{\sqrt{v/p}}, \quad w = v.$$

The Jacobian of the transformation is  $(w/p)^{1/2}$ , and the marginal pdf of  $T$  is given by

$$\begin{aligned} f_T(t) &= \int_0^\infty f_{U,V} \left( t \left( \frac{w}{p} \right)^{1/2}, w \right) \left( \frac{w}{p} \right)^{1/2} dw \\ &= \frac{1}{(2\pi)^{1/2}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2}} \int_0^\infty e^{-(1/2)t^2 w/p} w^{(p/2)-1} e^{-w/2} \left( \frac{w}{p} \right)^{1/2} dw \\ &= \frac{1}{(2\pi)^{1/2}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2} p^{1/2}} \int_0^\infty e^{-(1/2)(1+t^2/p)w} w^{((p+1)/2)-1} dw. \end{aligned}$$

Recognize the integrand as the kernel of a  $\text{gamma}((p+1)/2, 2/(1+t^2/p))$  pdf. We therefore have

$$f_T(t) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2} p^{1/2}} \Gamma\left(\frac{p+1}{2}\right) \left[ \frac{2}{1+t^2/p} \right]^{(p+1)/2},$$

which is equal to (5.4.6).

Student's  $t$  has no mgf because it does not have moments of all orders. In fact, if there are  $p$  degrees of freedom, then there are only  $p - 1$  moments. Hence, a  $t_1$  has no mean, a  $t_2$  has no variance, etc. It is easy to check (see Exercise 5.31) that if  $T_p$  is a random variable with a  $t_p$  distribution, then

$$(5.4.7) \quad \begin{aligned} ET_p &= 0 & \text{if } p > 1 \\ \text{Var } T_p &= \frac{p}{p-2} & \text{if } p > 2. \end{aligned}$$

Another important derived distribution is Snedecor's  $F$ , whose derivation is quite similar to that of Student's  $t$ . Its motivation, however, is somewhat different. The  $F$  distribution, named in honor of Sir Ronald Fisher, arises naturally as the distribution of a ratio of variances.

**Example 5.4.1:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu_X, \sigma_X^2)$  population, and let  $Y_1, \dots, Y_m$  be a random sample from an independent  $n(\mu_Y, \sigma_Y^2)$  population. If we were interested in comparing the variability of the populations, one quantity of interest would be the ratio  $\sigma_X^2/\sigma_Y^2$ . Information about this ratio is contained in  $S_X^2/S_Y^2$ , the ratio of sample variances. The  $F$  distribution allows us to compare these quantities by giving us a distribution of

$$(5.4.8) \quad \frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}.$$

Examination of (5.4.8) shows us how the  $F$  distribution is derived. The ratios  $S_X^2/\sigma_X^2$  and  $S_Y^2/\sigma_Y^2$  are each scaled chi squared variates, and they are independent. ||

**DEFINITION 5.4.2:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu_X, \sigma_X^2)$  population, and let  $Y_1, \dots, Y_m$  be a random sample from an independent  $n(\mu_Y, \sigma_Y^2)$  population. The random variable  $F = (S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$  has *Snedecor's F distribution with  $n - 1$  and  $m - 1$  degrees of freedom*. Equivalently, the random variable  $F$  has the  $F$  distribution with  $p$  and  $q$  degrees of freedom if it has pdf

$$(5.4.9) \quad f_F(x) = \frac{\Gamma(\frac{p+q}{2})}{\Gamma(\frac{p}{2}) \Gamma(\frac{q}{2})} \left(\frac{p}{q}\right)^{p/2} \frac{x^{(p/2)-1}}{[1 + (p/q)x]^{(p+q)/2}}, \quad 0 < x < \infty.$$

The  $F$  distribution can be derived in a more general setting than is done here. A variance ratio may have an  $F$  distribution even if the parent populations are not normal. Kelker (1970) has shown that as long as the parent populations have a certain type of symmetry (*spherical symmetry*), then the variance ratio will have an  $F$  distribution.

The derivation of the  $F$  pdf, starting from normal distributions, is similar to the derivation of Student's  $t$ . In fact, in one special case the  $F$  is a transform of the  $t$ . (See Theorem 5.4.2.) Similar to what we did for the  $t$ , we can reduce the task of

deriving the  $F$  pdf to that of finding the pdf of  $(U/p)/(V/q)$ , where  $U$  and  $V$  are independent,  $U \sim \chi_p^2$  and  $V \sim \chi_q^2$ . (See Exercise 5.30.)

**Example 5.4.1 (Continued):** To see how the  $F$  distribution may be used for inference about the true ratio of population variances, consider the following. The quantity  $(S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$  has an  $F_{n-1, m-1}$  distribution. (In general, we use the notation  $F_{p,q}$  to denote an  $F$  random variable with  $p$  and  $q$  degrees of freedom.) We can calculate

$$\begin{aligned} EF_{n-1, m-1} &= E\left(\frac{\chi_{n-1}^2/(n-1)}{\chi_{m-1}^2/(m-1)}\right) && \text{(by definition)} \\ &= E\left(\frac{\chi_{n-1}^2}{n-1}\right) E\left(\frac{m-1}{\chi_{m-1}^2}\right) && \text{(independence)} \\ &= \left(\frac{n-1}{n-1}\right) \left(\frac{m-1}{m-3}\right) && \text{(chi squared calculations)} \\ &= \frac{m-1}{m-3}. \end{aligned}$$

Note that this last expression is finite and positive only if  $m > 3$ . We have that

$$E\left(\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}\right) = EF_{n-1, m-1} = \frac{m-1}{m-3},$$

and, removing expectations, we have for reasonably large  $m$ ,

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \approx \frac{m-1}{m-3} \approx 1,$$

as we might expect. ||

The  $F$  distribution has many interesting properties, and is related to a number of other distributions. We summarize some of these facts in the next theorem, whose proof is left as an exercise. (See Exercises 5.30 and 5.31.)

#### **THEOREM 5.4.2:**

- a. If  $X \sim F_{p,q}$  then  $1/X \sim F_{q,p}$ , that is, the reciprocal of an  $F$  random variable is again an  $F$  random variable.
- b. If  $X \sim t_q$  then  $X^2 \sim F_{1,q}$ .
- c. If  $X \sim F_{p,q}$ , then  $(p/q)X/(1 + (p/q)X) \sim \text{beta}(p/2, q/2)$ . □

## 5.5 Order Statistics

Sample values such as the smallest, largest, or middle observation from a random sample can provide additional summary information. For example, the highest flood

waters or the lowest winter temperature recorded during the last 50 years might be useful data when planning for future emergencies. The median price of houses sold during the previous month might be useful for estimating the cost of living. These are all examples of *order statistics*.

**DEFINITION 5.5.1:** The *order statistics* of a random sample  $X_1, \dots, X_n$  are the sample values placed in ascending order. They are denoted by  $X_{(1)}, \dots, X_{(n)}$ .

The order statistics are random variables that satisfy  $X_{(1)} \leq \dots \leq X_{(n)}$ . In particular,

$$\begin{aligned} X_{(1)} &= \min_{1 \leq i \leq n} X_i, \\ X_{(2)} &= \text{second smallest } X_i, \\ &\vdots \\ X_{(n)} &= \max_{1 \leq i \leq n} X_i. \end{aligned}$$

Since they are random variables, we can discuss the probabilities that they take on various values. To calculate these probabilities we need the pdfs or pmfs of the order statistics. The formulas for the pdfs of the order statistics of a random sample from a continuous population will be the main topic later in this section, but first, we will mention some statistics that are easily defined in terms of the order statistics.

The *sample range*,  $R = X_{(n)} - X_{(1)}$ , is the distance between the smallest and largest observations. It is a measure of the dispersion in the sample and should reflect the dispersion in the population.

The *sample median*, which we will denote by  $M$ , is a number such that approximately one-half of the observations are less than  $M$  and one-half are greater. In terms of the order statistics,  $M$  is defined by

$$(5.5.1) \quad M = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ (X_{(n/2)} + X_{(n/2+1)}) / 2 & \text{if } n \text{ is even.} \end{cases}$$

The median is a measure of location that might be considered an alternative to the sample mean. One advantage of the sample median over the sample mean is that it is less affected by extreme observations. Envision a particular set of sample values and then consider the effect of increasing the largest observation. The sample median is unchanged by this change of the sample values. The sample mean, on the other hand, increases without bound as the largest observation increases. This insensitivity to extreme observations is sometimes considered an asset of the sample median. (A number associated with a statistic, called a *breakdown value*, is calculated as follows. The breakdown value, the idea of which is attributable to Hampel (1974), is the proportion of the sample that can be moved to infinity without the statistic

moving to infinity. It is thought that a large breakdown value is a good indication of the insensitivity of a statistic to underlying assumptions. It is easy to calculate the breakdown values of the mean and the median. It is 50% for the median and 0% for the mean. See Exercise 5.39 for more about breakdown values.)

Although related, the mean and median usually measure different things. For example, in recent baseball salary negotiations a major point of contention was the owners' contributions to the players' pension fund. The owners' view could be paraphrased as, "The average baseball player's annual salary is \$433,659 so, with that kind of money, the current pension is adequate." But the players' view was, "Over half of the players make less than \$250,000 annually and, because of the short professional life of most players, need the security of a larger pension." (These figures are for the 1988 season, not the year of the dispute.) Both figures were correct but the owners were discussing the mean while the players were discussing the median. About a dozen players with salaries over \$2 million can raise the average salary to \$433,659 while the majority of the players make less than \$250,000, including all rookies who make \$62,500. When discussing salaries, prices, or any variable with a few extreme values, the median gives a better indication of "typical" values than the mean. Other statistics which can be defined in terms of order statistics and are less sensitive to extreme values (such as the  $\alpha$ -trimmed mean discussed in Exercise 5.39) are discussed in texts such as Tukey (1977).

For any number  $p$  between 0 and 1, the  $(100p)$ th sample percentile is the observation such that approximately  $np$  of the observations are less than this observation and  $n(1 - p)$  of the observations are greater. The 50th sample percentile ( $p = .5$ ) is the sample median. For other values of  $p$ , we can more precisely define the sample percentiles in terms of the order statistics in the following way.

**DEFINITION 5.5.2:** The notation  $\{b\}$ , when appearing in a subscript, is defined to be *the number b rounded to the nearest integer in the usual way*. More precisely, if  $i$  is an integer and  $i - .5 \leq b < i + .5$ , then  $\{b\} = i$ .

The  $(100p)$ th sample percentile is  $X_{\{\{np\}\}}$  if  $p < .5$  and  $X_{(n+1-\{n(1-p)\})}$  if  $p > .5$ . For example, if  $n = 12$  and the 65th percentile is wanted, we note that  $12 \times (1 - .65) = 4.2$  and  $12 + 1 - 4 = 9$ . Thus the 65th percentile is  $X_{(9)}$ . The cases  $p < .5$  and  $p > .5$  are defined separately so that the sample percentiles exhibit the following symmetry. If the  $(100p)$ th sample percentile is the  $i$ th smallest observation then the  $(100(1 - p))$ th sample percentile should be the  $i$ th largest observation and the above definition achieves this. For example, if  $n = 11$ , the 30th sample percentile is  $X_{(3)}$  and the 70th sample percentile is  $X_{(9)}$ .

In addition to the median, two other sample percentiles are commonly identified. These are the *lower quartile* (25th percentile) and *upper quartile* (75th percentile). A measure of dispersion which is sometimes used is the *interquartile range*, the distance between the lower and upper quartiles.

Since the order statistics are functions of the sample, probabilities concerning order statistics can be computed in terms of probabilities for the sample. If  $X_1, \dots, X_n$  are iid discrete random variables, then the calculation of probabilities for the order statistics is mainly a counting task. These formulas are derived in Theorem 5.5.1.

If  $X_1, \dots, X_n$  are a random sample from a continuous population then convenient expressions for the pdf of one or more order statistics are derived in Theorems 5.5.2 and 5.5.3. These can then be used to derive the distribution of functions of the order statistics.

**THEOREM 5.5.1:** Let  $X_1, \dots, X_n$  be a random sample from a discrete distribution with pmf  $f_X(x_i) = p_i$  where  $x_1 < x_2 < \dots$  are the possible values of  $X$  in ascending order. Define

$$P_0 = 0$$

$$P_1 = p_1$$

$$P_2 = p_1 + p_2$$

 $\vdots$ 

$$P_i = p_1 + p_2 + \dots + p_i$$

 $\vdots$ 

Let  $X_{(1)}, \dots, X_{(n)}$  denote the order statistics from the sample. Then

$$(5.5.2) \quad P(X_{(j)} \leq x_i) = \sum_{k=j}^n \binom{n}{k} P_i^k (1 - P_i)^{n-k}$$

and

$$(5.5.3) \quad P(X_{(j)} = x_i) = \sum_{k=j}^n \binom{n}{k} [P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}].$$

*Proof:* Fix  $i$ , and let  $Y$  be a random variable that counts the number of  $X_1, \dots, X_n$  that are less than or equal to  $x_i$ . For each of  $X_1, \dots, X_n$  call the event  $\{X_j \leq x_i\}$  a “success” and  $\{X_j > x_i\}$  a “failure.” Then  $Y$  is the number of successes in  $n$  trials. The probability of a success is the same value, namely  $P_i = P(X_j \leq x_i)$ , for each trial, since  $X_1, \dots, X_n$  are identically distributed. The success or failure of the  $j$ th trial is independent of the outcome of any other trial since  $X_j$  is independent of the other  $X_i$ s. Thus,  $Y \sim \text{binomial}(n, P_i)$ .

The event  $\{X_{(j)} \leq x_i\}$  is equivalent to the event  $\{Y \geq j\}$ , that is, at least  $j$  of the sample values are less than or equal to  $x_i$ . Equation (5.5.2) expresses this binomial probability

$$P(X_{(j)} \leq x_i) = P(Y \geq j).$$

Equation (5.5.3) simply expresses the difference

$$P(X_{(j)} = x_i) = P(X_{(j)} \leq x_i) - P(X_{(j)} \leq x_{i-1}).$$

The case  $i = 1$  is exceptional in that  $P(X_{(j)} = x_1) = P(X_{(j)} \leq x_1)$ . The definition of  $P_0 = 0$  takes care of this exception in (5.5.3).  $\square$

If  $X_1, \dots, X_n$  are a random sample from a continuous population then the situation is simplified slightly by the fact that the probability is 0 that any two  $X_j$ 's are equal, freeing us from worrying about ties. Thus  $P(X_{(1)} < X_{(2)} < \dots < X_{(n)}) = 1$  and the sample space for  $(X_{(1)}, \dots, X_{(n)})$  is  $\{(x_1, \dots, x_n) : x_1 < x_2 < \dots < x_n\}$ . In Theorems 5.5.2 and 5.5.3 we derive the pdf for one and the joint pdf for two order statistics, again using binomial arguments.

**THEOREM 5.5.2:** Let  $X_{(1)}, \dots, X_{(n)}$  denote the order statistics of a random sample,  $X_1, \dots, X_n$ , from a continuous population with cdf  $F_X(x)$  and pdf  $f_X(x)$ . Then the pdf of  $X_{(j)}$  is

$$(5.5.4) \quad f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x) [F_X(x)]^{j-1} [1 - F_X(x)]^{n-j}.$$

*Proof:* We first find the cdf of  $X_{(j)}$  and then differentiate it to obtain the pdf. As in Theorem 5.5.1, let  $Y$  be a random variable that counts the number of  $X_1, \dots, X_n$  less than or equal to  $x$ . Then, defining a “success” as the event  $\{X_j \leq x\}$ , we see that  $Y \sim \text{binomial}(n, F_X(x))$ . (Note that we can write  $P_i = F_X(x_i)$  in Theorem 5.5.1. Also, although  $X_1, \dots, X_n$  are continuous random variables, the counting variable  $Y$  is discrete.) Thus,

$$F_{X_{(j)}}(x) = P(Y \geq j) = \sum_{k=j}^n \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k},$$

and the pdf of  $X_{(j)}$  is

$$\begin{aligned} f_{X_{(j)}}(x) &= \frac{d}{dx} F_{X_{(j)}}(x) \\ &= \sum_{k=j}^n \binom{n}{k} \left( k [F_X(x)]^{k-1} [1 - F_X(x)]^{n-k} f_X(x) \right. \\ &\quad \left. - (n-k) [F_X(x)]^k [1 - F_X(x)]^{n-k-1} f_X(x) \right) \quad (\text{chain rule}) \\ &= \binom{n}{j} j f_X(x) [F_X(x)]^{j-1} [1 - F_X(x)]^{n-j} \\ &\quad + \sum_{k=j+1}^n \binom{n}{k} k [F_X(x)]^{k-1} [1 - F_X(x)]^{n-k} f_X(x) \\ &\quad - \sum_{k=j}^{n-1} \binom{n}{k} (n-k) [F_X(x)]^k [1 - F_X(x)]^{n-k-1} f_X(x) \quad \begin{pmatrix} k = n \text{ term} \\ \text{is zero} \end{pmatrix} \end{aligned}$$

$$\begin{aligned}
 &= \frac{n!}{(j-1)!(n-j)!} f_X(x) [F_X(x)]^{j-1} [1 - F_X(x)]^{n-j} \\
 (5.5.5) \quad &+ \sum_{k=j}^{n-1} \binom{n}{k+1} (k+1) [F_X(x)]^k [1 - F_X(x)]^{n-k-1} f_X(x) \left( \begin{array}{c} \text{change dummy} \\ \text{variable} \end{array} \right) \\
 &- \sum_{k=j}^{n-1} \binom{n}{k} (n-k) [F_X(x)]^k [1 - F_X(x)]^{n-k-1} f_X(x).
 \end{aligned}$$

Noting that

$$\begin{aligned}
 (5.5.6) \quad \binom{n}{k+1} (k+1) &= \frac{n!}{k!(n-k-1)!} \\
 &= \binom{n}{k} (n-k),
 \end{aligned}$$

we see that the last two sums in (5.5.5) cancel. Thus, the pdf  $f_{X_{(j)}}(x)$  is given by the expression in (5.5.4).  $\square$

**Example 5.5.1:** Let  $X_1, \dots, X_n$  be iid uniform(0, 1), so  $f_X(x) = 1$  for  $x \in (0, 1)$  and  $F_X(x) = x$  for  $x \in (0, 1)$ . Using (5.5.4), we see that the pdf of the  $j$ th order statistic is

$$\begin{aligned}
 f_{X_{(j)}}(x) &= \frac{n!}{(j-1)!(n-j)!} x^{j-1} (1-x)^{n-j} \quad \text{for } x \in (0, 1) \\
 &= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1} (1-x)^{(n-j+1)-1}.
 \end{aligned}$$

Thus, the  $j$ th order statistic from a uniform(0, 1) sample has a beta( $j, n-j+1$ ) distribution. From this we can deduce that

$$\mathbb{E}X_{(j)} = \frac{j}{n+1} \quad \text{and} \quad \text{Var } X_{(j)} = \frac{j(n-j+1)}{(n+1)^2(n+2)}. \quad \parallel$$

The joint distribution of two or more order statistics can be used to derive the distribution of some of the statistics mentioned at the beginning of this section. The joint pdf of any two order statistics is given in the following theorem.

**THEOREM 5.5.3:** Let  $X_{(1)}, \dots, X_{(n)}$  denote the order statistics of a random sample,  $X_1, \dots, X_n$ , from a continuous population with cdf  $F_X(x)$  and pdf  $f_X(x)$ . Then the joint pdf of  $X_{(i)}$  and  $X_{(j)}$ ,  $1 \leq i < j \leq n$ , is

$$(5.5.7) \quad f_{X_{(i)}, X_{(j)}}(u, v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f_X(u) f_X(v) [F_X(u)]^{i-1} \\ \times [F_X(v) - F_X(u)]^{j-1-i} [1 - F_X(v)]^{n-j}$$

for  $-\infty < u < v < \infty$ .

*Proof:* We first obtain an expression for the joint cdf of  $X_{(i)}$  and  $X_{(j)}$  and then find the joint pdf by computing the mixed partial as indicated in (4.1.3). Let  $U$  be a random variable that counts the number of  $X_1, \dots, X_n$  less than or equal to  $u$  and  $V$  be a random variable that counts the number of  $X_1, \dots, X_n$  greater than  $u$  and less than or equal to  $v$ . Then  $(U, V, n - U - V)$  is a multinomial random vector with  $n$  trials and cell probabilities  $(F_X(u), F_X(v) - F_X(u), 1 - F_X(v))$ . The joint cdf can be expressed as

$$\begin{aligned} F_{X_{(i)}, X_{(j)}}(u, v) &= P(U \geq i, U + V \geq j) \\ &= \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} P(U = k, V = m) + P(U \geq j) \\ (5.5.8) \quad &= \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} \frac{n!}{k!m!(n-k-m)!} [F_X(u)]^k [F_X(v) - F_X(u)]^m [1 - F_X(v)]^{n-k-m} \\ &\quad + P(U \geq j). \end{aligned}$$

The joint pdf is

$$f_{X_{(i)}, X_{(j)}}(u, v) = \frac{\partial^2}{\partial u \partial v} F_{X_{(i)}, X_{(j)}}(u, v).$$

The mixed partial of  $P(U \geq j)$  is 0 since this term depends only on  $u$ , not  $v$ . The mixed partial of each of the terms in (5.5.8) yields four terms when the chain rule is applied. Some of these terms are 0. The rest of these terms all cancel with each other, using relationships like (5.5.6), except for the one term given in (5.5.7). Details are left to Exercise 5.40.  $\square$

The joint pdf of three or more order statistics could be derived using similar but even more involved arguments. Perhaps the other most useful pdf is  $f_{X_{(1)}, \dots, X_{(n)}}(x_1, \dots, x_n)$ , the joint pdf of all the order statistics, which is given by

$$f_{X_{(1)}, \dots, X_{(n)}}(x_1, \dots, x_n) = \begin{cases} n! f_X(x_1) \cdots f_X(x_n) & -\infty < x_1 < \cdots < x_n < \infty \\ 0 & \text{otherwise} \end{cases}$$

The  $n!$  naturally comes into this formula because, for any set of values,  $x_1, \dots, x_n$ , there are  $n!$  equally likely assignments for these values to  $X_1, \dots, X_n$  which all yield the same values for the order statistics. This joint pdf and the techniques from Chapter

4 can be used to derive marginal and conditional distributions and distributions of other functions of the order statistics. (See Exercises 5.41 and 5.42.)

We now use the joint pdf (5.5.7) to derive the distribution of some of the functions mentioned at the beginning of this section.

**Example 5.5.2:** Let  $X_1, \dots, X_n$  be iid uniform(0,  $a$ ) and let  $X_{(1)}, \dots, X_{(n)}$  denote the order statistics. The range was earlier defined as  $R = X_{(n)} - X_{(1)}$ . The *midrange*, a measure of location like the sample median or the sample mean, is defined by  $V = (X_{(1)} + X_{(n)})/2$ . We will derive the joint pdf of  $R$  and  $V$  from the joint pdf of  $X_{(1)}$  and  $X_{(n)}$ . From (5.5.7) we have that

$$\begin{aligned} f_{X_{(1)}, X_{(n)}}(x_1, x_n) &= \frac{n(n-1)}{a^2} \left(\frac{x_n}{a} - \frac{x_1}{a}\right)^{n-2} \\ &= \frac{n(n-1)(x_n - x_1)^{n-2}}{a^n}, \quad 0 < x_1 < x_n < a. \end{aligned}$$

Solving for  $X_{(1)}$  and  $X_{(n)}$ , we obtain  $X_{(1)} = V - R/2$  and  $X_{(n)} = V + R/2$ . The Jacobian for this transformation is  $-1$ . The transformation from  $(X_{(1)}, X_{(n)})$  to  $(R, V)$  maps  $\{(x_1, x_n) : 0 < x_1 < x_n < a\}$  onto the set  $\{(r, v) : 0 < r < a, r/2 < v < a - r/2\}$ . To see this note that obviously  $0 < r < a$  and for a fixed value of  $r$ ,  $v$  ranges from  $r/2$  (corresponding to  $x_1 = 0, x_n = r$ ) to  $a - r/2$  (corresponding to  $x_1 = a - r, x_n = a$ ). Thus, the joint pdf of  $(R, V)$  is

$$f_{R,V}(r, v) = \frac{n(n-1)r^{n-2}}{a^n}, \quad 0 < r < a, \quad r/2 < v < a - r/2.$$

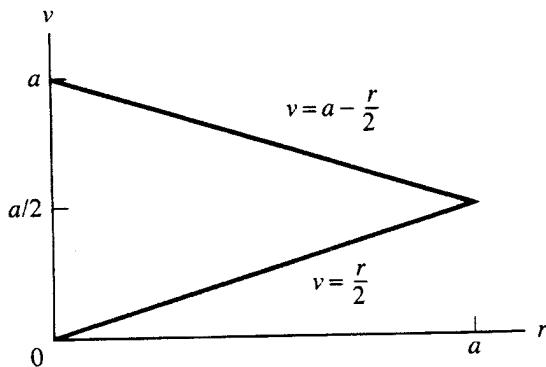
The marginal pdf of  $R$  is thus

$$\begin{aligned} f_R(r) &= \int_{r/2}^{a-r/2} \frac{n(n-1)r^{n-2}}{a^n} dv \\ (5.5.9) \quad &= \frac{n(n-1)r^{n-2}(a-r)}{a^n}, \quad 0 < r < a. \end{aligned}$$

If  $a = 1$ , we see that  $r$  has a beta( $n - 1, 2$ ) distribution. Or, for arbitrary  $a$ , it is easy to deduce from (5.5.9) that  $R/a$  has a beta distribution. Note that the constant  $a$  is a scale parameter.

The set where  $f_{R,V}(r, v) > 0$  is shown in Figure 5.5.1 (page 236), where we see that the range of integration of  $r$  depends on whether  $v > a/2$  or  $v \leq a/2$ . Thus, the marginal pdf of  $V$  is given by

$$\begin{aligned} f_V(v) &= \int_0^{2v} \frac{n(n-1)r^{n-2}}{a^n} dr \\ &= \frac{n(2v)^{n-1}}{a^n}, \quad 0 < v \leq a/2 \end{aligned}$$

FIGURE 5.5.1 Region on which  $f_{R,V}(r, v) > 0$  for Example 5.5.2

and

$$\begin{aligned} f_V(v) &= \int_0^{2(a-v)} \frac{n(n-1)r^{n-2}}{a^n} dr \\ &= \frac{n[2(a-v)]^{n-1}}{a^n}, \quad a/2 < v \leq a. \end{aligned}$$

This pdf is symmetric about  $a/2$  and has a peak at  $a/2$ . ||

### EXERCISES

- 5.1 Color blindness appears in 1% of the people in a certain population. How large must a sample be if the probability of its containing a color-blind person is to be .95 or more? (Assume that the population is large enough to be considered infinite, so that sampling can be considered to be with replacement.)
- 5.2 Suppose  $X_1, X_2, \dots$  are jointly continuous and independent, each distributed with marginal pdf  $f(x)$ . If the  $X_i$ 's represent annual rainfalls at a given location, find the distribution of the number of years until the first year's rainfall,  $X_1$ , is exceeded for the first time.
- 5.3 Let  $X_1, \dots, X_n$  be iid random variables with continuous cdf  $F_X$ , and suppose  $\text{E}X_i = \mu$ . Define the random variables  $Y_1, \dots, Y_n$  by

$$Y_i = \begin{cases} 1 & \text{if } X_i > \mu \\ 0 & \text{if } X_i \leq \mu \end{cases}.$$

Find the distribution of  $\sum_{i=1}^n Y_i$ .

- 5.4 A generalization of iid random variables is *exchangeable* random variables, an idea due to De Finetti (1972). A discussion of exchangeability can also be found in Feller (1971). The random variables  $X_1, \dots, X_n$  are *exchangeable* if any permutation of any subset of them of size  $k$  ( $k \leq n$ ) has the same distribution. In this exercise we will see an example of random variables that are exchangeable, but not iid. Let  $X_i|P \sim \text{iid Bernoulli}(P)$ ,  $i = 1, \dots, n$ , and let  $P \sim \text{uniform}(0, 1)$ .

- a. Show that the marginal distribution of any  $k$  of the  $X$ s is the same as

$$P(X_1 = x_1, \dots, X_k = x_k) = \int_0^1 p^t(1-p)^{k-t} dp = \frac{t!(k-t)!}{(k+1)!},$$

where  $t = \sum_{i=1}^k x_i$ . Hence, the  $X$ s are exchangeable.

- b. Show that, marginally,

$$P(X_1 = x_1, \dots, X_n = x_n) \neq \prod_{i=1}^n P(X_i = x_i),$$

so the distribution of the  $X$ s is exchangeable, but not iid.

(De Finetti proved an elegant characterization theorem for an infinite sequence of exchangeable random variables. He proved that any such sequence of exchangeable random variables is a mixture of iid random variables.)

- 5.5** Let  $X_1, \dots, X_n$  be iid with pdf or pmf  $f_X(x)$ , and let  $\bar{X}$  denote the sample mean. Show that

$$f_{\bar{X}}(x) = n f_{X_1 + \dots + X_n}(nx),$$

even if the mgf of  $X$  does not exist.

- 5.6** If  $X$  has pdf  $f_X(x)$  and  $Y$ , independent of  $X$ , has pdf  $f_Y(y)$ , establish formulas, similar to (5.2.3), for the random variable  $Z$  in each of the following situations.

- a.  $Z = X - Y$
- b.  $Z = XY$
- c.  $Z = X/Y$

- 5.7** In Example 5.2.2, a partial fraction decomposition is needed to derive the distribution of the sum of two independent Cauchy random variables. This exercise provides the details that are skipped in that example.

- a. Find the constants for the partial fraction decomposition. That is, find  $A$ ,  $B$ ,  $C$ , and  $D$  that satisfy

$$\begin{aligned} & \frac{1}{1 + (w/\sigma)^2} \frac{1}{1 + ((z-w)/\tau)^2} \\ &= \frac{Aw}{1 + (w/\sigma)^2} + \frac{B}{1 + (w/\sigma)^2} - \frac{Cw}{1 + ((z-w)/\tau)^2} - \frac{D}{1 + ((z-w)/\tau)^2}, \end{aligned}$$

where  $A$ ,  $B$ ,  $C$ , and  $D$  may depend on  $z$  but not on  $w$ .

- b. Using the decomposition of part (a) and the integration formulas given in Example 5.2.2, show that, if  $X$  is Cauchy( $0, \sigma$ ) and  $Y$  is Cauchy( $0, \tau$ ) independent of  $X$ , then  $Z = X + Y$  is Cauchy( $0, \sigma + \tau$ ).

- 5.8** Show that, for any random vector (not necessarily a random sample)  $(X_1, \dots, X_n)$ ,

$$\text{Var} \left( \sum_{i=1}^n X_i \right) = \sum_{i=1}^n \text{Var} X_i + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j).$$

- 5.9** Let  $X_1, \dots, X_n$  be a random sample, where  $\bar{X}$  and  $S^2$  are calculated in the usual way.

- a. Show that

$$S^2 = \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (X_i - X_j)^2.$$

Assume now that the  $X_i$ 's have a finite fourth moment, and denote  $\theta_1 = EX_i, \theta_j = E(X_i - \theta_1)^j, j = 2, 3, 4$ .

b. Show that  $\text{Var } S^2 = \frac{1}{n}(\theta_4 - \frac{n-3}{n-1}\theta_2^2)$ .

c. Find  $\text{Cov}(\bar{X}, S^2)$  in terms of  $\theta_1, \dots, \theta_4$ . Under what conditions is  $\text{Cov}(\bar{X}, S^2) = 0$ ?

- 5.10** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population.
- Find expressions for  $\theta_1, \dots, \theta_4$ , as defined in the previous exercise, in terms of  $\mu$  and  $\sigma^2$ .
  - Use the results of the previous exercise, together with the results of part (a), to calculate  $\text{Var } S^2$ .
  - Calculate  $\text{Var } S^2$  a completely different (and easier) way: Use the fact that  $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$ .
- 5.11** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  distribution, which is an exponential family. Show that the condition in Theorem 5.2.5, that the sample space of  $(T_1, \dots, T_k)$  contain an open subset of  $\mathbb{R}^k$ , is satisfied if  $n \geq k = 2$  but not if  $n = 1$ .
- 5.12** Suppose  $\bar{X}$  and  $S^2$  are calculated from a random sample  $X_1, \dots, X_n$  drawn from a population with finite variance  $\sigma^2$ . We know that  $ES^2 = \sigma^2$ . Prove that  $ES \leq \sigma$ , and, if  $\sigma^2 > 0$ , then  $ES < \sigma$ .
- 5.13** Formulate and prove a version of the WLLN with a weaker assumption than a finite variance. Use the Markov Inequality, instead of Chebychev's Inequality, in the proof.
- 5.14** In Example 5.3.3, find a subsequence of the  $Y_i$ 's that converges almost surely, that is, that converges pointwise.
- 5.15** Let  $X_1, X_2, \dots$  be a sequence of random variables that converges in probability to a constant  $a$ . Assume that  $P(X_i > 0) = 1$  for all  $i$ .
  - Show that the sequence  $Y_1, Y_2, \dots$ , defined by  $Y_i = \sqrt{X_i}$ , converges in probability to  $\sqrt{a}$ .
  - Show that, if  $a > 0$ , the sequence  $Y_1, Y_2, \dots$ , defined by  $Y_i = a/X_i$ , converges in probability to 1.
  - Use the results in parts (a) and (b) to prove the fact used in Example 5.3.5, that  $\sigma/S_n$  converges in probability to 1.
- 5.16** Establish the following recursion relations for means and variances. Let  $\bar{X}_n$  and  $S_n^2$  be the mean and variance, respectively, of  $X_1, \dots, X_n$ . Then suppose another observation,  $X_{n+1}$ , becomes available. Show that
  - $\bar{X}_{n+1} = \frac{\bar{X}_n + n\bar{X}_n}{n+1}$
  - $nS_{n+1}^2 = (n-1)S_n^2 + \left(\frac{n}{n+1}\right)(X_{n+1} - \bar{X}_n)^2$
- 5.17** A manufacturer of booklets packages them in boxes of 100. It is known that, on the average, the booklets weigh 1 ounce, with a standard deviation of .05 oz. The manufacturer is interested in calculating

$$P(100 \text{ booklets weigh more than } 100.4 \text{ oz}),$$

a number that would help detect whether too many booklets are being put in a box. Explain how you would calculate the (approximate?) value of this probability. Mention any relevant theorems or assumptions needed.

- 5.18 If  $\bar{X}_1$  and  $\bar{X}_2$  are the means of two independent samples of size  $n$  from a population with variance  $\sigma^2$ , find a value for  $n$  so that  $P(|\bar{X}_1 - \bar{X}_2| < \sigma/5) \approx .99$ . Justify your calculations.
- 5.19 Suppose  $\bar{X}$  is the mean of 100 observations from a population with mean  $\mu$  and variance  $\sigma^2 = 9$ . Find limits between which  $\bar{X} - \mu$  will lie with probability at least .90. Use both Chebychev's Inequality and the Central Limit Theorem, and comment on each.
- 5.20 Let  $X_1, \dots, X_n$  be a random sample from a population with mean  $\mu$  and variance  $\sigma^2$ . Show that

$$E \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = 0 \quad \text{and} \quad \text{Var} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = 1.$$

Thus, the normalization of  $\bar{X}_n$  in the Central Limit Theorem gives random variables that have the same mean and variance as the limiting  $n(0, 1)$  distribution.

- 5.21 Stirling's Formula (derived in Exercise 1.23), which gives an approximation for factorials, can be easily derived using the CLT.  
 a. Argue that, if  $X_i \sim \text{exponential}(1)$ ,  $i = 1, 2, \dots$ , all independent, then for every  $x$ ,

$$P \left( \frac{\bar{X}_n - 1}{1/\sqrt{n}} \leq x \right) \rightarrow P(Z \leq x),$$

where  $Z$  is a standard normal random variable.

- b. Show that differentiating both sides of the approximation in part (a) suggests

$$\frac{\sqrt{n}}{\Gamma(n)} (x\sqrt{n} + n)^{n-1} e^{-(x\sqrt{n} + n)} \approx \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

and that  $x = 0$  gives Stirling's Formula.

- 5.22 Given that  $N = n$ , the conditional distribution of  $Y$  is  $\chi^2_{2n}$ . The unconditional distribution of  $N$  is  $\text{Poisson}(\theta)$ .
- a. Calculate  $EY$  and  $\text{Var } Y$  (unconditional moments).
- b. Show that, as  $\theta \rightarrow \infty$ ,  $(Y - EY)/\sqrt{\text{Var } Y} \rightarrow n(0, 1)$  in distribution.
- 5.23 In Example 5.3.4, a normal approximation to the negative binomial distribution was given. Just as with the normal approximation to the binomial distribution given in Example 3.2.2, the approximation might be improved with a "continuity correction." For  $X_i$ 's defined as in Example 5.3.4, let  $V_n = \sum_{i=1}^n X_i$ . For  $n = 10$ ,  $p = .7$ , and  $r = 2$ , calculate  $P(V_n = v)$  for  $v = 0, 1, \dots, 10$  using each of the following three methods.
- a. Exact calculations.
- b. Normal approximation as given in Example 5.3.4.
- c. Normal approximation with continuity correction.
- 5.24 The generation of random (or *pseudo-random*) numbers that approximately follow a standard normal distribution has already been addressed in Exercises 4.29 and 4.30. We again look at the problem of generating standard normal random variables from uniform ones. One method (not one of the better ones) is to take  $X = \sum_{i=1}^{12} U_i - 6$ , where the  $U_i$ 's are iid  $\text{uniform}(0, 1)$ .
- a. Justify the fact that  $X$  is approximately  $n(0, 1)$ .
- b. Can you think of any obvious way in which the approximation fails?
- c. Show how good (or bad) the approximation is by comparing the first four moments. (The fourth moment is a lengthy calculation.)

- 5.25** An alternate proof of Theorem 5.3.3 (the Central Limit Theorem) was given by Tardiff (1981). Tardiff evaluates  $\lim_{n \rightarrow \infty} [M_Y(t/\sqrt{n})]^n$  by evaluating the limit of the logarithm using l'Hospital's Rule. (We use the notation of Theorem 5.3.3.)  
 a. Use the properties of logarithms and l'Hospital's Rule to show

$$\begin{aligned}\lim_{n \rightarrow \infty} \log \left( \left[ M_Y \left( \frac{t}{\sqrt{n}} \right) \right]^n \right) &= \lim_{n \rightarrow \infty} \frac{\log(M_Y(t/\sqrt{n}))}{1/n} \\ &= \frac{t}{2} \left( \lim_{n \rightarrow \infty} \frac{M'_Y(t/\sqrt{n})/M_Y(t/\sqrt{n})}{1/\sqrt{n}} \right).\end{aligned}$$

- b. Argue that  $M'_Y(0) = 0$  and thus l'Hospital's Rule can be applied again. Doing so, we obtain that the expression in part (a) is equal to

$$\frac{t^2}{2} \left( \lim_{n \rightarrow \infty} \frac{M_Y(t/\sqrt{n})M''_Y(t/\sqrt{n}) - [M'_Y(t/\sqrt{n})]^2}{[M_Y(t/\sqrt{n})]^2} \right).$$

- c. Show that  $M''_Y(t/\sqrt{n}) \rightarrow 1$  as  $n \rightarrow \infty$ , and hence this last limit is equal to 1. Thus, combining all of the previous calculations, show that

$$\lim_{n \rightarrow \infty} M_{\sqrt{n}(\bar{X}_n - \mu)/\sigma}(t) = e^{t^2/2}.$$

- 5.26** Let  $X_1, \dots, X_n$  be a random sample from a  $n(0, 1)$  population. Define

$$Y_1 = \left| \frac{1}{n} \sum_{i=1}^n X_i \right|, \quad Y_2 = \frac{1}{n} \sum_{i=1}^n |X_i|.$$

Calculate  $EY_1$  and  $EY_2$ , and establish an inequality between them.

- 5.27** Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$ . Find a function of  $S^2$ , the sample variance, say  $g(S^2)$ , that satisfies  $Eg(S^2) = \sigma$ . (Hint: Try  $g(S^2) = c\sqrt{S^2}$ , where  $c$  is a constant.)  
**5.28** a. Prove that the statement of Lemma 5.4.2 follows from the special case of  $\mu_i = 0$  and  $\sigma_i^2 = 1$ . That is, show that if  $X_j = \sigma_j Z_j + \mu_j$  and  $Z_j \sim n(0, 1)$ ,  $j = 1, \dots, n$ , all independent,  $a_{ij}, b_{rj}$  are constants, and

$$\text{Cov} \left( \sum_{j=1}^n a_{ij} Z_j, \sum_{j=1}^n b_{rj} Z_j \right) = 0 \Rightarrow \sum_{j=1}^n a_{ij} Z_j \text{ and } \sum_{j=1}^n b_{rj} Z_j \text{ are independent,}$$

then

$$\text{Cov} \left( \sum_{j=1}^n a_{ij} X_j, \sum_{j=1}^n b_{rj} X_j \right) = 0 \Rightarrow \sum_{j=1}^n a_{ij} X_j \text{ and } \sum_{j=1}^n b_{rj} X_j \text{ are independent.}$$

- b. Verify the expression for  $\text{Cov} \left( \sum_{j=1}^n a_{ij} X_j, \sum_{j=1}^n b_{rj} X_j \right)$  in Lemma 5.4.2.  
**5.29** Let  $X_i, i = 1, 2, 3$ , be independent with  $n(i, i^2)$  distributions. For each of the following situations, use the  $X_i$ s to construct a statistic with the indicated distribution.  
 a. Chi squared with 3 degrees of freedom.

- b.  $t$  distribution with 2 degrees of freedom.  
 c.  $F$  distribution with 1 and 2 degrees of freedom.
- 5.30** Let  $X$  be a random variable with an  $F_{p,q}$  distribution.
- Derive the pdf of  $X$ .
  - Derive the mean and variance of  $X$ .
  - Show that  $1/X$  has an  $F_{q,p}$  distribution.
  - Show that  $(p/q)X/[1 + (p/q)X]$  has a beta distribution with parameters  $p/2$  and  $q/2$ .
- 5.31** Let  $X$  be a random variable with a Student's  $t$  distribution with  $p$  degrees of freedom.
- Derive the mean and variance of  $X$ .
  - Show that  $X^2$  has an  $F$  distribution with 1 and  $p$  degrees of freedom.
  - Let  $f(x|p)$  denote the pdf of  $X$ . Show that

$$\lim_{p \rightarrow \infty} f(x|p) \rightarrow \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

at each value of  $x$ ,  $-\infty < x < \infty$ . This correctly suggests that as  $p \rightarrow \infty$ ,  $X$  converges in distribution to a  $n(0, 1)$  random variable. (*Hint:* Use Stirling's Formula.)

- d. Use the results of parts (a) and (b) to argue that, as  $p \rightarrow \infty$ ,  $X^2$  converges in distribution to a  $\chi_1^2$  random variable.  
 e. What might you conjecture about the distributional limit, as  $p \rightarrow \infty$ , of  $qF_{q,p}$ ?
- 5.32** a. Prove that the  $\chi^2$  distribution is *stochastically increasing* in its degrees of freedom; that is, if  $p > q$  then for any  $a$ ,  $P(\chi_p^2 > a) > P(\chi_q^2 > a)$ .  
 b. Use the results of part (a) to prove that for any  $\nu$ ,  $kF_{k,\nu}$  is stochastically increasing in  $k$ .  
 c. Show that for any  $k$ ,  $\nu$ , and  $\alpha$ ,  $kF_{\alpha,k,\nu} > (k-1)F_{\alpha,k-1,\nu}$ . (See the *Miscellanea* section for Chapter 8, and also Exercise 11.16.)
- 5.33** a. We can see that the  $t$  distribution is a mixture of normals using the following argument:

$$P(T_\nu \leq t) = P\left(\frac{Z}{\sqrt{\chi_\nu^2/\nu}} \leq t\right) = \int_0^\infty P(Z \leq t\sqrt{x}/\sqrt{\nu}) P(\chi_\nu^2 = x) dx,$$

where  $T_\nu$  is a  $t$  random variable with  $\nu$  degrees of freedom. Using the Fundamental Theorem of Calculus, and interpreting  $P(\chi_\nu^2 = \nu x)$  as a pdf, we obtain

$$f_{T_\nu}(t) = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2 x/2\nu} \frac{\sqrt{x}}{\sqrt{\nu}} \frac{1}{\Gamma(\nu/2) 2^{\nu/2}} (x)^{(\nu/2)-1} e^{-x/2} dx,$$

- a scale mixture of normals. Verify this formula by direct integration.  
 b. A similar formula holds for the  $F$  distribution; that is, it can be written as a mixture of chi squareds. If  $F_{1,\nu}$  is an  $F$  random variable with 1 and  $\nu$  degrees of freedom, then we can write

$$P(F_{1,\nu} \leq \nu t) = \int_0^\infty P(\chi_1^2 \leq ty) f_\nu(y) dy,$$

where  $f_\nu(y)$  is a  $\chi_\nu^2$  pdf. Use the Fundamental Theorem of Calculus to obtain an integral expression for the pdf of  $F_{1,\nu}$ , and show that the integral equals the pdf.

- c. Verify that the generalization of part (b),

$$P\left(F_{m,\nu} \leq \frac{\nu}{m}t\right) = \int_0^{\infty} P(\chi_m^2 \leq ty) f_{\nu}(y) dy,$$

is valid for all integers  $m > 1$ .

- 5.34 What is the probability that the larger of two continuous iid random variables will exceed the population median? Generalize this result to samples of size  $n$ .
- 5.35 Let  $X$  and  $Y$  be iid  $n(0, 1)$  random variables, and define  $Z = \min(X, Y)$ . Prove that  $Z^2 \sim \chi_1^2$ .
- 5.36 Let  $U_i, i = 1, 2, \dots$ , be independent uniform(0, 1) random variables, and let  $X$  have distribution

$$P(X = x) = \frac{c}{x!}, \quad x = 1, 2, 3, \dots$$

where  $c = 1/(e - 1)$ . Find the distribution of

$$Z = \min\{U_1, \dots, U_X\}.$$

(Hint: Note that the distribution of  $Z|X = x$  is that of the first order statistic from a sample of size  $x$ .)

- 5.37 Let  $X_1, \dots, X_n$  be a random sample from a population with pdf

$$f_X(x) = \begin{cases} 1/\theta & \text{if } 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

Let  $X_{(1)} < \dots < X_{(n)}$  be the order statistics. Show that  $X_{(1)}/X_{(n)}$  and  $X_{(n)}$  are independent random variables.

- 5.38 As a generalization of the previous exercise, let  $X_1, \dots, X_n$  be iid with pdf

$$f_X(x) = \begin{cases} \frac{a}{\theta^a} x^{a-1} & \text{if } 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

Let  $X_{(1)} < \dots < X_{(n)}$  be the order statistics. Show that  $X_{(1)}/X_{(2)}, X_{(2)}/X_{(3)}, \dots, X_{(n-1)}/X_{(n)}$ , and  $X_{(n)}$  are mutually independent random variables. Find the distribution of each of them.

- 5.39 This exercise has to do with the computation of *breakdown values*, as discussed in Section 5.5. Formally we can define a breakdown value in the following way. Let  $X_{(1)} < \dots < X_{(n)}$  be an ordered sample of size  $n$ , and let  $T_n$  be a statistic based on this sample.  $T_n$  has breakdown value  $b$ ,  $0 \leq b \leq 1$  if, for every  $\epsilon > 0$ ,

$$\lim_{X_{(\{(1-b)n\}} \rightarrow \infty}} T_n < \infty \quad \text{and} \quad \lim_{X_{(\{(1-(b+\epsilon))n\}} \rightarrow \infty}} T_n = \infty.$$

- a. If  $T_n = \bar{X}_n$ , the sample mean, show that  $b = 0$ .  
 b. If  $T_n = M_n$ , the sample median, show that  $b = .5$ .

An estimator that “splits the difference” between the mean and the median in terms of sensitivity is the  $\alpha$ -trimmed mean,  $0 < \alpha < \frac{1}{2}$ , defined as follows.  $\bar{X}_n^\alpha$ , the  $\alpha$ -trimmed mean, is computed by deleting the  $\alpha n$  smallest observations and the  $\alpha n$  largest observations, and taking the arithmetic mean of the remaining observations.

- c. If  $T_n = \bar{X}_n^\alpha$ , the  $\alpha$ -trimmed mean of the sample,  $0 < \alpha < \frac{1}{2}$ , show that  $0 < b < \frac{1}{2}$ .

- 5.40** Complete the proof of Theorem 5.5.3. That is, compute the mixed partial derivative and show that the appropriate terms cancel.
- 5.41** Let  $X_1, \dots, X_n$  be iid with pdf  $f_X(x)$  and cdf  $F_X(x)$ , and let  $X_{(1)} < \dots < X_{(n)}$  be the order statistics.
- Find an expression for the conditional pdf of  $X_{(i)}$  given  $X_{(j)}$  in terms of  $f_X$  and  $F_X$ .
  - Find the pdf of  $V|R = r$ , where  $V$  and  $R$  are defined in Example 5.5.2.
- 5.42** Let  $X_1, \dots, X_n$  be iid with pdf  $f_X(x)$  and cdf  $F_X(x)$ , and let  $X_{(i_1)} < \dots < X_{(i_l)}$  and  $X_{(j_1)} < \dots < X_{(j_m)}$  be any two disjoint groups of order statistics. In terms of the pdf  $f_X(x)$  and the cdf  $F_X(x)$ , find expressions for
- The marginal cdf and pdf of  $X_{(i_1)}, \dots, X_{(i_l)}$ .
  - The conditional cdf and pdf of  $X_{(i_1)}, \dots, X_{(i_l)}$  given  $X_{(j_1)}, \dots, X_{(j_m)}$ .

## Miscellanea

---

### *More on the Central Limit Theorem*

For the case of a sequence of iid random variables, necessary and sufficient conditions for convergence to normality are known, with probably the most famous result due to Lindeberg and Feller. The following special case is due to Lévy. Let  $X_1, X_2, \dots$  be an iid sequence with  $\text{EX}_i = \mu < \infty$ , and let  $V_n = \sum_{i=1}^n X_i$ . The sequence  $V_n$  will converge to a  $n(0, 1)$  random variable (when suitably normalized) if, and only if,

$$\lim_{t \rightarrow \infty} \frac{t^2 P(|X_1 - \mu| > t)}{\text{E}((X_1 - \mu)^2 I_{[-t, t]}(X_1 - \mu))} = 0.$$

Note that the condition is a variance condition. While it does not quite require that the variances be finite, it does require that they be “almost” finite. This is an important point in the convergence to normality—normality comes from summing up small disturbances.

Other types of central limit theorems abound—in particular, ones aimed at relaxing the independence assumption. While this assumption cannot be done away with, it can be made less restrictive. Chung (1974) has a full development of this topic.

### *The Independence of $\bar{X}$ and $S^2$*

S. M. Stigler (1984) outlines the proof of the independence of  $\bar{X}$  and  $S^2$  taught by William Kruskal. There are elements of Kruskal’s proof in our proof of Theorem 5.4.1, but there are also some differences. Kruskal uses induction to prove the entire theorem, while we use it for only one part.

### *Kruskal’s Proof of Theorem 5.4.1*

The entire proof relies on an induction argument and the following two facts, which are quite easy to establish:

1. Linear combinations of normal random variables are normal.
2. Uncorrelated normal random variables are independent.

Recall that  $\bar{X}_m$  and  $S_m^2$  are the sample mean and variance based on the first  $m$  observations. Refer to parts (a), (b), and (c) of Theorem 5.4.1. To proceed with the induction, first assume that  $n = 2$ . We have

$$\bar{X}_2 = \frac{X_1 + X_2}{2}, \quad S_2^2 = \frac{(X_1 - X_2)^2}{2}.$$

Then part (b) follows immediately, part (c) follows from the definition of  $\chi_1^2$ , and part (a) follows by showing that  $\text{Cov}(X_1 + X_2, X_1 - X_2) = 0$  and invoking the two facts listed above and Theorem 4.6.5.

Continuing with the induction, we assume that the theorem is true for  $n = k$ , and prove it true for  $n = k + 1$ . We first establish algebraically that

$$(5.M.1) \quad \bar{X}_{k+1} = \frac{X_{k+1} + k\bar{X}_k}{k+1}$$

and

$$(5.M.2) \quad kS_{k+1}^2 = (k-1)S_k^2 + \left( \frac{k}{k+1} \right) (X_{k+1} - \bar{X}_k)^2.$$

Then we notice the following:

- $\bar{X}_n$  and  $X_{n+1}$  are independent random variables, so part (b) follows from the facts (1) and (2) on page 243 and the computation of  $E\bar{X}_{k+1}$  and  $\text{Var } \bar{X}_{k+1}$ .
- $X_{k+1} - \bar{X}_k \sim N(0, (k+1)\sigma^2/k)$ , thus  $(k/(k+1))(X_{k+1} - \bar{X}_k)^2/\sigma^2 \sim \chi_1^2$ . Furthermore,  $X_{k+1}$  is independent of  $S_k^2$ ,  $\bar{X}_k$  is independent of  $S_k^2$  (by the induction hypothesis), so part (c) follows by dividing  $kS_k^2$  by  $\sigma^2$  in (5.M.2).
- The induction hypothesis and inspection of (5.M.1) show that  $\bar{X}_{k+1}$  is independent of  $S_k^2$ , and part (a) follows from the two facts by showing that

$$\text{Cov}(\bar{X}_{k+1}, X_{k+1} - \bar{X}_k) = 0.$$

### The Bias of $S^2$

Most of the calculations that we have done in this chapter have assumed that the observations are independent, and calculations of some expectations have relied on this fact. H. A. David (1985) pointed out that, if the observations are dependent, then  $S^2$  may be a biased estimate of  $\sigma^2$ . That is, it may not happen that  $ES^2 = \sigma^2$ . However, the range of the possible bias is easily calculated. If  $X_1, \dots, X_n$  are random variables (not necessarily independent) with mean  $\mu$  and variance  $\sigma^2$ , then

$$(n-1)ES^2 = E \left( \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \right) = n\sigma^2 - n\text{Var } \bar{X}.$$

$\text{Var } \bar{X}$  can vary, according to the amount and type of dependence, from 0 (if all of the variables are constant) to  $\sigma^2$  (if all of the variables are copies of  $X_1$ ). Substituting these values in the above equation, we get the range of  $ES^2$  under dependence as

$$0 \leq ES^2 \leq \frac{n}{n-1}\sigma^2.$$

### Chebychev's Inequality Revisited

In Section 4.7 we looked at Chebychev's Inequality and in Example 4.7.3 we saw a particularly useful form. That form still requires knowledge of the mean and variance of a random variable and in some cases we might be interested in bounds using estimated values for the mean and variance.

If  $X_1, \dots, X_n$  is a random sample from a population with mean  $\mu$  and variance  $\sigma^2$ , Chebychev's Inequality says

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Saw et al. (1984) showed that if we substitute  $\bar{X}$  for  $\mu$  and  $S^2$  for  $\sigma^2$ , we obtain

$$P(|X - \bar{X}| \geq kS) \leq \frac{1}{n+1} g\left(\frac{n(n+1)k^2}{n-1+(n+1)k^2}\right),$$

where

$$g(t) = \begin{cases} \nu & \text{if } \nu \text{ is even} \\ \nu & \text{if } \nu \text{ is odd and } t < a, \\ \nu - 1 & \text{if } \nu \text{ is odd and } t > a \end{cases}$$

and

$$\nu = \text{largest integer } < \frac{n+1}{t}, \quad a = \frac{(n+1)(n+1-\nu)}{1+\nu(n+1-\nu)}.$$

# 6 Principles of Data Reduction

*"One forms provisional theories and waits for time or fuller knowledge to explode them. A bad habit, Mr. Ferguson, but human nature is weak."*

Sherlock Holmes

*The Adventure of the Sussex Vampire*

An experimenter uses the information in a sample  $X_1, \dots, X_n$  to make inferences about an unknown parameter  $\theta$ . If the sample size  $n$  is large, then the observed sample  $x_1, \dots, x_n$  is a long list of numbers that may be hard to interpret. An experimenter might wish to summarize the information in a sample by determining a few key features of the sample values. This is usually done by computing statistics, functions of the sample. For example, the sample mean, the sample variance, the largest observation, and the smallest observation are four statistics that might be used to summarize some key features of the sample. Recall that we use boldface letters to denote multiple variates, so  $\mathbf{X}$  denotes the random variables  $X_1, \dots, X_n$ , and  $\mathbf{x}$  denotes the sample  $x_1, \dots, x_n$ .

Any statistic,  $T(\mathbf{X})$ , defines a form of data reduction or data summary. An experimenter who uses only the observed value of the statistic,  $T(\mathbf{x})$ , rather than the entire observed sample,  $\mathbf{x}$ , will treat as equal two samples,  $\mathbf{x}$  and  $\mathbf{y}$ , that satisfy  $T(\mathbf{x}) = T(\mathbf{y})$  even though the actual sample values may be different in some ways.

Data reduction in terms of a particular statistic can be thought of as a partition of the sample space  $\mathcal{X}$ . Let  $\mathcal{T} = \{t: t = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$  be the image of  $\mathcal{X}$  under  $T(\mathbf{x})$ . Then  $T(\mathbf{x})$  partitions the sample space into sets  $A_t, t \in \mathcal{T}$ , defined by  $A_t = \{\mathbf{x}: T(\mathbf{x}) = t\}$ . The statistic summarizes the data in that, rather than reporting the entire sample  $\mathbf{x}$ , it reports only that  $T(\mathbf{x}) = t$  or, equivalently,  $\mathbf{x} \in A_t$ . For example, if  $T(\mathbf{x}) = x_1 + \dots + x_n$ , then  $T(\mathbf{x})$  does not report the actual sample values but only the sum. There may be many different sample points that have the same sum. The advantages and consequences of this type of data reduction are the topics of this chapter.

We study three principles of data reduction. We are interested in methods of data reduction that do not discard important information about the unknown parameter  $\theta$  and methods that successfully discard information that is irrelevant as far as gaining knowledge about  $\theta$  is concerned. The Sufficiency Principle promotes a method of data reduction that does not discard information about  $\theta$  while achieving some summarization of the data. The Likelihood Principle describes a function of the parameter, determined by the observed sample, that contains all the information about  $\theta$  that is available from the sample. The Invariance Principle prescribes yet another method of data reduction that still preserves some important features of the model.

## 6.1 The Sufficiency Principle

A *sufficient statistic* for a parameter  $\theta$  is a statistic that, in a certain sense, captures all the information about  $\theta$  contained in the sample. Any additional information in the sample, besides the value of the sufficient statistic, does not contain any more information about  $\theta$ . These considerations lead to the data reduction technique known as the Sufficiency Principle.

**SUFFICIENCY PRINCIPLE:** If  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$  then any inference about  $\theta$  should depend on the sample  $\mathbf{X}$  only through the value  $T(\mathbf{X})$ . That is, if  $\mathbf{x}$  and  $\mathbf{y}$  are two sample points such that  $T(\mathbf{x}) = T(\mathbf{y})$ , then the inference about  $\theta$  should be the same whether  $\mathbf{X} = \mathbf{x}$  or  $\mathbf{X} = \mathbf{y}$  is observed.

In this section we investigate some aspects of sufficient statistics and the Sufficiency Principle.

### 6.1.1 Sufficient Statistics

A sufficient statistic is formally defined in the following way.

**DEFINITION 6.1.1:** A statistic  $T(\mathbf{X})$  is a *sufficient statistic for  $\theta$*  if the conditional distribution of the sample  $\mathbf{X}$  given the value of  $T(\mathbf{X})$  does not depend on  $\theta$ .

If  $T(\mathbf{X})$  has a continuous distribution, then  $P_\theta(T(\mathbf{X}) = t) = 0$  for all values of  $t$ . A more sophisticated notion of conditional probability than that introduced in Chapter 1 is needed to fully understand Definition 6.1.1 in this case. A discussion of this can be found in more advanced texts such as Lehmann (1986). We will do our calculations in the discrete case and will point out analogous results that are true in the continuous case.

To understand Definition 6.1.1, let  $t$  be a possible value of  $T(\mathbf{X})$ , that is, a value such that  $P_\theta(T(\mathbf{X}) = t) > 0$ . We wish to consider the conditional probability  $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t)$ . If  $\mathbf{x}$  is a sample point such that  $T(\mathbf{x}) \neq t$ , then clearly  $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t) = 0$ . Thus, we are interested in  $P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$ . By the definition, if  $T(\mathbf{X})$  is a sufficient statistic this conditional probability is the same for all values of  $\theta$  so we have omitted the subscript.

A sufficient statistic captures all the information about  $\theta$  in this sense. Consider Experimenter 1, who observes  $\mathbf{X} = \mathbf{x}$  and, of course, can compute  $T(\mathbf{X}) = T(\mathbf{x})$ . To make an inference about  $\theta$  he can use the information that  $\mathbf{X} = \mathbf{x}$  and  $T(\mathbf{X}) = T(\mathbf{x})$ . Now consider Experimenter 2, who is not told the value of  $\mathbf{X}$  but only that  $T(\mathbf{X}) = T(\mathbf{x})$ . Experimenter 2 knows  $P(\mathbf{X} = \mathbf{y}|T(\mathbf{X}) = T(\mathbf{x}))$ , a probability distribution on  $A_{T(\mathbf{x})} = \{\mathbf{y} : T(\mathbf{y}) = T(\mathbf{x})\}$ , because this can be computed from the model without knowledge of the true value of  $\theta$ . Thus, Experimenter 2 can use this distribution and a randomization device, such as a random number table, to generate an observation  $\mathbf{Y}$  satisfying  $P(\mathbf{Y} = \mathbf{y}|T(\mathbf{X}) = T(\mathbf{x})) = P(\mathbf{X} = \mathbf{y}|T(\mathbf{X}) = T(\mathbf{x}))$ . It turns out that, for each value of  $\theta$ ,  $\mathbf{X}$  and  $\mathbf{Y}$  have the same unconditional probability distribution, as we shall see below. So Experimenter 1, who knows  $\mathbf{X}$ , and Experimenter 2, who knows  $\mathbf{Y}$ , have equivalent information about  $\theta$ . But surely the use of the random

number table to generate  $\mathbf{Y}$  has not added to Experimenter 2's knowledge of  $\theta$ . All his knowledge about  $\theta$  is contained in the knowledge that  $T(\mathbf{X}) = T(\mathbf{x})$ . So Experimenter 2, who knows only  $T(\mathbf{X}) = T(\mathbf{x})$ , has just as much information about  $\theta$  as does Experimenter 1, who knows the entire sample  $\mathbf{X} = \mathbf{x}$ .

To complete the above argument, we need to show that  $\mathbf{X}$  and  $\mathbf{Y}$  have the same unconditional distribution, that is,  $P_\theta(\mathbf{X} = \mathbf{x}) = P_\theta(\mathbf{Y} = \mathbf{x})$  for all  $\mathbf{x}$  and  $\theta$ . Note that the events  $\{\mathbf{X} = \mathbf{x}\}$  and  $\{\mathbf{Y} = \mathbf{x}\}$  are both subsets of the event  $\{T(\mathbf{X}) = T(\mathbf{x})\}$ . Also recall that

$$P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) = P(\mathbf{Y} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$$

and these conditional probabilities do not depend on  $\theta$ . Thus we have

$$\begin{aligned} P_\theta(\mathbf{X} = \mathbf{x}) &= P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})) \\ &= P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))P_\theta(T(\mathbf{X}) = T(\mathbf{x})) \quad \left( \begin{array}{l} \text{definition of} \\ \text{conditional probability} \end{array} \right) \\ &= P(\mathbf{Y} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))P_\theta(T(\mathbf{X}) = T(\mathbf{x})) \\ &= P_\theta(\mathbf{Y} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})) \\ &= P_\theta(\mathbf{Y} = \mathbf{x}). \end{aligned}$$

To use Definition 6.1.1 to verify that a statistic  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$ , we must verify that for any fixed values of  $\mathbf{x}$  and  $t$ , the conditional probability  $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t)$  is the same for all values of  $\theta$ . Now, this probability is 0 for all values of  $\theta$  if  $T(\mathbf{x}) \neq t$ . So, we must verify only that  $P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$  does not depend on  $\theta$ . But since  $\{\mathbf{X} = \mathbf{x}\}$  is a subset of  $\{T(\mathbf{X}) = T(\mathbf{x})\}$ ,

$$\begin{aligned} P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) &= \frac{P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x}))}{P_\theta(T(\mathbf{X}) = T(\mathbf{x}))} \\ &= \frac{P_\theta(\mathbf{X} = \mathbf{x})}{P_\theta(T(\mathbf{X}) = T(\mathbf{x}))} \\ &= \frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} \end{aligned}$$

where  $p(\mathbf{x}|\theta)$  is the joint pmf of the sample  $\mathbf{X}$  and  $q(t|\theta)$  is the pmf of  $T(\mathbf{X})$ . Thus,  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$  if, and only if, for every  $\mathbf{x}$  the above ratio of pmfs is constant as a function of  $\theta$ . If  $\mathbf{X}$  and  $T(\mathbf{X})$  have continuous distributions, then the above conditional probabilities cannot be interpreted in the sense of Chapter 1. But it is still appropriate to use the above criterion to determine if  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$ .

**THEOREM 6.1.1:** If  $p(\mathbf{x}|\theta)$  is the joint pdf or pmf of  $\mathbf{X}$ , and  $q(t|\theta)$  is the pdf or pmf of  $T(\mathbf{X})$ , then  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$  if, and only if, for every  $\mathbf{x}$  in the sample space the ratio  $p(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$  is constant as a function of  $\theta$ .  $\square$

We now use Theorem 6.1.1 to verify that certain common statistics are sufficient statistics.

**Example 6.1.1:** Let  $X_1, \dots, X_n$  be iid Bernoulli random variables with parameter  $\theta$ ,  $0 < \theta < 1$ . We will show that  $T(\mathbf{X}) = X_1 + \dots + X_n$  is a sufficient statistic for  $\theta$ . Note that  $T(\mathbf{X})$  counts the number of  $X_i$ 's that equal 1, so  $T(\mathbf{X})$  has a binomial( $n, \theta$ ) distribution. The ratio of pmfs is thus

$$\begin{aligned} \frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{\prod \theta^{x_i} (1-\theta)^{1-x_i}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} && (\text{define } t = \sum x_i) \\ &= \frac{\theta^{\sum x_i} (1-\theta)^{\sum(1-x_i)}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} && (\prod \theta^{x_i} = \theta^{\sum x_i}) \\ &= \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} \\ &= \frac{1}{\binom{n}{t}} \\ &= \frac{1}{\binom{n}{\sum x_i}}. \end{aligned}$$

Since this ratio does not depend on  $\theta$ , by Theorem 6.1.1,  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$ . The interpretation is this: The total number of ones in this Bernoulli sample contains all the information about  $\theta$  that is in the data. Other features of the data, such as the exact value of  $X_3$ , contain no additional information. ||

**Example 6.1.2:** Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$ , where  $\sigma^2$  is known. We wish to show that the sample mean,  $T(\mathbf{X}) = \bar{X} = (X_1 + \dots + X_n)/n$ , is a sufficient statistic for  $\mu$ . The joint pdf of the sample  $\mathbf{X}$  is

$$\begin{aligned} f(\mathbf{x}|\mu) &= \prod_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp(-(x_i - \mu)^2/(2\sigma^2)) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^n (x_i - \mu)^2/(2\sigma^2)\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu)^2/(2\sigma^2)\right) && (\text{add and subtract } \bar{x}) \\ (6.1.1) \quad &= (2\pi\sigma^2)^{-n/2} \exp\left(-\left(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)/(2\sigma^2)\right). \end{aligned}$$

The last equality is true because the cross-product term  $\sum_{i=1}^n (x_i - \bar{x})(\bar{x} - \mu)$  may be rewritten as  $(\bar{x} - \mu)\sum_{i=1}^n (x_i - \bar{x})$ , and  $\sum_{i=1}^n (x_i - \bar{x}) = 0$ . Recall that the sample

mean  $\bar{X}$  has a  $n(\mu, \sigma^2/n)$  distribution. Thus, the ratio of pdfs is

$$\begin{aligned} \frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{(2\pi\sigma^2)^{-n/2} \exp\left(-\left(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)/(2\sigma^2)\right)}{(2\pi\sigma^2/n)^{-1/2} \exp(-n(\bar{x} - \mu)^2/(2\sigma^2))} \\ &= n^{-1/2} (2\pi\sigma^2)^{-(n-1)/2} \exp\left(-\sum_{i=1}^n (x_i - \bar{x})^2/(2\sigma^2)\right), \end{aligned}$$

which does not depend on  $\mu$ . By Theorem 6.1.1, the sample mean is a sufficient statistic for  $\mu$ . ||

It may be unwieldy to use the definition of a sufficient statistic to find a sufficient statistic for a particular model. To use the definition, we must guess a statistic  $T(\mathbf{X})$  to be sufficient, find the pmf or pdf of  $T(\mathbf{X})$ , and check that the ratio of pdfs or pmfs does not depend on  $\theta$ . The first step requires a good deal of intuition and the second sometimes requires some tedious analysis. Fortunately, the next theorem, due to Halmos and Savage (1949), allows us to find a sufficient statistic by simple inspection of the pdf or pmf of the sample.

**THEOREM 6.1.2 (Factorization Theorem):** Let  $f(\mathbf{x}|\theta)$  denote the joint pdf or pmf of a sample  $\mathbf{X}$ . A statistic  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$  if and only if there exist functions  $g(t|\theta)$  and  $h(\mathbf{x})$  such that, for all sample points  $\mathbf{x}$  and all parameter points  $\theta$ ,

$$(6.1.2) \quad f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}).$$

*Proof:* We give the proof only for discrete distributions.

Suppose  $T(\mathbf{X})$  is a sufficient statistic. Choose  $g(t|\theta) = P_\theta(T(\mathbf{X}) = t)$  and  $h(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$ . Because  $T(\mathbf{X})$  is sufficient, the conditional probability defining  $h(\mathbf{x})$  does not depend on  $\theta$ . Thus this choice of  $h(\mathbf{x})$  and  $g(t|\theta)$  is legitimate, and for this choice we have

$$\begin{aligned} f(\mathbf{x}|\theta) &= P_\theta(\mathbf{X} = \mathbf{x}) \\ &= P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})) \\ &= P_\theta(T(\mathbf{X}) = T(\mathbf{x}))P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) \quad (\text{sufficiency}) \\ &= g(T(\mathbf{x})|\theta)h(\mathbf{x}). \end{aligned}$$

So factorization (6.1.2) has been exhibited. We also see from the last two lines above that

$$P_\theta(T(\mathbf{X}) = T(\mathbf{x})) = g(T(\mathbf{x})|\theta),$$

so  $g(T(\mathbf{x})|\theta)$  is the pmf of  $T(\mathbf{X})$ .

Now assume the factorization (6.1.2) exists. Let  $q(t|\theta)$  be the pmf of  $T(\mathbf{X})$ . To show that  $T(\mathbf{X})$  is sufficient we examine the ratio  $f(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$ . Define  $A_{T(\mathbf{x})} = \{\mathbf{y} : T(\mathbf{y}) = T(\mathbf{x})\}$ . Then

$$\begin{aligned}\frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{q(T(\mathbf{x})|\theta)} && \text{(since (6.1.2) is satisfied)} \\ &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sum_{A_{T(\mathbf{x})}} g(T(\mathbf{y})|\theta)h(\mathbf{y})} && \text{(definition of the pmf of } T\text{)} \\ &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{g(T(\mathbf{x})|\theta)\sum_{A_{T(\mathbf{x})}} h(\mathbf{y})} && \text{(since } T\text{ is constant on } A_{T(\mathbf{x})}\text{)} \\ &= \frac{h(\mathbf{x})}{\sum_{A_{T(\mathbf{x})}} h(\mathbf{y})}.\end{aligned}$$

Since the ratio does not depend on  $\theta$ , by Theorem 6.1.1,  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$ .  $\square$

To use the Factorization Theorem to find a sufficient statistic, we factor the joint pdf of the sample into two parts, with one part not depending on  $\theta$ . The part that does not depend on  $\theta$  constitutes the  $h(\mathbf{x})$  function. The other part, the one that depends on  $\theta$ , usually depends on the sample  $\mathbf{x}$  only through some function  $T(\mathbf{x})$  and this function is a sufficient statistic for  $\theta$ . This is illustrated in the following example.

**Example 6.1.2 (Continued):** For the normal model described earlier, we saw that the pdf could be factored as

$$(6.1.3) \quad f(\mathbf{x}|\mu) = (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^n (x_i - \bar{x})^2/(2\sigma^2)\right) \exp(-n(\bar{x} - \mu)^2/(2\sigma^2)).$$

We can define

$$h(\mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^n (x_i - \bar{x})^2/(2\sigma^2)\right),$$

which does not depend on the unknown parameter  $\mu$ . The factor in (6.1.3) that contains  $\mu$  depends on the sample  $\mathbf{x}$  only through the function  $T(\mathbf{x}) = \bar{x}$ , the sample mean. So we have

$$g(t|\mu) = \exp(-n(t - \mu)^2/(2\sigma^2))$$

and note that

$$f(\mathbf{x}|\mu) = h(\mathbf{x})g(T(\mathbf{x})|\mu).$$

Thus, by the Factorization Theorem,  $T(\mathbf{X}) = \bar{X}$  is a sufficient statistic for  $\mu$ . ||

The Factorization Theorem requires that the equality  $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$  hold for all  $\mathbf{x}$  and  $\theta$ . If the set of  $\mathbf{x}$  on which  $f(\mathbf{x}|\theta)$  is positive depends on  $\theta$ , care must be taken in the definition of  $h$  and  $g$  to ensure that the product is 0 where  $f$  is 0. Of course, correct definition of  $h$  and  $g$  makes the sufficient statistic evident, as the next example illustrates.

**Example 6.1.3:** Let  $X_1, \dots, X_n$  be iid observations from the discrete uniform distribution on  $1, \dots, \theta$ . That is, the unknown parameter,  $\theta$ , is a positive integer and the pmf of  $X_i$  is

$$f(x|\theta) = \begin{cases} \theta^{-1} & x = 1, \dots, \theta \\ 0 & \text{otherwise} \end{cases}.$$

Thus the joint pmf of  $X_1, \dots, X_n$  is

$$f(\mathbf{x}|\theta) = \begin{cases} \theta^{-n} & x_i \in \{1, \dots, \theta\} \text{ for } i = 1, \dots, n \\ 0 & \text{otherwise} \end{cases}.$$

The restriction " $x_i \in \{1, \dots, \theta\}$  for  $i = 1, \dots, n$ " can be re-expressed as " $x_i \in \{1, 2, \dots\}$  for  $i = 1, \dots, n$  (note that there is no  $\theta$  in this restriction) and  $\max_i x_i \leq \theta$ ." If we define  $T(\mathbf{x}) = \max_i x_i$ ,

$$h(\mathbf{x}) = \begin{cases} 1 & x_i \in \{1, 2, \dots\} \text{ for } i = 1, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

and

$$g(t|\theta) = \begin{cases} \theta^{-n} & t \leq \theta \\ 0 & \text{otherwise} \end{cases},$$

it is easily verified that  $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$  for all  $\mathbf{x}$  and  $\theta$ . Thus, the largest order statistic,  $T(\mathbf{X}) = \max_i X_i$ , is a sufficient statistic in this problem.

This type of analysis can sometimes be carried out more clearly and concisely using indicator functions. Recall that  $I_A(x)$  is the indicator function of the set  $A$ , that is, it is equal to 1 if  $x \in A$  and equal to 0 otherwise. Let  $\mathcal{N} = \{1, 2, \dots\}$  be the set of positive integers and let  $\mathcal{N}_\theta = \{1, 2, \dots, \theta\}$ . Then the joint pmf of  $X_1, \dots, X_n$  is

$$\begin{aligned} f(\mathbf{x}|\theta) &= \prod_{i=1}^n \theta^{-1} I_{\mathcal{N}_\theta}(x_i) \\ &= \theta^{-n} \prod_{i=1}^n I_{\mathcal{N}_\theta}(x_i). \end{aligned}$$

Defining  $T(\mathbf{x}) = \max_i x_i$ , it is clear that

$$\prod_{i=1}^n I_{N_\theta}(x_i) = \left( \prod_{i=1}^n I_N(x_i) \right) I_{N_\theta}(T(\mathbf{x})).$$

Thus we have the factorization

$$f(\mathbf{x}|\theta) = \theta^{-n} I_{N_\theta}(T(\mathbf{x})) \left( \prod_{i=1}^n I_N(x_i) \right).$$

The first factor depends on  $x_1, \dots, x_n$  only through the value of  $T(\mathbf{x}) = \max_i x_i$  and the second factor does not depend on  $\theta$ . By the Factorization Theorem,  $T(\mathbf{X}) = \max_i X_i$  is a sufficient statistic for  $\theta$ . ||

In all the previous examples, the sufficient statistic is a real-valued function of the sample. All the information about  $\theta$  in the sample  $\mathbf{x}$  is summarized in the single number  $T(\mathbf{x})$ . Sometimes, the information cannot be summarized in one number and several numbers are required instead. In such cases, a sufficient statistic is a vector, say  $T(\mathbf{X}) = (T_1(\mathbf{X}), \dots, T_r(\mathbf{X}))$ . This situation often occurs when the parameter is also a vector, say  $\theta = (\theta_1, \dots, \theta_s)$ , and it is usually the case that the sufficient statistic and the parameter vectors are of equal length, that is,  $r = s$ . Different combinations of lengths are possible, however, as the exercises and Examples 6.1.7, 6.1.9, and 6.1.11 illustrate. The Factorization Theorem may be used to find a vector-valued sufficient statistic, as in Example 6.1.4.

**Example 6.1.4:** Again assume that  $X_1, \dots, X_n$  are iid  $n(\mu, \sigma^2)$  but, unlike Example 6.1.2, assume that both  $\mu$  and  $\sigma^2$  are unknown so the parameter vector is  $\theta = (\mu, \sigma^2)$ . Now when using the Factorization Theorem, any part of the joint pdf that depends on either  $\mu$  or  $\sigma^2$  must be included in the  $g$  function. From (6.1.1) it is clear that the pdf depends on the sample  $\mathbf{x}$  only through the two values  $T_1(\mathbf{x}) = \bar{x}$  and  $T_2(\mathbf{x}) = s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)$ . Thus we can define  $h(\mathbf{x}) = 1$  and

$$\begin{aligned} g(t|\theta) &= g(t_1, t_2|\mu, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp(- (n(t_1 - \mu)^2 + (n-1)t_2) / (2\sigma^2)). \end{aligned}$$

Then it can be seen that

$$(6.1.4) \quad f(\mathbf{x}|\mu, \sigma^2) = g(T_1(\mathbf{x}), T_2(\mathbf{x})|\mu, \sigma^2)h(\mathbf{x}).$$

Thus, by the Factorization Theorem,  $T(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) = (\bar{X}, S^2)$  is a sufficient statistic for  $(\mu, \sigma^2)$  in this normal model. ||

Example 6.1.4 demonstrates that, for the normal model, the common practice of summarizing a data set by reporting only the sample mean and variance is justified.

The sufficient statistic  $(\bar{X}, S^2)$  contains all the information about  $(\mu, \sigma^2)$  that is available in the sample. The experimenter should remember, however, that the definition of a sufficient statistic is model-dependent. For another model, that is, another family of densities, the sample mean and variance may not be a sufficient statistic for the population mean and variance. The experimenter who calculates only  $\bar{X}$  and  $S^2$  and totally ignores the rest of the data would be placing strong faith in the normal model assumption.

It is easy to find a sufficient statistic for an exponential family of distributions using the Factorization Theorem. The proof of the following important result is left as Exercise 6.4.

**THEOREM 6.1.3:** Let  $X_1, \dots, X_n$  be iid observations from a pdf or pmf,  $f(x|\theta)$ . Suppose  $f(x|\theta)$  belongs to an exponential family given by

$$f(x|\theta) = h(x)c(\theta) \exp \left( \sum_{i=1}^k w_i(\theta)t_i(x) \right).$$

Then

$$T(\mathbf{X}) = \left( \sum_{j=1}^n t_1(X_j), \dots, \sum_{j=1}^n t_k(X_j) \right)$$

is a sufficient statistic for  $\theta$ . □

### 6.1.2 Minimal Sufficient Statistics

In the preceding section we found one sufficient statistic for each model considered. In any problem there are, in fact, many sufficient statistics.

It is always true that the complete sample,  $\mathbf{X}$ , is a sufficient statistic. We can factor the pdf or pmf of  $\mathbf{X}$  as  $f(\mathbf{x}|\theta) = f(T(\mathbf{x})|\theta)h(\mathbf{x})$  where  $T(\mathbf{x}) = \mathbf{x}$  and  $h(\mathbf{x}) = 1$  for all  $\mathbf{x}$ . By the Factorization Theorem,  $T(\mathbf{X}) = \mathbf{X}$  is a sufficient statistic.

Also, it follows that any one-to-one function of a sufficient statistic is a sufficient statistic. Suppose  $T(\mathbf{X})$  is a sufficient statistic and define  $T^*(\mathbf{x}) = r(T(\mathbf{x}))$  for all  $\mathbf{x}$  where  $r$  is a one-to-one function with inverse  $r^{-1}$ . Then by the Factorization Theorem there exist  $g$  and  $h$  such that

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}) = g(r^{-1}(T^*(\mathbf{x}))|\theta)h(\mathbf{x}).$$

Defining  $g^*(t|\theta) = g(r^{-1}(t)|\theta)$ , we see that

$$f(\mathbf{x}|\theta) = g^*(T^*(\mathbf{x})|\theta)h(\mathbf{x}).$$

So, by the Factorization Theorem,  $T^*(\mathbf{X})$  is a sufficient statistic.

Because of the numerous sufficient statistics in a problem, we might ask whether one sufficient statistic is any better than another. Recall that the purpose of a sufficient statistic is to achieve data reduction without loss of information about the parameter  $\theta$ ; thus, a statistic that achieves the most data reduction while still retaining all the information about  $\theta$  might be considered preferable. The definition of such a statistic is formalized now.

**DEFINITION 6.1.2:** A sufficient statistic  $T(\mathbf{X})$  is called a *minimal sufficient statistic* if, for any other sufficient statistic  $T'(\mathbf{X})$ ,  $T(\mathbf{x})$  is a function of  $T'(\mathbf{x})$ .

To say that  $T(\mathbf{x})$  is a function of  $T'(\mathbf{x})$  simply means that if  $T'(\mathbf{x}) = T'(\mathbf{y})$  then  $T(\mathbf{x}) = T(\mathbf{y})$ . In terms of the partition sets described at the beginning of the chapter, if  $\{B_{t'} : t' \in T'\}$  are partition sets for  $T'(\mathbf{x})$ , and  $\{A_t : t \in T\}$  are the partition sets for  $T(\mathbf{x})$ , then Definition 6.1.2 states that every  $B_{t'}$  is a subset of some  $A_t$ . Thus, the partition associated with a minimal sufficient statistic is the *coarsest* possible partition for a sufficient statistic and a minimal sufficient statistic achieves the greatest possible data reduction for a sufficient statistic.

**Example 6.1.5:** Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$ ,  $\sigma^2$  known, the model considered in Example 6.1.2. Using factorization (6.1.3), we concluded that  $T(\mathbf{X}) = \bar{X}$  is a sufficient statistic for  $\mu$ . Instead, we could write down factorization (6.1.4) for this problem ( $\sigma^2$  is a known value now) and correctly conclude that  $T'(\mathbf{X}) = (\bar{X}, S^2)$  is a sufficient statistic for  $\mu$  in this problem. Clearly  $T(\mathbf{X})$  achieves a greater data reduction than  $T'(\mathbf{X})$  since we do not know the sample variance if we know only  $T(\mathbf{X})$ . We can write  $T(\mathbf{x})$  as a function of  $T'(\mathbf{x})$  by defining the function  $r(a, b) = a$ . Then  $T(\mathbf{x}) = \bar{x} = r(\bar{x}, s^2) = r(T'(\mathbf{x}))$ . Since  $T(\mathbf{X})$  and  $T'(\mathbf{X})$  are both sufficient statistics, they both contain the same information about  $\mu$ . Thus, the additional information about the value of  $S^2$ , the sample variance, does not add to our knowledge of  $\mu$  since the population variance  $\sigma^2$  is known. Of course, if  $\sigma^2$  is unknown, as in Example 6.1.4,  $T(\mathbf{X}) = \bar{X}$  is not a sufficient statistic and  $T'(\mathbf{X})$  contains more information about the parameter  $(\mu, \sigma^2)$  than does  $T(\mathbf{X})$ . ||

Using Definition 6.1.2 to find a minimal sufficient statistic is impractical, as was using Definition 6.1.1 to find sufficient statistics. We would need to guess that  $T(\mathbf{X})$  was a minimal sufficient statistic and then verify the condition in the definition. (Note that we did not show that  $\bar{X}$  is a minimal sufficient statistic in Example 6.1.5.) Fortunately, the following result of Lehmann and Scheffé (1950, Theorem 6.3) gives an easier way to find a minimal sufficient statistic.

**THEOREM 6.1.4:** Let  $f(\mathbf{x}|\theta)$  be the pmf or pdf of a sample  $\mathbf{X}$ . Suppose there exists a function  $T(\mathbf{x})$  such that, for two sample points  $\mathbf{x}$  and  $\mathbf{y}$ , the ratio  $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$  is constant as a function of  $\theta$  if and only if  $T(\mathbf{x}) = T(\mathbf{y})$ . Then  $T(\mathbf{X})$  is a minimal sufficient statistic for  $\theta$ .

*Proof:* To simplify the proof, we assume  $f(\mathbf{x}|\theta) > 0$  for all  $\mathbf{x} \in \mathcal{X}$  and  $\theta$ .

First we show that  $T(\mathbf{X})$  is a sufficient statistic. Let  $T = \{t : t = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$  be the image of  $\mathcal{X}$  under  $T(\mathbf{x})$ . Define the partition sets induced by  $T(\mathbf{x})$  as

$A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$ . For each  $A_t$ , choose and fix one element  $\mathbf{x}_t \in A_t$ . For any  $\mathbf{x} \in \mathcal{X}$ ,  $\mathbf{x}_{T(\mathbf{x})}$  is the fixed element that is in the same set,  $A_t$ , as  $\mathbf{x}$ . Since  $\mathbf{x}$  and  $\mathbf{x}_{T(\mathbf{x})}$  are in the same set  $A_t$ ,  $T(\mathbf{x}) = T(\mathbf{x}_{T(\mathbf{x})})$  and, hence,  $f(\mathbf{x}|\theta)/f(\mathbf{x}_{T(\mathbf{x})}|\theta)$  is constant as a function of  $\theta$ . Thus, we can define a function on  $\mathcal{X}$  by  $h(\mathbf{x}) = f(\mathbf{x}|\theta)/f(\mathbf{x}_{T(\mathbf{x})}|\theta)$  and  $h$  does not depend on  $\theta$ . Define a function on  $\mathcal{T}$  by  $g(t|\theta) = f(\mathbf{x}_t|\theta)$ . Then it can be seen that

$$f(\mathbf{x}|\theta) = \frac{f(\mathbf{x}_{T(\mathbf{x})}|\theta)f(\mathbf{x}|\theta)}{f(\mathbf{x}_{T(\mathbf{x})}|\theta)} = g(T(\mathbf{x})|\theta)h(\mathbf{x})$$

and, by the Factorization Theorem,  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$ .

Now to show that  $T(\mathbf{X})$  is minimal, let  $T'(\mathbf{X})$  be any other sufficient statistic. By the Factorization Theorem, there exist functions  $g'$  and  $h'$  such that  $f(\mathbf{x}|\theta) = g'(T'(\mathbf{x})|\theta)h'(\mathbf{x})$ . Let  $\mathbf{x}$  and  $\mathbf{y}$  be any two sample points with  $T'(\mathbf{x}) = T'(\mathbf{y})$ . Then

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{g'(T'(\mathbf{x})|\theta)h'(\mathbf{x})}{g'(T'(\mathbf{y})|\theta)h'(\mathbf{y})} = \frac{h'(\mathbf{x})}{h'(\mathbf{y})}.$$

Since this ratio does not depend on  $\theta$ , the assumptions of the theorem imply that  $T(\mathbf{x}) = T(\mathbf{y})$ . Thus,  $T(\mathbf{x})$  is a function of  $T'(\mathbf{x})$  and  $T(\mathbf{x})$  is minimal.  $\square$

**Example 6.1.6:** Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$ , both  $\mu$  and  $\sigma^2$  unknown. Let  $\mathbf{x}$  and  $\mathbf{y}$  denote two sample points and let  $(\bar{x}, s_x^2)$  and  $(\bar{y}, s_y^2)$  be the sample means and variances corresponding to the  $\mathbf{x}$  and  $\mathbf{y}$  samples, respectively. Then, using (6.1.4), we see that the ratio of densities is

$$\begin{aligned} \frac{f(\mathbf{x}|\mu, \sigma^2)}{f(\mathbf{y}|\mu, \sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2} \exp(-[n(\bar{x} - \mu)^2 + (n-1)s_x^2]/(2\sigma^2))}{(2\pi\sigma^2)^{-n/2} \exp(-[n(\bar{y} - \mu)^2 + (n-1)s_y^2]/(2\sigma^2))} \\ &= \exp([ -n(\bar{x}^2 - \bar{y}^2) + 2n\mu(\bar{x} - \bar{y}) - (n-1)(s_x^2 - s_y^2) ]/(2\sigma^2)). \end{aligned}$$

This ratio will be constant as a function of  $\mu$  and  $\sigma^2$  if and only if  $\bar{x} = \bar{y}$  and  $s_x^2 = s_y^2$ . Thus, by Theorem 6.1.4,  $(\bar{X}, S^2)$  is a minimal sufficient statistic for  $(\mu, \sigma^2)$ .  $\parallel$

If the set of  $\mathbf{x}$ s on which the pdf or pmf is positive depends on the parameter  $\theta$ , then, for the ratio in Theorem 6.1.4 to be constant as a function of  $\theta$ , the numerator and denominator must be positive for exactly the same values of  $\theta$ . This restriction is usually reflected in a minimal sufficient statistic, as the next example illustrates.

**Example 6.1.7:** Suppose  $X_1, \dots, X_n$  are iid uniform observations on the interval  $(\theta, \theta + 1)$ ,  $-\infty < \theta < \infty$ . Then the joint pdf of  $\mathbf{X}$  is

$$f(\mathbf{x}|\theta) = \begin{cases} 1 & \theta < x_i < \theta + 1, i = 1, \dots, n \\ 0 & \text{otherwise} \end{cases},$$

which can be written as

$$f(\mathbf{x}|\theta) = \begin{cases} 1 & \max_i x_i - 1 < \theta < \min_i x_i \\ 0 & \text{otherwise} \end{cases}$$

Thus, for two sample points  $\mathbf{x}$  and  $\mathbf{y}$ , the numerator and denominator of the ratio  $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$  will be positive for the same values of  $\theta$  if and only if  $\min_i x_i = \min_i y_i$  and  $\max_i x_i = \max_i y_i$ . And, if the minima and maxima are equal, then the ratio is constant and, in fact, equals 1. Thus, letting  $X_{(1)} = \min_i X_i$  and  $X_{(n)} = \max_i X_i$ , we have that  $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$  is a minimal sufficient statistic. This is a case in which the dimension of a minimal sufficient statistic does not match the dimension of the parameter. ||

A minimal sufficient statistic is not unique. Any one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic. So, for example,  $T'(\mathbf{X}) = (X_{(n)} - X_{(1)}, (X_{(n)} + X_{(1)})/2)$  is also a minimal sufficient statistic in Example 6.1.7 and  $T'(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$  is also a minimal sufficient statistic in Example 6.1.6.

### 6.1.3 Ancillary Statistics

In the preceding sections, we considered sufficient statistics. Such statistics, in a sense, contain all the information about  $\theta$  that is available in the sample. In this section we introduce a different sort of statistic, one that has a complementary purpose.

**DEFINITION 6.1.3:** A statistic  $S(\mathbf{X})$  whose distribution does not depend on the parameter  $\theta$  is called an *ancillary statistic*.

Alone, an ancillary statistic contains no information about  $\theta$ . An ancillary statistic is an observation on a random variable whose distribution is fixed and known, unrelated to  $\theta$ . Paradoxically, an ancillary statistic, when used in conjunction with other statistics, sometimes does contain valuable information for inferences about  $\theta$ . We will investigate this behavior in the next section. For now, we just give some examples of ancillary statistics.

**Example 6.1.8:** As in Example 6.1.7, let  $X_1, \dots, X_n$  be iid uniform observations on the interval  $(\theta, \theta + 1)$ ,  $-\infty < \theta < \infty$ . Let  $X_{(1)} < \dots < X_{(n)}$  be the order statistics from the sample. We show below that the range statistic,  $R = X_{(n)} - X_{(1)}$ , is an ancillary statistic by showing that the pdf of  $R$  does not depend on  $\theta$ . Recall that the cdf of each  $X_i$  is

$$F(x|\theta) = \begin{cases} 0 & x \leq \theta \\ x - \theta & \theta < x < \theta + 1 \\ 1 & \theta + 1 \leq x \end{cases}$$

Thus, the joint pdf of  $X_{(1)}$  and  $X_{(n)}$ , as given by (5.5.7), is

$$g(x_{(1)}, x_{(n)}|\theta) = \begin{cases} n(n-1)(x_{(n)} - x_{(1)})^{n-2} & \theta < x_{(1)} < x_{(n)} < \theta + 1 \\ 0 & \text{otherwise} \end{cases}$$

Making the transformation  $R = X_{(n)} - X_{(1)}$  and  $M = (X_{(1)} + X_{(n)})/2$ , which has the inverse transformation  $X_{(1)} = (2M - R)/2$  and  $X_{(n)} = (2M + R)/2$  with Jacobian 1, we see that the joint pdf of  $R$  and  $M$  is

$$h(r, m|\theta) = \begin{cases} n(n-1)r^{n-2} & 0 < r < 1, \theta + (r/2) < m < \theta + 1 - (r/2) \\ 0 & \text{otherwise} \end{cases}$$

(Notice the rather involved region of positivity for  $h(r, m|\theta)$ .) Thus, the pdf for  $R$  is

$$\begin{aligned} h(r|\theta) &= \int_{\theta+(r/2)}^{\theta+1-(r/2)} n(n-1)r^{n-2} dm \\ &= n(n-1)r^{n-2}(1-r), \quad 0 < r < 1. \end{aligned}$$

This is a beta pdf with  $\alpha = n - 1$  and  $\beta = 2$ . More importantly, the pdf is the same for all  $\theta$ . Thus, the distribution of  $R$  does not depend on  $\theta$  and  $R$  is ancillary. ||

In Example 6.1.8 the range statistic is ancillary because the model considered there is a location parameter model. The ancillarity of  $R$  does not depend on the uniformity of the  $X_i$ s, but rather on the parameter of the distribution being a location parameter. We now consider the general location parameter model.

**Example 6.1.9:** Let  $X_1, \dots, X_n$  be iid observations from a location parameter family with cdf  $F(x - \theta)$ ,  $-\infty < \theta < \infty$ . We will show that the range,  $R = X_{(n)} - X_{(1)}$ , is an ancillary statistic. We use Theorem 3.4.2 and work with  $Z_1, \dots, Z_n$ , iid observations from  $F(x)$  (corresponding to  $\theta = 0$ ) with  $X_1 = Z_1 + \theta, \dots, X_n = Z_n + \theta$ . Thus the cdf of the range statistic,  $R$ , is

$$\begin{aligned} F_R(r|\theta) &= P_\theta(R \leq r) \\ &= P_\theta(\max_i X_i - \min_i X_i \leq r) \\ &= P_\theta(\max_i (Z_i + \theta) - \min_i (Z_i + \theta) \leq r) \\ &= P_\theta(\max_i Z_i - \min_i Z_i + \theta - \theta \leq r) \\ &= P_\theta(\max_i Z_i - \min_i Z_i \leq r). \end{aligned}$$

The last probability does not depend on  $\theta$  because the distribution of  $Z_1, \dots, Z_n$  does not depend on  $\theta$ . Thus, the cdf of  $R$  does not depend on  $\theta$  and, hence,  $R$  is an ancillary statistic. ||

**Example 6.1.10:** Scale parameter families also have certain kinds of ancillary statistics. Let  $X_1, \dots, X_n$  be iid observations from a scale parameter family with cdf  $F(x/\sigma)$ ,  $\sigma > 0$ . Then any statistic that depends on the sample only through the  $n - 1$  values  $X_1/X_n, \dots, X_{n-1}/X_n$  is an ancillary statistic. For example,

$$\frac{X_1 + \cdots + X_n}{X_n} = \frac{X_1}{X_n} + \cdots + \frac{X_{n-1}}{X_n} + 1$$

is an ancillary statistic. To see this fact, let  $Z_1, \dots, Z_n$  be iid observations from  $F(x)$  (corresponding to  $\sigma = 1$ ) with  $X_i = \sigma Z_i$ . The joint cdf of  $X_1/X_n, \dots, X_{n-1}/X_n$  is

$$\begin{aligned} F(y_1, \dots, y_{n-1} | \sigma) &= P_\sigma(X_1/X_n \leq y_1, \dots, X_{n-1}/X_n \leq y_{n-1}) \\ &= P_\sigma(\sigma Z_1/(\sigma Z_n) \leq y_1, \dots, \sigma Z_{n-1}/(\sigma Z_n) \leq y_{n-1}) \\ &= P_\sigma(Z_1/Z_n \leq y_1, \dots, Z_{n-1}/Z_n \leq y_{n-1}). \end{aligned}$$

The last probability does not depend on  $\sigma$  because the distribution of  $Z_1, \dots, Z_n$  does not depend on  $\sigma$ . So the distribution of  $X_1/X_n, \dots, X_{n-1}/X_n$  is independent of  $\sigma$ , as is the distribution of any function of these quantities.

In particular, let  $X_1$  and  $X_2$  be iid  $n(0, \sigma^2)$  observations. From the above result, we see that  $X_1/X_2$  has a distribution which is the same for every value of  $\sigma$ . But, in Example 4.3.4, we saw that, if  $\sigma = 1$ ,  $X_1/X_2$  has a Cauchy(0, 1) distribution. Thus, for any  $\sigma > 0$ , the distribution of  $X_1/X_2$  is this same Cauchy distribution. ||

In this section, we have given examples, some rather general, of statistics that are ancillary for various models. In the next section we will consider the relationship between sufficient statistics and ancillary statistics.

### 6.1.4 Sufficient, Ancillary, and Complete Statistics

A minimal sufficient statistic is a statistic that has achieved the maximal amount of data reduction possible while still retaining all the information about the parameter  $\theta$ . Intuitively, a minimal sufficient statistic eliminates all the extraneous information in the sample, retaining only that piece with information about  $\theta$ . Since the distribution of an ancillary statistic does not depend on  $\theta$ , it might be suspected that a minimal sufficient statistic is unrelated to (or mathematically speaking, independent of) an ancillary statistic. However, this is not necessarily the case. In this section, we investigate this relationship in some detail.

We have already discussed a situation in which an ancillary statistic is not independent of a minimal sufficient statistic. Recall Example 6.1.7 in which  $X_1, \dots, X_n$  were iid observations from a uniform( $\theta, \theta + 1$ ) distribution. At the end of Section 6.1.2, we noted that the statistic  $(X_{(n)} - X_{(1)}, (X_{(n)} + X_{(1)})/2)$  is a minimal sufficient statistic and in Example 6.1.8 we showed that  $X_{(n)} - X_{(1)}$  is an ancillary statistic. Thus, in this case, the ancillary statistic is an important component of the minimal sufficient statistic. Certainly, the ancillary statistic and the minimal sufficient statistic are not independent.

To emphasize the point that an ancillary statistic can sometimes give important information for inferences about  $\theta$ , we give another example.

**Example 6.1.11:** Let  $X_1$  and  $X_2$  be iid observations from the discrete distribution that satisfies

$$P_\theta(X = \theta) = P_\theta(X = \theta + 1) = P_\theta(X = \theta + 2) = \frac{1}{3}$$

where  $\theta$ , the unknown parameter, is any integer. Let  $X_{(1)} \leq X_{(2)}$  be the order statistics for the sample. It can be shown with an argument similar to that in Example 6.1.7 that  $(R, M)$ , where  $R = X_{(2)} - X_{(1)}$  and  $M = (X_{(1)} + X_{(2)})/2$ , is a minimal sufficient statistic. Since this is a location parameter family, by Example 6.1.8,  $R$  is an ancillary statistic. To see how  $R$  might give information about  $\theta$ , even though it is ancillary, consider a sample point  $(r, m)$ , where  $m$  is an integer. First considering only  $m$ , for this sample point to have positive probability,  $\theta$  must be one of three values. Either  $\theta = m$  or  $\theta = m - 1$  or  $\theta = m - 2$ . With only the information that  $M = m$ , all three  $\theta$  values are possible values. But now suppose we get the additional information that  $R = 2$ . Then it must be the case that  $X_{(1)} = m - 1$  and  $X_{(2)} = m + 1$ . With this additional information, the only possible value for  $\theta$  is  $\theta = m - 1$ . Thus, the knowledge of the value of the ancillary statistic  $R$  has increased our knowledge about  $\theta$ . Of course, the knowledge of  $R$  alone would give us no information about  $\theta$ . (The idea that an ancillary statistic gives information about the *precision* of an estimate of  $\theta$  is not new. See Cox (1971) or Efron and Hinkley (1978) for more ideas.) ||

For many important situations, however, our intuition that a minimal sufficient statistic is independent of any ancillary statistic is correct. A description of situations in which this occurs relies on the next definition.

**DEFINITION 6.1.4:** Let  $f(t|\theta)$  be a family of pdfs or pmfs for a statistic  $T(\mathbf{X})$ . The family of probability distributions is called *complete* if  $E_\theta g(T) = 0$  for all  $\theta$  implies  $P_\theta(g(T) = 0) = 1$  for all  $\theta$ . Equivalently,  $T(\mathbf{X})$  is called a *complete statistic*.

Notice that completeness is a property of a family of probability distributions, not of a particular distribution. For example, if  $X$  has a  $n(0, 1)$  distribution, then defining  $g(x) = x$ , we have that  $Eg(X) = EX = 0$ . But the function  $g(x) = x$  satisfies  $P(g(X) = 0) = P(X = 0) = 0$ , not 1. However, this is a particular distribution, not a family of distributions. If  $X$  has a  $n(\theta, 1)$  distribution,  $-\infty < \theta < \infty$ , we shall see that no function of  $X$ , except one that is 0 with probability 1 for all  $\theta$ , satisfies  $E_\theta g(X) = 0$  for all  $\theta$ . Thus, the family of  $n(\theta, 1)$  distributions,  $-\infty < \theta < \infty$ , is complete.

**Example 6.1.12:** Let  $T$  have a binomial( $n, p$ ) distribution,  $0 < p < 1$ . Let  $g$  be a function such that  $E_p g(T) = 0$ . Then

$$\begin{aligned} 0 = E_p g(T) &= \sum_{t=0}^n g(t) \binom{n}{t} p^t (1-p)^{n-t} \\ &= (1-p)^n \sum_{t=0}^n g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t \end{aligned}$$

for all  $p$ ,  $0 < p < 1$ . The factor  $(1 - p)^n$  is not 0 for any  $p$  in this range. Thus it must be that

$$\begin{aligned} 0 &= \sum_{t=0}^n g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t \\ &= \sum_{t=0}^n g(t) \binom{n}{t} r^t \end{aligned}$$

for all  $r$ ,  $0 < r < \infty$ . But the last expression is a polynomial of degree  $n$  in  $r$  where the coefficient of  $r^t$  is  $g(t)\binom{n}{t}$ . For the polynomial to be 0 for all  $r$ , each coefficient must be 0. Since none of the  $\binom{n}{t}$  terms is zero, this implies that  $g(t) = 0$  for  $t = 0, 1, \dots, n$ . Since  $T$  takes on the values  $0, 1, \dots, n$  with probability 1, this yields that  $P_p(g(T) = 0) = 1$  for all  $p$ , the desired conclusion. Hence,  $T$  is a complete statistic. ||

**Example 6.1.13:** Let  $X_1, \dots, X_n$  be iid uniform( $0, \theta$ ) observations,  $0 < \theta < \infty$ . Using an argument similar to that in Example 6.1.3, it can be shown that  $T(\mathbf{X}) = \max_i X_i$  is a sufficient statistic and, by Theorem 5.5.2, the pdf of  $T(\mathbf{X})$  is

$$f(t|\theta) = \begin{cases} nt^{n-1}\theta^{-n} & 0 < t < \theta \\ 0 & \text{otherwise} \end{cases}$$

Suppose  $g(t)$  is a function satisfying  $E_\theta g(T) = 0$  for all  $\theta$ . Since  $E_\theta g(T)$  is constant as a function of  $\theta$ , its derivative with respect to  $\theta$  is 0. Thus we have that

$$\begin{aligned} 0 &= \frac{d}{d\theta} E_\theta g(T) = \frac{d}{d\theta} \int_0^\theta g(t) nt^{n-1}\theta^{-n} dt \\ &= (\theta^{-n}) \frac{d}{d\theta} \int_0^\theta ng(t)t^{n-1} dt \\ &\quad + \left( \frac{d}{d\theta} \theta^{-n} \right) \int_0^\theta ng(t)t^{n-1} dt \\ &\quad \left( \begin{array}{l} \text{applying the product} \\ \text{rule for differentiation} \end{array} \right) \\ &= \theta^{-n} ng(\theta)\theta^{n-1} + 0 \\ &= \theta^{-1}ng(\theta). \end{aligned}$$

The first term in the next-to-last line is the result of an application of the Fundamental Theorem of Calculus. The second term is 0 because the integral is, except for a constant, equal to  $E_\theta g(T)$ , which is 0. Since  $\theta^{-1}ng(\theta) = 0$  and  $\theta^{-1}n \neq 0$ , it must be that  $g(\theta) = 0$ . This is true for every  $\theta > 0$ ; hence,  $T$  is a complete statistic.

(On a somewhat pedantic note, realize that the Fundamental Theorem of Calculus does not apply to all functions, but only to functions that are *Riemann-integrable*. The equation

$$\frac{d}{d\theta} \int_0^\theta g(t)dt = g(\theta)$$

is valid only at points of continuity of Riemann-integrable  $g$ . Thus, strictly speaking, the above argument does not show that  $T$  is a complete statistic, since the condition of completeness applies to all functions, not just Riemann-integrable ones. From a more practical view, however, this distinction is not of concern since the condition of Riemann-integrability is so general that it includes virtually any function we could think of.) ||

We now use completeness to state a condition under which a minimal sufficient statistic is independent of every ancillary statistic.

**THEOREM 6.1.5 (Basu's Theorem):** If  $T(\mathbf{X})$  is a complete and minimal sufficient statistic, then  $T(\mathbf{X})$  is independent of every ancillary statistic.

*Proof:* We give the proof only for discrete distributions.

Let  $S(\mathbf{X})$  be any ancillary statistic. Then  $P(S(\mathbf{X}) = s)$  does not depend on  $\theta$  since  $S(\mathbf{X})$  is ancillary. Also the conditional probability,

$$P(S(\mathbf{X}) = s | T(\mathbf{X}) = t) = P(\mathbf{X} \in \{\mathbf{x}: S(\mathbf{x}) = s\} | T(\mathbf{X}) = t),$$

does not depend on  $\theta$  because  $T(\mathbf{X})$  is a sufficient statistic (recall the definition!). Thus, to show that  $S(\mathbf{X})$  and  $T(\mathbf{X})$  are independent, it suffices to show that

$$(6.1.5) \quad P(S(\mathbf{X}) = s | T(\mathbf{X}) = t) = P(S(\mathbf{X}) = s)$$

for all possible values  $t \in \mathcal{T}$ . Now,

$$P(S(\mathbf{X}) = s) = \sum_{t \in \mathcal{T}} P(S(\mathbf{X}) = s | T(\mathbf{X}) = t) P_\theta(T(\mathbf{X}) = t).$$

Furthermore, since  $\sum_{t \in \mathcal{T}} P_\theta(T(\mathbf{X}) = t) = 1$ , we can write

$$P(S(\mathbf{X}) = s) = \sum_{t \in \mathcal{T}} P(S(\mathbf{X}) = s) P_\theta(T(\mathbf{X}) = t).$$

Therefore, if we define the statistic

$$g(t) = P(S(\mathbf{X}) = s | T(\mathbf{X}) = t) - P(S(\mathbf{X}) = s),$$

the above two equations show that

$$E_\theta g(T) = \sum_{t \in T} g(t) P_\theta(T(\mathbf{X}) = t) = 0, \quad \text{for all } \theta.$$

Since  $T(\mathbf{X})$  is a complete statistic, this implies that  $g(t) = 0$  for all possible values  $t \in T$ . Hence (6.1.5) is verified.  $\square$

It should be noted that the “minimality” of the sufficient statistic was not used in the proof of Basu’s Theorem. Indeed, the theorem is true with the word “minimal” omitted. But, for the problems we will consider, a sufficient statistic will be complete only if it is minimal. Thus, we have stated Basu’s Theorem in this way as a reminder that the statistic  $T(\mathbf{X})$  in the theorem is a minimal sufficient statistic. More about the relationship between complete statistics and minimal sufficient statistics can be found in Lehmann and Scheffé (1950).

Basu’s Theorem is useful in that it allows us to deduce the independence of two statistics without ever finding the joint distribution of the two statistics. To use Basu’s Theorem, we need to show that a statistic is complete, which is sometimes a rather difficult analysis problem. Fortunately, most problems we are concerned with are covered by the following theorem. We will not prove this theorem but note that its proof depends on the uniqueness of a Laplace transform, a property that was mentioned in Section 2.3.

**THEOREM 6.1.6:** Let  $X_1, \dots, X_n$  be iid observations from an exponential family with pdf or pmf of the form

$$f(x|\theta) = h(x)c(\theta)\exp(w(\theta)t(x)).$$

Then the statistic

$$T(\mathbf{X}) = \sum_{i=1}^n t(X_i)$$

is complete.  $\square$

As an example of the use of Basu’s Theorem, Theorem 6.1.6, and many of the earlier results in this chapter, consider this example.

**Example 6.1.14:** Let  $X_1, \dots, X_n$  be iid exponential observations with parameter  $\theta$ . Consider computing the expected value of

$$g(\mathbf{X}) = \frac{X_n}{X_1 + \dots + X_n}.$$

We first note that the exponential distributions form a scale parameter family and thus, by Example 6.1.10,  $g(\mathbf{X})$  is an ancillary statistic. The exponential distributions also form an exponential family with  $t(x) = x$  and so, by Theorem 6.1.6,

$$T(\mathbf{X}) = \sum_{i=1}^n X_i$$

is a complete statistic and, by Theorem 6.1.3,  $T(\mathbf{X})$  is a sufficient statistic. (As noted after Basu's Theorem, we need not verify that  $T(\mathbf{X})$  is minimal although it could easily be verified using Theorem 6.1.4.) Hence, by Basu's Theorem,  $T(\mathbf{X})$  and  $g(\mathbf{X})$  are independent. Thus we have

$$\theta = E_\theta X_n = E_\theta T(\mathbf{X})g(\mathbf{X}) = (E_\theta T(\mathbf{X}))(E_\theta g(\mathbf{X})) = n\theta E_\theta g(\mathbf{X}).$$

Hence, for any  $\theta$ ,  $E_\theta g(\mathbf{X}) = n^{-1}$ . ||

**Example 6.1.15:** As another example of the use of Basu's Theorem, we consider the independence of  $\bar{X}$  and  $S^2$ , the sample mean and variance, when sampling from a  $n(\mu, \sigma^2)$  population. We have, of course, already shown that these statistics are independent in Theorem 5.4.1, but we will illustrate the use of Basu's Theorem in this important context. First consider  $\sigma^2$  fixed and let  $\mu$  vary,  $-\infty < \mu < \infty$ . By Example 6.1.2,  $\bar{X}$  is a sufficient statistic for  $\mu$ . Theorem 6.1.6 may be used to deduce that the family of  $n(\mu, \sigma^2/n)$  distributions,  $-\infty < \mu < \infty, \sigma^2/n$  known, is a complete family. Since this is the distribution of  $\bar{X}$ ,  $\bar{X}$  is a complete statistic. Now consider  $S^2$ . An argument similar to those used in Examples 6.1.9 and 6.1.10 could be used to show that in any location parameter family (remember  $\sigma^2$  is fixed,  $\mu$  is the location parameter),  $S^2$  is an ancillary statistic. Or, for this normal model, we can use Theorem 5.4.1 to see that the distribution of  $S^2$  depends on the fixed quantity  $\sigma^2$  but not on the parameter  $\mu$ . Either way,  $S^2$  is ancillary and so, by Basu's Theorem,  $S^2$  is independent of the complete sufficient statistic  $\bar{X}$ . For any  $\mu$  and the fixed  $\sigma^2$ ,  $\bar{X}$  and  $S^2$  are independent. But since  $\sigma^2$  was arbitrary, we have that the sample mean and variance are independent for any choice of  $\mu$  and  $\sigma^2$ . Note that neither  $\bar{X}$  nor  $S^2$  is ancillary in this model when both  $\mu$  and  $\sigma^2$  are unknown. Yet, by this argument, we are still able to use Basu's Theorem to deduce independence. This kind of argument is sometimes useful, but the fact remains that it is often harder to show that a statistic is complete than it is to show that two statistics are independent. ||

Basu's Theorem gives one relationship between sufficient statistics and ancillary statistics using the concept of complete statistics. There are other possible definitions of ancillarity and completeness. Some relationships between sufficiency and ancillarity for these definitions are discussed by Lehmann (1981).

## 6.2 The Likelihood Principle

In this section we study a specific, important statistic called the likelihood function, that also can be used to summarize data. There are many ways to use the likelihood

function, some of which are mentioned in this section and some in later chapters. But the main consideration in this section is an argument which indicates that, if certain other principles are accepted, the likelihood function *must* be used as a data reduction device.

### 6.2.1 The Likelihood Function

**DEFINITION 6.2.1:** Let  $f(\mathbf{x}|\theta)$  denote the joint pdf or pmf of the sample  $\mathbf{X} = (X_1, \dots, X_n)$ . Then, given that  $\mathbf{X} = \mathbf{x}$  is observed, the function of  $\theta$  defined by

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

is called the *likelihood function*.

If  $\mathbf{X}$  is a discrete random vector then  $L(\theta|\mathbf{x}) = P_\theta(\mathbf{X} = \mathbf{x})$ . If we compare the likelihood function at two parameter points and find that

$$P_{\theta_1}(\mathbf{X} = \mathbf{x}) = L(\theta_1|\mathbf{x}) > L(\theta_2|\mathbf{x}) = P_{\theta_2}(\mathbf{X} = \mathbf{x}),$$

then the sample we actually observed is more likely to have occurred if  $\theta = \theta_1$  than if  $\theta = \theta_2$ , which can be interpreted as saying that  $\theta_1$  is a more plausible value for the true value of  $\theta$  than is  $\theta_2$ . Many different ways have been proposed to use this information, but certainly it seems reasonable to examine the probability of the sample we actually observed under various possible values of  $\theta$ . This is the information provided by the likelihood function.

If  $X$  is a continuous, real-valued random variable and if the pdf of  $X$  is continuous in  $x$  then, for small  $\epsilon$ ,  $P_\theta(x - \epsilon < X < x + \epsilon)$  is approximately  $2\epsilon f(x|\theta) = 2\epsilon L(\theta|\mathbf{x})$  (this follows from the definition of a derivative). Thus,

$$\frac{P_{\theta_1}(x - \epsilon < X < x + \epsilon)}{P_{\theta_2}(x - \epsilon < X < x + \epsilon)} \approx \frac{L(\theta_1|\mathbf{x})}{L(\theta_2|\mathbf{x})}$$

and comparison of the likelihood function at two parameter values again gives an approximate comparison of the probability of the observed sample value,  $\mathbf{x}$ .

Definition 6.2.1 almost seems to be defining the likelihood function to be the same as the pdf or pmf. The only distinction between these two functions is which variable is considered fixed and which is varying. When we consider the pdf or pmf  $f(\mathbf{x}|\theta)$ , we are considering  $\theta$  as fixed and  $\mathbf{x}$  as the variable; when we consider the likelihood function  $L(\theta|\mathbf{x})$ , we are considering  $\mathbf{x}$  to be the observed sample point and  $\theta$  to be varying over all possible parameter values.

**Example 6.2.1:** Let  $X$  have a negative binomial distribution with  $r = 3$  and success probability  $p$ . If  $x = 2$  is observed then the likelihood function is the fifth-degree polynomial on  $0 \leq p \leq 1$  defined by

$$L(p|2) = P_p(X=2) = \binom{4}{2} p^3(1-p)^2.$$

In general, if  $X = x$  is observed, then the likelihood function is the polynomial of degree  $3 + x$ ,

$$L(p|x) = \binom{3+x-1}{x} p^3(1-p)^x. \quad ||$$

The Likelihood Principle specifies how the likelihood function should be used as a data reduction device.

**Likelihood Principle:** If  $x$  and  $y$  are two sample points such that  $L(\theta|x)$  is proportional to  $L(\theta|y)$ , that is, there exists a constant  $C(x, y)$  such that

$$(6.2.1) \quad L(\theta|x) = C(x, y)L(\theta|y) \quad \text{for all } \theta,$$

then the conclusions drawn from  $x$  and  $y$  should be identical.

Note that the constant  $C(x, y)$  in (6.2.1) may be different for different  $(x, y)$  pairs but  $C(x, y)$  does not depend on  $\theta$ .

In the special case of  $C(x, y) = 1$ , the Likelihood Principle states that if two sample points result in the same likelihood function then they contain the same information about  $\theta$ . But the Likelihood Principle goes further. It states that even if two sample points only have proportional likelihoods then they contain equivalent information about  $\theta$ . The rationale is this: The likelihood function is used to compare the plausibility of various parameter values, and, if  $L(\theta_2|x) = 2L(\theta_1|x)$  then, in some sense,  $\theta_2$  is twice as plausible as  $\theta_1$ . If (6.2.1) is also true then  $L(\theta_2|y) = 2L(\theta_1|y)$ . Thus, whether we observe  $x$  or  $y$  we conclude that  $\theta_2$  is twice as plausible as  $\theta_1$ .

We carefully used the word “plausible” rather than “probable” in the preceding paragraph because we often think of  $\theta$  as a fixed (albeit unknown) value. Furthermore, although  $f(x|\theta)$ , as a function of  $x$ , is a pdf, there is no guarantee that  $L(\theta|x)$ , as a function of  $\theta$ , is a pdf.

One form of inference, called *fiducial inference*, explicitly interprets likelihoods as probabilities for  $\theta$ . That is,  $L(\theta|x)$  is multiplied by  $M(x) = (\int_{-\infty}^{\infty} L(\theta|x)d\theta)^{-1}$  (the integral is replaced by a sum if the parameter space is countable) and then  $M(x)L(\theta|x)$  is interpreted as a pdf for  $\theta$  (provided, of course, that  $M(x)$  is finite!). Clearly,  $L(\theta|x)$  and  $L(\theta|y)$  satisfying (6.2.1) will yield the same pdf since the constant  $C(x, y)$  will simply be absorbed into the normalizing constant. Most statisticians do not subscribe to the fiducial theory of inference but it has a long history, dating back at least to the work of Fisher in the 1930s. We will, for history’s sake, compute one fiducial distribution.

**Example 6.2.2:** Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$ ,  $\sigma^2$  known. Using expression (6.1.3) for  $L(\mu|x)$ , we note first that (6.2.1) is satisfied if and only if  $\bar{x} = \bar{y}$ , in which case

$$C(\mathbf{x}, \mathbf{y}) = \exp \left( -\sum_{i=1}^n (x_i - \bar{x})^2 / (2\sigma^2) + \sum_{i=1}^n (y_i - \bar{y})^2 / (2\sigma^2) \right).$$

Thus, the Likelihood Principle states that the same conclusion about  $\mu$  should be drawn for any two sample points satisfying  $\bar{x} = \bar{y}$ . To compute the fiducial pdf for  $\mu$ , we see that if we define  $M(\mathbf{x}) = n^{n/2} \exp(\sum_{i=1}^n (x_i - \bar{x})^2 / (2\sigma^2))$  then  $M(\mathbf{x})L(\mu|\mathbf{x})$  (as a function of  $\mu$ ) is a  $n(\bar{x}, \sigma^2/n)$  pdf. This is the *fiducial distribution* of  $\mu$  and a fiducialist can make the following probability calculation regarding  $\mu$ .

The parameter  $\mu$  has a  $n(\bar{x}, \sigma^2/n)$  distribution. Hence,  $(\mu - \bar{x})/(\sigma/\sqrt{n})$  has a  $n(0, 1)$  distribution. Thus we have

$$\begin{aligned}.95 &= P \left( -1.96 < \frac{\mu - \bar{x}}{\sigma/\sqrt{n}} < 1.96 \right) \\&= P(-1.96\sigma/\sqrt{n} < \mu - \bar{x} < 1.96\sigma/\sqrt{n}) \\&= P(\bar{x} - 1.96\sigma/\sqrt{n} < \mu < \bar{x} + 1.96\sigma/\sqrt{n}).\end{aligned}$$

This algebra is similar to earlier calculations but the interpretation is quite different. Here  $\bar{x}$  is a fixed, known number, the observed data value, and  $\mu$  is the variable with the normal probability distribution. ||

We will discuss other, more common, uses of the likelihood function in later chapters when we discuss specific methods of inference. But now we consider an argument that shows that the Likelihood Principle is a necessary consequence of two other fundamental principles.

### 6.2.2 The Formal Likelihood Principle

For discrete distributions, the Likelihood Principle can be derived from two intuitively simpler ideas. This is also true, with some qualifications, for continuous distributions. In this subsection we will deal only with discrete distributions. Berger and Wolpert (1984) provide a thorough discussion of the Likelihood Principle in both the discrete and continuous cases. These results were first proved by Birnbaum (1962) in a landmark paper, but our presentation more closely follows that of Berger and Wolpert.

Formally, we define an experiment  $E$  to be a triple  $(\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$ , where  $\mathbf{X}$  is a random vector with pmf  $f(\mathbf{x}|\theta)$  for some  $\theta$  in the parameter space  $\Theta$ . An experimenter, knowing what experiment  $E$  was performed, and having observed a particular sample  $\mathbf{X} = \mathbf{x}$ , will make some inference or draw some conclusion about  $\theta$ . This conclusion we denote by  $\text{Ev}(E, \mathbf{x})$ , which stands for the *evidence about  $\theta$  arising from  $E$  and  $\mathbf{x}$* .

**Example 6.2.3:** Let  $E$  be the experiment consisting of observing  $X_1, \dots, X_n$  iid  $n(\mu, \sigma^2)$ ,  $\sigma^2$  known. Since the sample mean,  $\bar{X}$ , is a sufficient statistic for  $\mu$  and  $E\bar{X} = \mu$ , we might use the observed value  $\bar{X} = \bar{x}$  as an estimate of  $\mu$ . To give a measure of the accuracy of this estimate, it is common to report the standard deviation of  $\bar{X}$ ,  $\sigma/\sqrt{n}$ . Thus we could define  $\text{Ev}(E, \mathbf{x}) = (\bar{x}, \sigma/\sqrt{n})$ . Here we see that the  $\bar{x}$

coordinate depends on the observed sample  $\mathbf{x}$  while the  $\sigma/\sqrt{n}$  coordinate depends on the knowledge of  $E$ . ||

To relate the concept of an evidence function to something familiar we now restate the Sufficiency Principle of Section 6.1 in terms of these concepts.

**FORMAL SUFFICIENCY PRINCIPLE:** Consider experiment  $E = (\mathbf{X}, \theta, \{f(\mathbf{x}|\theta)\})$  and suppose  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$ . If  $\mathbf{x}$  and  $\mathbf{y}$  are sample points satisfying  $T(\mathbf{x}) = T(\mathbf{y})$ , then  $\text{Ev}(E, \mathbf{x}) = \text{Ev}(E, \mathbf{y})$ .

Thus, the *Formal* Sufficiency Principle goes slightly further than the Sufficiency Principle of Section 6.1. In Section 6.1 no mention was made of the experiment. Here, we are agreeing to equate evidence if the sufficient statistics match. The Likelihood Principle can be derived from the Formal Sufficiency Principle and the following principle, an eminently reasonable one.

**CONDITIONALITY PRINCIPLE:** Let  $E_1 = (\mathbf{X}_1, \theta, \{f_1(\mathbf{x}_1|\theta)\})$  and  $E_2 = (\mathbf{X}_2, \theta, \{f_2(\mathbf{x}_2|\theta)\})$  be two experiments, where only the unknown parameter  $\theta$  need be common between the two experiments. Consider the mixed experiment in which the random variable  $J$  is observed, where  $P(J = 1) = P(J = 2) = \frac{1}{2}$  (independent of  $\theta, \mathbf{X}_1$ , or  $\mathbf{X}_2$ ), and then experiment  $E_J$  is performed. Formally, the experiment performed is  $E^* = (\mathbf{X}^*, \theta, \{f^*(\mathbf{x}^*|\theta)\})$ , where  $\mathbf{X}^* = (j, \mathbf{X}_j)$  and  $f^*(\mathbf{x}^*|\theta) = f^*((j, \mathbf{x}_j)|\theta) = \frac{1}{2}f_j(\mathbf{x}_j|\theta)$ . Then

$$(6.2.2) \quad \text{Ev}(E^*, (j, \mathbf{x}_j)) = \text{Ev}(E_j, \mathbf{x}_j).$$

The Conditionality Principle simply says that if one of two experiments is randomly chosen and the chosen experiment is done, yielding data  $\mathbf{x}$ , the information about  $\theta$  *depends only on the experiment performed*. That is, it is the same information as would have been obtained if it were decided (nonrandomly) to do that experiment from the beginning, and data  $\mathbf{x}$  had been observed. The fact that this experiment was performed, rather than some other, has not increased, decreased, or changed knowledge of  $\theta$ .

**Example 6.2.4:** Suppose the parameter of interest is the probability  $p$ ,  $0 < p < 1$ , where  $p$  denotes the probability that a particular coin will land “Heads” when it is flipped. Let  $E_1$  be the experiment consisting of tossing the coin 20 times and recording the number of Heads in those 20 tosses.  $E_1$  is a binomial experiment and  $\{f_1(x_1|p)\}$  is the family of  $\text{binomial}(20, p)$  pmfs. Let  $E_2$  be the experiment consisting of tossing the coin until the seventh Head occurs and recording the number of Tails before the seventh Head.  $E_2$  is a negative binomial experiment. Now suppose the experimenter uses a random number table to choose between these two experiments, happens to choose  $E_2$ , and collects data consisting of the seventh Head occurring on trial 20. The Conditionality Principle says that the information about  $\theta$  that the experimenter now has,  $\text{Ev}(E^*, (2, 13))$ , is the same as that which he would have,  $\text{Ev}(E_2, 13)$ , if he had just chosen to do the negative binomial experiment and had never contemplated the binomial experiment. ||

The following Formal Likelihood Principle can now be derived from the Formal Sufficiency Principle and the Conditionality Principle.

**FORMAL LIKELIHOOD PRINCIPLE:** Consider two experiments,  $E_1 = (X_1, \theta, \{f_1(x_1|\theta)\})$  and  $E_2 = (X_2, \theta, \{f_2(x_2|\theta)\})$ , where the unknown parameter  $\theta$  is the same in both experiments. Suppose  $x_1^*$  and  $x_2^*$  are sample points from  $E_1$  and  $E_2$ , respectively, such that

$$(6.2.3) \quad L(\theta|x_2^*) = CL(\theta|x_1^*)$$

for all  $\theta$  and for some constant  $C$  that may depend on  $x_1^*$  and  $x_2^*$  but not  $\theta$ . Then

$$\text{Ev}(E_1, x_1^*) = \text{Ev}(E_2, x_2^*).$$

The Formal Likelihood Principle is different from the Likelihood Principle in Section 6.2.1 because the Formal Likelihood Principle concerns two experiments, whereas the Likelihood Principle concerns only one. The Likelihood Principle, however, can be derived from the Formal Likelihood Principle by letting  $E_2$  be an exact replicate of  $E_1$ . Thus, the two-experiment setting in the Formal Likelihood Principle is something of an artifact and the important consequence is the following corollary, whose proof is left as an exercise. (See Exercise 6.25.)

**Likelihood Principle Corollary:** If  $E = (X, \theta, \{f(x|\theta)\})$  is an experiment, then  $\text{Ev}(E, x)$  should depend on  $E$  and  $x$  only through  $L(\theta|x)$ .

Now we prove Birnbaum's Theorem and then investigate its somewhat surprising consequences.

**Theorem 6.2.1 (Birnbaum's Theorem):** The Formal Likelihood Principle follows from the Formal Sufficiency Principle and the Conditionality Principle. The converse is also true.

*Proof:* Let  $E_1, E_2, x_1^*$ , and  $x_2^*$  be as defined in the Formal Likelihood Principle and let  $E^*$  be the mixed experiment from the Conditionality Principle. On the sample space of  $E^*$  define the statistic

$$T(j, x_j) = \begin{cases} (1, x_1^*) & \text{if } j = 1 \text{ and } x_1 = x_1^* \text{ or if } j = 2 \text{ and } x_2 = x_2^* \\ (j, x_j) & \text{otherwise} \end{cases}.$$

We will use the Factorization Theorem to prove that  $T(J, X_J)$  is a sufficient statistic in the  $E^*$  experiment. Define

$$h(j, x_j) = \begin{cases} C & \text{if } (j, x_j) = (2, x_2^*) \\ 1 & \text{otherwise} \end{cases},$$

where  $C$  is the constant from (6.2.3), and define

$$g(t|\theta) = g((j, x_j)|\theta) = f^*((j, x_j)|\theta).$$

For all sample points except  $(2, \mathbf{x}_2^*)$  (but including  $(1, \mathbf{x}_1^*)$ ),  $T(j, \mathbf{x}_j) = (j, \mathbf{x}_j)$  so

$$g(T(j, \mathbf{x}_j)|\theta)h(j, \mathbf{x}_j) = g((j, \mathbf{x}_j)|\theta)(1) = f^*((j, \mathbf{x}_j)|\theta).$$

For  $(2, \mathbf{x}_2^*)$  we also have

$$\begin{aligned} g(T(2, \mathbf{x}_2^*)|\theta)h(2, \mathbf{x}_2^*) &= g((1, \mathbf{x}_1^*)|\theta)C && \text{(definition of } T\text{)} \\ &= f^*((1, \mathbf{x}_1^*)|\theta)C && \text{(definition of } g\text{)} \\ &= C \frac{1}{2} f_1(\mathbf{x}_1^*|\theta) && \left( \begin{array}{l} \text{definition of } f^* \text{ in} \\ \text{Conditionality Principle} \end{array} \right) \\ &= C \frac{1}{2} L(\theta|\mathbf{x}_1^*) && \text{(definition of } L\text{)} \\ &= \frac{1}{2} L(\theta|\mathbf{x}_2^*) && \text{(by (6.2.3))} \\ &= \frac{1}{2} f_2(\mathbf{x}_2^*|\theta) && \text{(definition of } L\text{)} \\ &= f^*((2, \mathbf{x}_2^*)|\theta). && \text{(definition of } f^*\text{)} \end{aligned}$$

Thus, by the Factorization Theorem,  $T(J, \mathbf{X}_J)$  is a sufficient statistic for  $\theta$ . Now, by the Formal Sufficiency Principle,

$$(6.2.4) \quad \text{Ev}(E^*, (1, \mathbf{x}_1^*)) = \text{Ev}(E^*, (2, \mathbf{x}_2^*)).$$

Using (6.2.2), the Conditionality Principle, we have

$$\text{Ev}(E^*, (1, \mathbf{x}_1^*)) = \text{Ev}(E_1, \mathbf{x}_1^*),$$

and

$$\text{Ev}(E^*, (2, \mathbf{x}_2^*)) = \text{Ev}(E_2, \mathbf{x}_2^*).$$

Hence, using (6.2.4), this yields

$$\text{Ev}(E_1, \mathbf{x}_1^*) = \text{Ev}(E_2, \mathbf{x}_2^*),$$

the Formal Likelihood Principle.

To prove the converse, first consider the Conditionality Principle. Let one experiment be the  $E^*$  experiment and the other  $E_j$ . Then

$$\begin{aligned} L(\theta|(j, \mathbf{x}_j)) &= f^*((j, \mathbf{x}_j)|\theta) && \text{(definition of } L\text{)} \\ &= \frac{1}{2} f_j(\mathbf{x}_j|\theta) && \text{(definition of } f^*\text{)} \\ &= \frac{1}{2} L(\theta|\mathbf{x}_j). && \text{(definition of } L\text{)} \end{aligned}$$

Letting  $(j, \mathbf{x}_j)$  and  $\mathbf{x}_j$  play the roles of  $\mathbf{x}_1^*$  and  $\mathbf{x}_2^*$  in the Formal Sufficiency Principle we can conclude

$$\text{Ev}(E^*, (j, \mathbf{x}_j)) = \text{Ev}(E_j, \mathbf{x}_j),$$

the Conditionality Principle. Now consider the Formal Sufficiency Principle. If  $T(\mathbf{X})$  is sufficient and  $T(\mathbf{x}) = T(\mathbf{y})$ , then

$$L(\theta|\mathbf{x}) = CL(\theta|\mathbf{y}),$$

where  $C = h(\mathbf{x})/h(\mathbf{y})$  and  $h$  is the function from the Factorization Theorem. Hence, by the Formal Likelihood Principle,

$$\text{Ev}(E, \mathbf{x}) = \text{Ev}(E, \mathbf{y}),$$

the Formal Sufficiency Principle. □

**Example 6.2.4 (Continued):** Consider again the binomial and negative binomial experiments with the two sample points  $x_1 = 7$  (7 out of 20 Heads in the binomial experiment), and  $x_2 = 13$  (the 7th Head occurs on the 20th flip of the coin). The likelihood functions are

$$L(p|x_1 = 7) = \binom{20}{7} p^7(1-p)^{13} \quad \text{for the binomial experiment,}$$

and

$$L(p|x_2 = 13) = \binom{19}{6} p^7(1-p)^{13} \quad \text{for the negative binomial experiment.}$$

These are proportional likelihood functions so the Formal Likelihood Principle states that the same conclusion regarding  $p$  should be made in both cases. In particular, the Formal Likelihood Principle asserts that the fact that in the first case sampling ended because 20 trials were completed and in the second case sampling stopped because the 7th Head was observed is immaterial as far as our conclusions about  $p$  are concerned. Lindley and Phillips (1976) give a thorough discussion of the binomial-negative binomial inference problem. ||

This point, of equivalent inferences from different experiments, may be amplified by considering the sufficient statistic,  $T$ , defined in the proof of Birnbaum's Theorem and the sample points  $\mathbf{x}_1^* = 7$  and  $\mathbf{x}_2^* = 13$ . For any sample points in the mixed experiment, other than  $(1, 7)$  or  $(2, 13)$ ,  $T$  tells which experiment, binomial or negative binomial, was performed, and the result of the experiment. But for  $(1, 7)$  and  $(2, 13)$  we have  $T(1, 7) = T(2, 13) = (1, 7)$ . If we use only the sufficient statistic to make an inference and if  $T = (1, 7)$ , then all we know is that 7 out of 20 heads

were observed. We do not know whether the 7 or the 20 was the fixed quantity. But the Formal Sufficiency Principle applied to the mixture experiment states that the same conclusion regarding  $p$  should be obtained in either case. Then the Conditionality Principle asserts that the same conclusion should be reached whether just the binomial experiment is done (no prior random choice of experiments) and  $X_1 = 7$  is observed or just the negative binomial experiment is done and  $X_2 = 13$  is observed.

Many common statistical procedures violate the Formal Likelihood Principle. With these procedures, different conclusions would be reached for the two experiments discussed in Example 6.2.4. This violation of the Formal Likelihood Principle may seem strange because, by Birnbaum's Theorem, we are then violating either the Sufficiency Principle or the Conditionality Principle. Let us examine these two principles more closely.

The Formal Sufficiency Principle is, in essence, the same as that discussed in Section 6.1. There, we saw that all the information about  $\theta$  is contained in the sufficient statistic, and knowledge of the entire sample cannot add any information. Thus, basing evidence on the sufficient statistic is an eminently plausible principle. One shortcoming of this principle, one that invites violation, is that it is very model-dependent. As mentioned in the discussion after Example 6.1.4, belief in this principle necessitates belief in the model, something that may not be easy to do.

Most data analysts perform some sort of "model checking" when analyzing a set of data. Most model checking is, necessarily, based on statistics other than a sufficient statistic. For example, it is common practice to examine *residuals* from a model, statistics that measure variation in the data not accounted for by the model. (We will see residuals in more detail in Chapters 11 and 12.) Such a practice immediately violates the Sufficiency Principle, since the residuals are not based on sufficient statistics. (Of course, such a practice directly violates the Likelihood Principle also.) Thus, it must be realized that *before* considering the Sufficiency Principle (or the Likelihood Principle), we must be comfortable with the model.

The Conditionality Principle, stated informally, says that "only the experiment actually performed matters." That is, in Example 6.2.4, if we did the binomial experiment, and not the negative binomial experiment, then the (not done) negative binomial experiment should in no way influence our conclusion about  $\theta$ . This principle, also, seems to be eminently plausible.

How, then, can statistical practice violate the Formal Likelihood Principle, when it would mean violating either the Principle of Sufficiency or Conditionality? Several authors have addressed this question, among them Durbin (1970) and Kalbfleisch (1975). One argument, put forth by Kalbfleisch, is that the proof of the Formal Likelihood Principle is not compelling. This is because the Sufficiency Principle is applied in ignorance of the Conditionality Principle. The sufficient statistic,  $T(J, X_J)$ , used in the proof of Theorem 6.2.1, is defined on the mixture experiment. If the Conditionality Principle were invoked first, then separate sufficient statistics would have to be defined for each experiment. In this case, the Formal Likelihood Principle would no longer follow. (A key argument in the proof of Birnbaum's Theorem is that  $T(J, X_J)$  can take on the same value for sample points from each experiment. This cannot happen with separate sufficient statistics.)

At any rate, since many intuitively appealing inference procedures do violate the Likelihood Principle, it is not universally accepted by all statisticians. Yet it is mathematically appealing and does suggest a useful data reduction technique.

### 6.3 The Invariance Principle

The previous two sections both describe data reduction principles in the following way. A function  $T(\mathbf{x})$  of the sample is specified and the principle states that if  $\mathbf{x}$  and  $\mathbf{y}$  are two sample points with  $T(\mathbf{x}) = T(\mathbf{y})$  then the same inference about  $\theta$  should be made whether  $\mathbf{x}$  or  $\mathbf{y}$  is observed. The function  $T(\mathbf{x})$  is a sufficient statistic when the Sufficiency Principle is used. The “value” of  $T(\mathbf{x})$  is the set of all likelihood functions proportional to  $L(\theta|\mathbf{x})$  if the Likelihood Principle is used. The Invariance Principle describes a data reduction technique in a slightly different way. In any application of the Invariance Principle, a function  $T(\mathbf{x})$  is specified but, if  $T(\mathbf{x}) = T(\mathbf{y})$ , then the Invariance Principle states that the inference made if  $\mathbf{x}$  is observed should have a *certain relationship* to the inference made if  $\mathbf{y}$  is observed, although the two inferences may not be the same. This restriction on the inference procedure sometimes leads to a simpler analysis, just as do the data reduction principles discussed in earlier sections.

Although commonly combined into what is called the Invariance Principle, the data reduction technique we will now describe actually combines two different invariance considerations.

The first type of invariance might be called *measurement invariance*. It prescribes that the inference made should not depend on the measurement scale that is used. For example, suppose two foresters are going to estimate the average diameter of trees in a forest. The first uses data on tree diameters expressed in inches and the second uses the same data expressed in meters. Now both are asked to produce an estimate in inches. (The second might conveniently estimate the average diameter in meters and then transform the estimate to inches.) Measurement invariance requires that both foresters produce the same estimates. No doubt, almost all would agree that this type of invariance is reasonable.

The second type of invariance might be called *formal invariance*. It states that if two inference problems have the same formal structure in terms of the mathematical model used, then the same inference procedure should be used in both problems. The elements of the model that must be the same are:  $\Theta$ , the parameter space;  $\{f(\mathbf{x}|\theta) : \theta \in \Theta\}$ , the set of pdfs or pmfs for the sample; and the set of *allowable inferences and consequences of wrong inferences*. This last element has not been discussed much prior to this; for this section we will assume that the set of possible inferences is the same as  $\Theta$ , that is, an inference is simply a choice of an element of  $\Theta$  as an estimate or guess at the true value of  $\theta$ . Formal invariance is concerned only with the mathematical entities involved, not the physical description of the experiment. For example,  $\Theta$  may be  $\Theta = \{\theta : \theta > 0\}$  in two problems. But in one problem  $\theta$  may be the average price of a dozen eggs in the United States (measured in cents) and in another problem  $\theta$  may refer to the average height of giraffes in Kenya (measured in meters). Yet, formal invariance equates these two parameter spaces since they both refer to the same set of real numbers.

**INVARIANCE PRINCIPLE:** If  $Y = g(X)$  is a change of measurement scale such that the model for  $Y$  has the same formal structure as the model for  $X$ , then an inference procedure should be both measurement invariant and formally invariant.

We will now illustrate how these two concepts of invariance can work together to provide useful data reduction.

**Example 6.3.1:** Let  $X$  have a binomial distribution with sample size  $n$  known and success probability  $p$  unknown. Let  $T(x)$  be the estimate of  $p$  that is used when  $X = x$  is observed. Rather than using the number of successes,  $X$ , to make an inference about  $p$ , we could use the number of failures,  $Y = n - X$ .  $Y$  also has a binomial distribution with parameters  $(n, q = 1 - p)$ . Let  $T^*(y)$  be the estimate of  $q$  that is used when  $Y = y$  is observed so that  $1 - T^*(y)$  is the estimate of  $p$  when  $Y = y$  is observed. If  $x$  successes are observed then the estimate of  $p$  is  $T(x)$ . But if there are  $x$  successes then there are  $n - x$  failures and  $1 - T^*(n - x)$  is also an estimate of  $p$ . Measurement invariance requires that these two estimates be equal, that is,  $T(x) = 1 - T^*(n - x)$ , since the change from  $X$  to  $Y$  is just a change in measurement scale. Furthermore, the formal structures of the inference problems based on  $X$  and  $Y$  are the same.  $X$  and  $Y$  both have  $\text{binomial}(n, \theta)$  distributions,  $0 \leq \theta \leq 1$ . So formal invariance requires that  $T(z) = T^*(z)$  for all  $z = 0, \dots, n$ . Thus, measurement and formal invariance together require that

$$(6.3.1) \quad T(x) = 1 - T^*(n - x) = 1 - T(n - x).$$

If we consider only estimators satisfying (6.3.1) then we have greatly reduced and simplified the set of estimators we are willing to consider. Whereas the specification of an arbitrary estimator requires the specification of  $T(0), T(1), \dots, T(n)$ , the specification of an estimator satisfying (6.3.1) requires the specification only of  $T(0), T(1), \dots, T([n/2])$ , where  $[n/2]$  is the greatest integer not larger than  $n/2$ . The remaining values of  $T(x)$  are determined by those already specified and (6.3.1). For example,  $T(n) = 1 - T(0)$  and  $T(n - 1) = 1 - T(1)$ . This is the type of data reduction that is always achieved by the Invariance Principle. The inference to be made for some sample points determines the inference to be made for other sample points.

Two estimators that are invariant for this problem are  $T_1(x) = x/n$  and  $T_2(x) = .9(x/n) + .1(.5)$ . The estimator  $T_1(x)$  uses the sample proportion of successes to estimate  $p$ .  $T_2(x)$  “shrinks” the sample proportion toward .5, an estimator which might be sensible if there is reason to think that  $p$  is near .5. Condition (6.3.1) is easily verified for both of these estimators and so they are both invariant. An estimator that is not invariant is  $T_3(x) = .8(x/n) + .2(1)$ . Condition (6.3.1) is not satisfied since  $T_3(0) = .2 \neq 0 = 1 - T_3(n - 0)$ . ||

A key to the invariance argument in Example 6.3.1 and to any invariance argument is the choice of the transformations. The data transformation used in Example 6.3.1 is  $Y = n - X$ . The transformations (changes of measurement scale) used in any application of the Invariance Principle are described by a set of functions on the sample space called a *group of transformations*.

**DEFINITION 6.3.1:** A set of functions  $\{g(x) : g \in \mathcal{G}\}$  from the sample space  $\mathcal{X}$  onto  $\mathcal{X}$  is called a *group of transformations* of  $\mathcal{X}$  if

- i. (*Inverse*) For every  $g \in \mathcal{G}$  there is a  $g' \in \mathcal{G}$  such that  $g'(g(x)) = x$  for all  $x \in \mathcal{X}$ .
- ii. (*Composition*) For every  $g \in \mathcal{G}$  and  $g' \in \mathcal{G}$  there exists  $g'' \in \mathcal{G}$  such that  $g'(g(x)) = g''(x)$  for all  $x \in \mathcal{X}$ .
- Sometimes the third requirement,
- iii. (*Identity*) The identity,  $e(x)$ , defined by  $e(x) = x$  is an element of  $\mathcal{G}$ ,

is stated as part of the definition of a group. But (iii) is a consequence of (i) and (ii) and need not be verified separately. (See Exercise 6.28.)

**Example 6.3.1 (Continued):** For this problem, only two transformations are involved so we may set  $\mathcal{G} = \{g_1, g_2\}$ , with  $g_1(x) = n - x$  and  $g_2(x) = x$ . Conditions (i) and (ii) are easily verified. The choice of  $g' = g$  verifies (i), that is, each element is its own inverse. For example,

$$g_1(g_1(x)) = g_1(n - x) = n - (n - x) = x.$$

In (ii), if  $g' = g$  then  $g'' = g_2$  while if  $g' \neq g$  then  $g'' = g_1$  satisfies the equality. For example, take  $g' \neq g = g_1$ . Then

$$g_2(g_1(x)) = g_2(n - x) = n - x = g_1(x). \quad ||$$

To use the Invariance Principle, we must be able to apply formal invariance to the transformed problem. That is, after changing the measurement scale we must still have the same formal structure. This requirement is summarized in the next definition.

**DEFINITION 6.3.2:** Let  $\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$  be a set of pdfs or pmfs for  $X$  and let  $\mathcal{G}$  be a group of transformations of the sample space  $\mathcal{X}$ . Then  $\mathcal{F}$  is *invariant under the group  $\mathcal{G}$*  if for every  $\theta \in \Theta$  and  $g \in \mathcal{G}$  there exists a unique  $\theta' \in \Theta$  such that  $Y = g(X)$  has the distribution  $f(y|\theta')$  if  $X$  has the distribution  $f(x|\theta)$ .

**Example 6.3.1 (Continued):** In the binomial problem, we must check both  $g_1$  and  $g_2$ . If  $X \sim \text{binomial}(n, p)$  then  $g_1(X) = n - X \sim \text{binomial}(n, 1 - p)$  so  $p' = 1 - p$  where  $p$  plays the role of  $\theta$  in Definition 6.3.2. Also  $g_2(X) = X \sim \text{binomial}(n, p)$  so  $p' = p$  in this case. Thus the set of binomial pmfs is invariant under the group  $\mathcal{G} = \{g_1, g_2\}$ .  $||$

In Example 6.3.1, the group of transformations had only two elements. In many cases, the group of transformations is infinite, as the next examples illustrate.

**Example 6.3.2:** Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$ , both  $\mu$  and  $\sigma^2$  unknown. Consider the group of transformations defined by  $\mathcal{G} = \{g_a(x), -\infty < a < \infty\}$  where

$g_a(x_1, \dots, x_n) = (x_1 + a, \dots, x_n + a)$ . To verify that this set of transformations is a group, conditions (i) and (ii) from Definition 6.3.1 must be verified. For (i) note that

$$\begin{aligned} g_{-a}(g_a(x_1, \dots, x_n)) &= g_{-a}(x_1 + a, \dots, x_n + a) \\ &= (x_1 + a - a, \dots, x_n + a - a) \\ &= (x_1, \dots, x_n). \end{aligned}$$

So if  $g = g_a$  then  $g' = g_{-a}$  satisfies (i). For (ii) note that

$$\begin{aligned} g_{a_2}(g_{a_1}(x_1, \dots, x_n)) &= g_{a_2}(x_1 + a_1, \dots, x_n + a_1) \\ &= (x_1 + a_1 + a_2, \dots, x_n + a_1 + a_2) \\ &= g_{a_1+a_2}(x_1, \dots, x_n). \end{aligned}$$

So if  $g = g_{a_1}$  and  $g' = g_{a_2}$  then  $g'' = g_{a_1+a_2}$  satisfies (ii), and Definition 6.3.1 is verified.  $\mathcal{G}$  is a group of transformations.

The set  $\mathcal{F}$  in this problem is the set of all joint densities  $f(x_1, \dots, x_n | \mu, \sigma^2)$  for  $X_1, \dots, X_n$  defined by “ $X_1, \dots, X_n$  are iid  $n(\mu, \sigma^2)$  for some  $-\infty < \mu < \infty$  and  $\sigma^2 > 0$ .” For any  $a$ ,  $-\infty < a < \infty$ , the random variables  $Y_1, \dots, Y_n$  defined by

$$(Y_1, \dots, Y_n) = g_a(X_1, \dots, X_n) = (X_1 + a, \dots, X_n + a)$$

are iid  $n(\mu + a, \sigma^2)$  random variables. Thus, the joint distribution of  $\mathbf{Y} = g_a(\mathbf{X})$  is in  $\mathcal{F}$  and hence  $\mathcal{F}$  is invariant under  $\mathcal{G}$ . In terms of the notation in Definition 6.3.2, if  $\theta = (\mu, \sigma^2)$  then  $\theta' = (\mu + a, \sigma^2)$ . ||

In Example 6.3.2 we have shown only that a particular model is invariant under a particular group of transformations. We now investigate how the Invariance Principle affects different types of data reduction for different inference problems.

**Example 6.3.3:** Suppose that for the model in Example 6.3.2, the inference to be made is an estimate of the mean  $\mu$ . Let  $T(\mathbf{x})$  be the estimate used if  $\mathbf{X} = \mathbf{x}$  is observed. If  $g_a(\mathbf{X}) = \mathbf{Y} = \mathbf{y}$  is observed, then let  $T^*(\mathbf{y})$  be the estimate of  $\mu + a$ , the mean of each  $Y_i$ . If  $\mu + a$  is estimated by  $T^*(\mathbf{y})$ , then  $\mu$  would be estimated by  $T^*(\mathbf{y}) - a$ . Measurement invariance thus requires that

$$T(\mathbf{x}) = T^*(\mathbf{y}) - a,$$

that is,

$$(6.3.2) \quad T(x_1, \dots, x_n) = T^*(x_1 + a, \dots, x_n + a) - a,$$

for all  $\mathbf{x} = (x_1, \dots, x_n)$  and all  $a$ . But both problems, that based on  $\mathbf{X}$  and that based on  $\mathbf{Y}$ , have the same formal structure. They both involve estimating the mean of  $n$  iid normal observations with unknown mean and variance. Thus formal invariance requires that

$$(6.3.3) \quad T(\mathbf{x}) = T^*(\mathbf{x})$$

for all  $\mathbf{x}$ . Thus, combining (6.3.2) and (6.3.3), the Invariance Principle requires that

$$(6.3.4) \quad T(x_1, \dots, x_n) + a = T(x_1 + a, \dots, x_n + a)$$

for all  $(x_1, \dots, x_n)$  and all  $a$ . The data reduction that has been accomplished is that, once  $T(\mathbf{x})$  has been specified, then  $T(\mathbf{y})$  is specified for all  $\mathbf{y}$ s that can be derived from  $\mathbf{x}$  by the addition of a constant  $a$  to each coordinate. Two estimates of  $\mu$  which are invariant in the sense of (6.3.4) are the sample mean,  $T(x_1, \dots, x_n) = \bar{x} = (x_1 + \dots + x_n)/n$ , and the sample midrange,  $T(x_1, \dots, x_n) = (x_{(1)} + x_{(n)})/2$ , where  $x_{(1)} \leq \dots \leq x_{(n)}$  are the ordered values of  $x_1, \dots, x_n$ . An estimate which is not invariant in the sense of (6.3.4) is  $T(\mathbf{x}) = .9\bar{x}$ . To see that this estimate is not invariant, note that  $T(x_1, \dots, x_n) + a = .9\bar{x} + a \neq .9(\bar{x} + a) = T(x_1 + a, \dots, x_n + a)$ , unless  $a = 0$ . ||

**Example 6.3.4:** Consider again the model and group described in Example 6.3.2, but now suppose the inference to be made is an estimate of  $\sigma^2$ . Because the variance of  $Y_i = X_i + a$  is the same as the variance of  $X_i$ , measurement invariance requires that  $T(x_1, \dots, x_n) = T^*(x_1 + a, \dots, x_n + a)$ , where  $T(\mathbf{x})$  is the variance estimate based on  $\mathbf{X} = \mathbf{x}$  and  $T^*(\mathbf{y})$  is the variance estimate based on  $\mathbf{Y} = \mathbf{y}$ . Formal invariance again requires that  $T(\mathbf{x}) = T^*(\mathbf{x})$  for all  $\mathbf{x}$ , just as in Example 6.3.3. Thus for the inference problem, an estimate of  $\sigma^2$  is invariant if

$$(6.3.5) \quad T(x_1, \dots, x_n) = T(x_1 + a, \dots, x_n + a)$$

for all  $(x_1, \dots, x_n)$  and  $a$ . The two samples  $(x_1, \dots, x_n)$  and  $(x_1 + a, \dots, x_n + a)$  are located at different points on the real line but the dispersion or spread among the sample values is the same for both samples. Thus, it is reasonable to require an estimate of the variance to be the same for both samples. The sample variance  $s^2$  is an invariant estimate of  $\sigma^2$  in the sense of (6.3.5) because

$$\begin{aligned} s^2(x_1 + a, \dots, x_n + a) &= \frac{1}{n-1} \sum_{i=1}^n \left( (x_i + a) - \left( \frac{\sum_{j=1}^n (x_j + a)}{n} \right) \right)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n \left( x_i + a - \left( \frac{\sum_{j=1}^n x_j}{n} + a \right) \right)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n \left( x_i - \left( \frac{\sum_{j=1}^n x_j}{n} \right) \right)^2 \\ &= s^2(x_1, \dots, x_n). \end{aligned} \quad ||$$

Application of the Invariance Principle leads to different effects in (6.3.4) and (6.3.5). The effect described in (6.3.4), in which the estimate changes in a prescribed

way as the data are transformed, is sometimes called *equivariance*. Equation (6.3.1) also describes an equivariant estimate, this time for the binomial problem. The term *invariant* is sometimes reserved for situations like that described in (6.3.5), in which the estimate remains unchanged as the data are transformed. Specifically, an estimate satisfying (6.3.5) is called a *location-invariant estimate*.

Sometimes, the Invariance Principle is combined with the Sufficiency Principle to achieve greater data reduction. In most cases, the Sufficiency Principle is applied first and then the Invariance Principle is applied to the sufficient statistics. This is the more popular order not only because the Sufficiency Principle usually produces the greater data reduction, but also because of its more fundamental appeal. Usually the same results will be obtained regardless of the order in which the principles are used. (General conditions under which this holds are given in Hall, Wijsman, and Ghosh (1965).) An example of the use of the Sufficiency Principle followed by the Invariance Principle is given in the following example.

**Example 6.3.5:** Consider again the normal model from Example 6.3.2. By Example 6.1.4, the sample mean and variance,  $(\bar{X}, S^2)$ , constitute a sufficient statistic for this model. Now we apply the Invariance Principle to the observation  $(\bar{x}, s^2)$ . For  $-\infty < a < \infty$ , define  $g_a(\bar{x}, s^2) = (\bar{x} + a, s^2)$ . Arguments similar to those in Example 6.3.2 show that this set of transformations is a group. The set of possible joint distributions for  $(\bar{X}, S^2)$  is defined by the fact that  $\bar{X}$  and  $S^2$  are independent,  $\bar{X} \sim n(\mu, \sigma^2/n)$  and  $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$ . This set of joint distributions is invariant under the group we have defined because the joint distribution of  $g_a(\bar{X}, S^2) = (\bar{X} + a, S^2)$  is described by the fact that  $\bar{X} + a$  and  $S^2$  are independent,  $\bar{X} + a \sim n(\mu + a, \sigma^2/n)$  and  $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$ . This set of possible joint distributions is the same as in the untransformed problem. Now suppose we wish to estimate  $\sigma^2$ . An argument using the Invariance Principle, as in Example 6.3.4, shows that any invariant estimate of  $\sigma^2$  must satisfy

$$(6.3.6) \quad T(\bar{x}, s^2) = T(\bar{x} + a, s^2)$$

for all  $\bar{x}$ ,  $s^2$ , and  $a$ . Since  $a$  can be any real number, (6.3.6) implies that  $T(\bar{x}, s^2)$  depends on  $(\bar{x}, s^2)$  only through the value of  $s^2$ . Thus, sufficiency and invariance require that an estimate of  $\sigma^2$  be a function only of the sample variance  $s^2$ . This is a greater degree of data reduction than was obtained using either the Invariance Principle in Example 6.3.4 or the Sufficiency Principle in Example 6.1.4. ||

Remember, once again, that the Invariance Principle is composed of two distinct types of invariance. One type, measurement invariance, is intuitively reasonable. When many people think of the Invariance Principle, they think that it refers only to measurement invariance. If this were the case, the Invariance Principle would probably be universally accepted. But the other invariance, formal invariance, is quite different. It equates any two problems with the same mathematical structure, regardless of the physical reality they are trying to explain. It says that one inference procedure is appropriate *even if the physical realities are quite different*, an assumption that is sometimes difficult to justify.

Measurement invariance requires the same inference for two equivalent data points:  $x$ , measurements expressed in one scale, and  $y$ , *exactly the same measurements* expressed in a different scale. Formal invariance, in the end, leads to a relationship between the inferences at two *different* data points in the same measurement scale, a useful data reduction but not as intuitively justifiable. Example 6.3.6 illustrates measurement-invariant procedures that are not formally invariant.

**Example 6.3.6:** Suppose an experimenter wishes to estimate  $\theta$ , the mean boiling point of water, based on a single observation  $X$ , the boiling point measured in degrees Celsius. The experimenter knows that the number to be estimated is approximately 100°C although it may not equal this value exactly because of the altitude and impurities in the water. Thus, he decides to use the estimate  $T(x) = .5x + .5(100)$ . If the measurement scale is changed to degrees Fahrenheit, the experimenter would use  $T^*(y) = .5y + .5(212)$  to estimate the mean boiling point expressed in degrees Fahrenheit. Or, if he wanted to express the estimate in degrees Celsius, he would use  $\frac{5}{9}(T^*(y) - 32)$ , the familiar relation between degrees Celsius and degrees Fahrenheit. This procedure is measurement invariant in that the same answer will be obtained for the same data. That is, since  $x = \frac{5}{9}(y - 32)$ ,

$$\begin{aligned}\frac{5}{9}(T^*(y) - 32) &= \frac{5}{9} \left( T^* \left( \frac{9}{5}x + 32 \right) - 32 \right) \\ &= \frac{5}{9} \left( \left( .5 \left( \frac{9}{5}x + 32 \right) + .5(212) \right) - 32 \right) \\ &= .5x + .5(100) \\ &= T(x).\end{aligned}$$

If, for example, we assumed that  $X \sim n(\theta, \sigma^2)$ ,  $\theta$  and  $\sigma^2$  unknown, then  $Y = \frac{9}{5}X + 32$  also has a normal distribution with unknown mean and variance. Formal invariance would thus require that  $T(x) = T^*(x)$  for all  $x$ . Obviously the estimators we have defined above do not satisfy this. So they are not invariant in the sense of the Invariance Principle. ||

The Invariance Principle is a data reduction technique that restricts inference by prescribing what other inferences must be made at related sample points. For example, (6.3.4) gives the relationship between inferences at the related sample points  $(x_1, \dots, x_n)$  and  $(x_1 + a, \dots, x_n + a)$ . In this sense, the Invariance Principle is similar to both the Sufficiency Principle and the Likelihood Principle. All three principles prescribe similar relationships between inferences at different sample points. The Sufficiency Principle states that the inference should be the same at any two sample points that yield the same value of a sufficient statistic. The Likelihood Principle makes the same prescription for any two sample points that yield proportional likelihood functions. Thus, all three data reduction techniques restrict the set of allowable inferences and, in this way, simplify the analysis of the problem.

**EXERCISES**

- 6.1** Let  $X$  be one observation from a  $n(0, \sigma^2)$  population. Is  $|X|$  a sufficient statistic?
- 6.2** Let  $X_1, \dots, X_n$  be independent random variables with densities

$$f_{X_i}(x|\theta) = \begin{cases} e^{i\theta-x} & x \geq i\theta \\ 0 & x < i\theta \end{cases}.$$

Prove that  $T = \min_i(X_i/i)$  is a sufficient statistic for  $\theta$ .

- 6.3** Let  $X_1, \dots, X_n$  be a random sample from the pdf

$$f(x|\mu, \sigma) = \frac{1}{\sigma} e^{-(x-\mu)/\sigma}, \quad \mu < x < \infty, \quad 0 < \sigma < \infty.$$

Find a two-dimensional sufficient statistic for  $(\mu, \sigma)$ .

- 6.4** Prove Theorem 6.1.3.

- 6.5** Let  $X_1, \dots, X_n$  be independent random variables with pdfs

$$f(x_i|\theta) = \begin{cases} \frac{1}{2i\theta} & -i(\theta-1) < x_i < i(\theta+1) \\ 0 & \text{otherwise} \end{cases},$$

where  $\theta > 0$ . Find a two-dimensional sufficient statistic for  $\theta$ .

- 6.6** Let  $X_1, \dots, X_n$  be a random sample from a gamma( $\alpha, \beta$ ) population. Find a two-dimensional sufficient statistic for  $(\alpha, \beta)$ .

- 6.7** Let  $X_1, \dots, X_n$  be a random sample from a population with pdf or pmf  $f(x|\theta)$ . Show that the order statistics,  $T(X_1, \dots, X_n) = (X_{(1)}, \dots, X_{(n)})$ , are a sufficient statistic for  $\theta$ .

- 6.8** Let  $f(x, y|\theta_1, \theta_2, \theta_3, \theta_4)$  be the bivariate pdf for the uniform distribution on the rectangle with lower left corner  $(\theta_1, \theta_2)$  and upper right corner  $(\theta_3, \theta_4)$  in  $\mathbb{R}^2$ . The parameters satisfy  $\theta_1 < \theta_3$  and  $\theta_2 < \theta_4$ . Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be a random sample from this pdf. Find a four-dimensional sufficient statistic for  $(\theta_1, \theta_2, \theta_3, \theta_4)$ .

- 6.9** For each of the following distributions let  $X_1, \dots, X_n$  be a random sample. Find a minimal sufficient statistic for  $\theta$ .

a.  $f(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty$  (normal)

b.  $f(x|\theta) = e^{-(x-\theta)}, \quad \theta < x < \infty, \quad -\infty < \theta < \infty$  (location exponential)

c.  $f(x|\theta) = \frac{e^{-(x-\theta)}}{(1+e^{-(x-\theta)})^2}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty$  (logistic)

d.  $f(x|\theta) = \frac{1}{\pi[1+(x-\theta)^2]}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty$  (Cauchy)

e.  $f(x|\theta) = \frac{1}{2} e^{-|x-\theta|}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty$  (double exponential)

- 6.10** Show that the minimal sufficient statistic for the uniform( $\theta, \theta + 1$ ), found in Example 6.1.7, is not complete.

- 6.11** Refer to the pdfs given in Exercise 6.9. For each, let  $X_{(1)} < \dots < X_{(n)}$  be the ordered sample, and define  $Y_i = X_{(n)} - X_{(i)}, i = 1, \dots, n-1$ .

- a. For each of the pdfs in Exercise 6.9, verify that the set  $(Y_1, \dots, Y_{n-1})$  is ancillary for  $\theta$ . Try to prove a general theorem, like Example 6.1.9, that handles all these families at once.

- b. In each case determine whether the set  $(Y_1, \dots, Y_{n-1})$  is independent of the minimal sufficient statistic.

- 6.12** A natural ancillary statistic in most problems is the *sample size*. For example, let  $N$  be a random variable taking values  $1, 2, \dots$  with probabilities  $p_1, p_2, \dots$  where  $\sum p_i = 1$ .

Having observed  $N = n$ , perform  $n$  Bernoulli trials with success probability  $\theta$ , getting  $X$  successes.

a. Prove that the pair  $(X, N)$  is minimal sufficient and  $N$  is ancillary for  $\theta$ . (Note the similarity to some of the hierarchical models discussed in Section 4.4.)

b. Prove that the estimator  $X/N$  is unbiased for  $\theta$  and has variance  $\theta(1 - \theta)E(1/N)$ .

- 6.13 Suppose  $X_1$  and  $X_2$  are iid observations from the pdf  $f(x|\alpha) = \alpha x^{\alpha-1} e^{-x^\alpha}$ ,  $x > 0$ ,  $\alpha > 0$ . Show that  $(\log X_1)/(\log X_2)$  is an ancillary statistic.

- 6.14 Let  $X_1, \dots, X_n$  be a random sample from a location family. Show that  $M - \bar{X}$  is an ancillary statistic where  $M$  is the sample median.

- 6.15 Let  $X_1, \dots, X_n$  be iid  $n(\theta, a\theta^2)$ , where  $a$  is a known constant and  $\theta > 0$ . Show that the statistic  $T = (\bar{X}, S^2)$  is a sufficient statistic for  $\theta$ , but the family of distributions is not complete.

- 6.16 Let  $X_1, \dots, X_n$  be iid with geometric distribution

$$P_\theta(X = x) = \theta(1 - \theta)^{x-1}, \quad x = 1, 2, \dots, \quad 0 < \theta < 1.$$

Show that  $\sum X_i$  is sufficient for  $\theta$ , and find the family of distributions of  $\sum X_i$ . Is the family complete?

- 6.17 Let  $X_1, \dots, X_n$  be iid Poisson( $\lambda$ ). Show that the family of distributions of  $\sum X_i$  is complete. Prove completeness without using Theorem 6.1.6.

- 6.18 The random variable  $X$  takes the values 0, 1, 2 according to one of the following distributions:

	$P(X = 0)$	$P(X = 1)$	$P(X = 2)$	
Distribution 1	$p$	$3p$	$1 - 4p$	$0 < p < \frac{1}{4}$
Distribution 2	$p$	$p^2$	$1 - p - p^2$	$0 < p < \frac{1}{2}$

In each case determine whether the family of distributions of  $X$  is complete.

- 6.19 For each of the following pdfs let  $X_1, \dots, X_n$  be iid observations. Find a complete sufficient statistic, or show that one does not exist.

a.  $f(x|\theta) = \frac{2x}{\theta^2}$ ,  $0 < x < \theta$ ,  $\theta > 0$

b.  $f(x|\theta) = \frac{\theta}{(1+x)^{1+\theta}}$ ,  $0 < x < \infty$ ,  $\theta > 0$

c.  $f(x|\theta) = \frac{(\log \theta)\theta^x}{\theta-1}$ ,  $0 < x < 1$ ,  $\theta > 1$

d.  $f(x|\theta) = e^{-(x-\theta)} \exp(-e^{-(x-\theta)})$ ,  $-\infty < x < \infty$ ,  $-\infty < \theta < \infty$

e.  $f(x|\theta) = \binom{2}{x} \theta^x (1-\theta)^{2-x}$ ,  $x = 0, 1, 2$ ,  $0 \leq \theta \leq 1$

- 6.20 Let  $X$  be one observation from the pdf

$$f(x|\theta) = \left(\frac{\theta}{2}\right)^{|x|} (1-\theta)^{1-|x|}, \quad x = -1, 0, 1, \quad 0 \leq \theta \leq 1.$$

- a. Is  $X$  a complete sufficient statistic?

- b. Is  $|X|$  a complete sufficient statistic?

- c. Does  $f(x|\theta)$  belong to the exponential class?

- 6.21 Let  $X_1, \dots, X_n$  be a random sample from a population with pdf

$$f(x|\theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad \theta > 0.$$

- a. Is  $\sum X_i$  sufficient for  $\theta$ ?  
 b. Find a complete sufficient statistic for  $\theta$ .
- 6.22 Let  $X_1, \dots, X_n$  be a random sample from a uniform distribution on the interval  $(\theta, 2\theta)$ ,  $\theta > 0$ . Find a minimal sufficient statistic for  $\theta$ . Is the statistic complete?
- 6.23 Consider the following family of distributions:

$$\mathcal{P} = \{P_\lambda(X = x) : P_\lambda(X = x) = \lambda^x e^{-\lambda} / x!; x = 0, 1, 2, \dots; \lambda = 0 \text{ or } 1\}.$$

This is a Poisson family with  $\lambda$  restricted to be 0 or 1. Show that the family  $\mathcal{P}$  is *not complete*, demonstrating that completeness can be dependent on the range of the parameter. (See Exercises 6.15 and 6.17.)

- 6.24 Let  $X_1, \dots, X_n$  be a random sample from the pdf  $f(x|\mu) = e^{-(x-\mu)}$ , where  $-\infty < \mu < x < \infty$ .
- Show that  $X_{(1)} = \min_i X_i$  is a complete sufficient statistic.
  - Use Basu's Theorem to show that  $X_{(1)}$  and  $S^2$  are independent.
- 6.25 Prove the Likelihood Principle Corollary. That is, assuming both the Formal Sufficiency Principle and the Conditionality Principle, prove that if  $E = (\mathbf{X}, \theta, \{f(x|\theta)\})$  is an experiment then  $\text{Ev}(E, \mathbf{x})$  should depend on  $E$  and  $\mathbf{x}$  only through  $L(\theta|\mathbf{x})$ .
- 6.26 Consider the model in Exercise 6.12. Show that the Formal Likelihood Principle implies that any conclusions about  $\theta$  should not depend on the fact that the sample size  $n$  was chosen randomly. That is, the likelihood for  $(n, \mathbf{x})$ , a sample point from Exercise 6.12, is proportional to the likelihood for the sample point  $\mathbf{x}$ , a sample point from a fixed-sample-size binomial( $n, \theta$ ) experiment.
- 6.27 A risky experimental treatment is to be given to at most three patients. The treatment will be given to one patient. If it is a success, then it will be given to a second. If it is a success, it will be given to a third patient. Model the outcomes for the patients as independent Bernoulli( $p$ ) random variables. Identify the four sample points in this model and show that, according to the Formal Likelihood Principle, the inference about  $p$  should not depend on the fact that the sample size was determined by the data.
- 6.28 In Definition 6.3.1, show that (iii) is implied by (i) and (ii).
- 6.29 Let  $X_1, \dots, X_n$  be iid observations from a location-scale family. Let  $T_1(X_1, \dots, X_n)$  and  $T_2(X_1, \dots, X_n)$  be two statistics that both satisfy

$$T_i(ax_1 + b, \dots, ax_n + b) = aT_i(x_1, \dots, x_n)$$

for all values of  $x_1, \dots, x_n$  and  $b$  and for any  $a > 0$ .

- Show that  $T_1/T_2$  is an ancillary statistic.
- Let  $R$  be the sample range and  $S$  be the sample standard deviation. Verify that  $R$  and  $S$  satisfy the above condition so that  $R/S$  is an ancillary statistic.

## Miscellanea

---

### The Converse of Basu's Theorem

An interesting statistical fact is that the converse of Basu's Theorem is false. That is, if  $T(\mathbf{X})$  is independent of every ancillary statistic, it does not necessarily follow that  $T(\mathbf{X})$  is a complete, minimal sufficient statistic. A particularly nice treatment of the topic is given by Lehmann (1981). He makes the point that one reason that the converse fails is that ancillarity

is a property of the *entire distribution* of a statistic, whereas completeness is a property dealing only with *expectations*. Consider the following modification of the definition of ancillarity.

**Definition:** A statistic  $V(\mathbf{X})$  is called *first-order ancillary* if  $E_\theta V(\mathbf{X})$  is independent of  $\theta$ .

Lehmann then proves the following theorem, which is somewhat of a converse to Basu's Theorem.

**Theorem:** Let  $T$  be a statistic with  $\text{Var } T < \infty$ . A *necessary and sufficient* condition for  $T$  to be complete is that every bounded first-order ancillary  $V$  is uncorrelated (for all  $\theta$ ) with every bounded real-valued function of  $T$ .  $\square$

Lehmann also notes that a type of converse is also obtainable if, instead of modifying the definition of ancillarity, the definition of completeness is modified.

### **Confusion About Ancillarity**

One of the problems with the concept of *ancillarity* is that there are many different definitions of ancillarity, and different properties are given in these definitions. As was seen in this chapter, ancillarity is confusing enough with one definition—with five or six the situation becomes hopeless.

As told by Buehler (1982), the concept of ancillarity goes back to Sir Ronald Fisher (1925), "who left a characteristic trail of intriguing concepts but no definition." Buehler goes on to tell of at least *three* definitions of ancillarity, crediting, among others, Basu (1959) and Cox and Hinkley (1974). Buehler gives eight properties of ancillary statistics, and lists 25 examples.

### **The Likelihood Function as a Minimal Sufficient Statistic**

There is a striking similarity between the statement of Theorem 6.1.4 and the Likelihood Principle. Both relate to the ratio  $L(\theta|\mathbf{x})/L(\theta|\mathbf{y})$ , one to describe a minimal sufficient statistic, and the other to describe the Likelihood Principle. In fact, these theorems can be combined, with a bit of care, into the fact that a statistic  $T(\mathbf{x})$  is a minimal sufficient statistic if, and only if, it is a one-to-one function of  $L(\theta|\mathbf{x})$  (where two sample points that satisfy (6.2.1) are said to have the same likelihood function). Example 6.2.2 and Exercise 6.9 illustrate this point.

### **More on Sufficiency**

We may ask, "If there are *sufficient* statistics, why aren't there *necessary* statistics?" In fact, there are. According to Dynkin (1951), we have the following definition.

**Definition:** A statistic is said to be *necessary* if it can be written as a function of every sufficient statistic.

Comparing the definition of a necessary statistic and the definition of a minimal sufficient statistic, it should come as no surprise that we have the following theorem.

**Theorem:** A statistic is a minimal sufficient statistic if, and only if, it is a necessary and sufficient statistic.

# 7 Point Estimation

*“...when you have eliminated all which is impossible, then whatever remains, however improbable, must be the truth.”*

**Sherlock Holmes**

*The Adventure of the Blanched Soldier*

## 7.1 Introduction

This chapter is divided into two parts. The first part deals with methods for finding estimators, and the second part deals with evaluating these (and other) estimators. In general these two activities are intertwined. Often the methods of evaluating estimators will suggest new ones. However, for the time being, we will make the distinction between finding estimators and evaluating them.

The rationale behind point estimation is quite simple. When sampling is from a population described by a pdf or pmf  $f(x|\theta)$ , knowledge of  $\theta$  yields knowledge of the entire population. Hence, it is natural to seek a method of finding a good estimator of the point  $\theta$ , that is, a good point estimator. It is also the case that the parameter  $\theta$  has a meaningful physical interpretation (as in the case of a population mean) so there is direct interest in obtaining a good point estimate of  $\theta$ . It may also be the case that some function of  $\theta$ , say  $\tau(\theta)$ , is of interest. The methods described in this chapter can also be used to obtain estimators of  $\tau(\theta)$ .

The following definition of a point estimator may seem unnecessarily vague. However, at this point, we want to be careful not to eliminate any candidates from consideration.

**DEFINITION 7.1.1:** A *point estimator* is any function  $W(X_1, \dots, X_n)$  of a sample. That is, any statistic is a point estimator.

Notice that the definition makes no mention of any correspondence between the estimator and the parameter it is to estimate. While it might be argued that such a statement should be included in the definition, such a statement would restrict the available set of estimators. Also, there is no mention in the definition of the range of the statistic  $W(X_1, \dots, X_n)$ . While, in principle, the range of the statistic should coincide with that of the parameter, we will see that this is not always the case.

There is one distinction that must be made clear, the difference between an estimate and an estimator. An *estimator* is a function of the sample, while an *estimate* is the realized value of an estimator (that is, a number) that is obtained when a sample is actually taken. Notationally, when a sample is taken, an estimator is a function

of the random variables  $X_1, \dots, X_n$ , while an estimate is a function of the realized values  $x_1, \dots, x_n$ .

In many cases, there will be an obvious or natural candidate for a point estimator of a particular parameter. For example, the sample mean is a natural candidate for a point estimator of the population mean. However, when we leave a simple case like this, intuition may not only desert us, it may also lead us astray. Therefore, it is useful to have some techniques that will at least give us some reasonable candidates for consideration. Be advised that these techniques do not carry any guarantees with them. The point estimators that they yield still must be evaluated before their worth is established.

## 7.2 Methods of Finding Estimators

In some cases it is an easy task to decide how to estimate a parameter, and often intuition alone can lead us to very good estimators. For example, estimating a parameter with its sample analogue is usually reasonable. In particular, the sample mean is a good estimate for the population mean. In more complicated models, ones that often arise in practice, we need a more methodical way of estimating parameters. In this section we detail four methods of finding estimators.

### 7.2.1 Method of Moments

The method of moments is, perhaps, the oldest method of finding point estimators, dating back at least to Karl Pearson in the late 1800s. It has the virtue of being quite simple to use and almost always yields some sort of estimate. In many cases, unfortunately, this method yields estimators that may be improved upon. However, it is a good place to start when other methods prove intractable.

Let  $X_1, \dots, X_n$  be a sample from a population with pdf or pmf  $f(x|\theta_1, \dots, \theta_k)$ . Method of moments estimators are found by equating the first  $k$  sample moments to the corresponding  $k$  population moments, and solving the resulting system of simultaneous equations. More precisely, define

$$\begin{aligned}
 m_1 &= \frac{1}{n} \sum_{i=1}^n X_i^1, & \mu_1 &= EX^1; \\
 m_2 &= \frac{1}{n} \sum_{i=1}^n X_i^2, & \mu_2 &= EX^2; \\
 &\vdots && \\
 m_k &= \frac{1}{n} \sum_{i=1}^n X_i^k, & \mu_k &= EX^k.
 \end{aligned}
 \tag{7.2.1}$$

The population moment  $\mu_j$  will typically be a function of  $\theta_1, \dots, \theta_k$ , say  $\mu_j(\theta_1, \dots, \theta_k)$ . The method of moments estimator  $(\tilde{\theta}_1, \dots, \tilde{\theta}_k)$  of  $(\theta_1, \dots, \theta_k)$  is obtained by solving the following system of equations for  $(\theta_1, \dots, \theta_k)$  in terms of  $(m_1, \dots, m_k)$ :

$$(7.2.2) \quad \begin{aligned} m_1 &= \mu_1(\theta_1, \dots, \theta_k), \\ m_2 &= \mu_2(\theta_1, \dots, \theta_k), \\ &\vdots \\ m_k &= \mu_k(\theta_1, \dots, \theta_k). \end{aligned}$$

**Example 7.2.1:** Suppose  $X_1, \dots, X_n$  are iid  $n(\theta, \sigma^2)$ . In the preceding notation,  $\theta_1 = \theta$  and  $\theta_2 = \sigma^2$ . We have  $m_1 = \bar{X}$ ,  $m_2 = (1/n) \sum X_i^2$ ,  $\mu_1 = \theta$ ,  $\mu_2 = \theta^2 + \sigma^2$ , and hence we must solve

$$\bar{X} = \theta, \quad \frac{1}{n} \sum X_i^2 = \theta^2 + \sigma^2.$$

Solving for  $\theta$  and  $\sigma^2$  yields the method of moments estimators

$$\tilde{\theta} = \bar{X} \quad \text{and} \quad \tilde{\sigma}^2 = \frac{1}{n} \sum X_i^2 - \bar{X}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2. \quad ||$$

In this simple example, the method of moments solution coincides with our intuition, and perhaps give some credence to both. The method is somewhat more helpful, however, when no obvious estimator suggests itself.

**Example 7.2.2:** Let  $X_1, \dots, X_n$  be iid binomial( $k, p$ ), that is,

$$P(X_i = x | k, p) = \binom{k}{x} p^x (1-p)^{k-x}, \quad x = 0, 1, \dots, k.$$

Here we assume that both  $k$  and  $p$  are unknown and we desire point estimators for both parameters. (This somewhat unusual application of the binomial model has been used to estimate crime rates for crimes that are known to have many unreported occurrences. For such a crime, both the true reporting rate,  $p$ , and the total number of occurrences,  $k$ , are unknown.)

Equating the first two sample moments to those of the population yields the system of equations

$$\begin{aligned} \bar{X} &= kp, \\ \frac{1}{n} \sum X_i^2 &= kp(1-p) + k^2 p^2, \end{aligned}$$

which now must be solved for  $k$  and  $p$ . After a little algebra, we obtain the method of moments estimators

$$\tilde{k} = \frac{\bar{X}^2}{\bar{X} - (1/n) \sum (X_i - \bar{X})^2},$$

and

$$\tilde{p} = \frac{\bar{X}}{\tilde{k}}.$$

Admittedly, these are not the best estimators for the population parameters. In particular, it is possible to get negative estimates of  $k$  and  $p$  which, of course, must be positive numbers. (This is a case where the range of the estimator does not coincide with the range of the parameter it is estimating.) However, in fairness to the method of moments, note that negative estimates will occur only when the sample mean is smaller than the sample variance, indicating a large degree of variability in the data. The method of moments has, in this case, at least given us a set of candidates for point estimators of  $k$  and  $p$ . Although our intuition may have given us a candidate for an estimator of  $p$ , coming up with an estimator of  $k$  is much more difficult. ||

The method of moments can be very useful in obtaining approximations to the distributions of statistics. This technique, sometimes called "moment matching," gives us an approximation that is based on matching moments of distributions. In theory, the moments of the distribution of any statistic can be matched to those of any distribution but, in practice, it is best to use distributions that are similar. The following example illustrates one of the most famous uses of this technique, the approximation of Satterthwaite (1946). It is still used today (see Exercise 8.52).

**Example 7.2.3:** If  $Y_i, i = 1, \dots, k$ , are independent  $\chi_{r_i}^2$  random variables, we have already seen (Lemma 5.4.1) that the distribution of  $\sum Y_i$  is also chi squared, with degrees of freedom equal to  $\sum r_i$ . Unfortunately, the distribution of  $\sum a_i Y_i$ , where the  $a_i$ s are known constants, is, in general, quite difficult to obtain. It does seem reasonable, however, to assume that a  $\chi_\nu^2$ , for some value of  $\nu$ , will provide a good approximation.

This is almost Satterthwaite's problem. He was interested in approximating the denominator of a  $t$  statistic, and  $\sum a_i Y_i$  represented the square of the denominator of his statistic. Hence, for given  $a_1, \dots, a_k$ , he wanted to find a value of  $\nu$  so that

$$\sum_{i=1}^k a_i Y_i \sim \frac{\chi_\nu^2}{\nu}. \quad (\text{approximately})$$

Since  $E(\chi_\nu^2/\nu) = 1$ , to match first moments we need

$$\mathbb{E} \left( \sum_{i=1}^k a_i Y_i \right) = \sum_{i=1}^k a_i \mathbb{E} Y_i = \sum_{i=1}^k a_i r_i = 1,$$

which gives us a constraint on the  $a_i$ 's but gives us no information on how to estimate  $\nu$ . To do this we must match second moments, and we need

$$\mathbb{E} \left( \sum_{i=1}^k a_i Y_i \right)^2 = \mathbb{E} \left( \frac{\chi_\nu^2}{\nu} \right)^2 = \frac{2}{\nu} + 1.$$

Applying the method of moments, we drop the first expectation and solve for  $\nu$ , yielding

$$\hat{\nu} = \frac{2}{(\sum_{i=1}^k a_i Y_i)^2 - 1}.$$

Thus, straightforward application of the method of moments yields an estimator of  $\nu$ , but one that can be negative. We might suppose that Satterthwaite was aghast at this possibility, for this is not the estimator he proposed. Working much harder, he customized the method of moments in the following way. Write

$$\begin{aligned} \mathbb{E} \left( \sum a_i Y_i \right)^2 &= \text{Var} \left( \sum a_i Y_i \right) + \left( \mathbb{E} \sum a_i Y_i \right)^2 \\ &= \left( \mathbb{E} \sum a_i Y_i \right)^2 \left[ \frac{\text{Var}(\sum a_i Y_i)}{(\mathbb{E} \sum a_i Y_i)^2} + 1 \right] \\ &= \left[ \frac{\text{Var}(\sum a_i Y_i)}{(\mathbb{E} \sum a_i Y_i)^2} + 1 \right]. \end{aligned} \quad (\mathbb{E} \sum a_i Y_i = 1)$$

Now equate second moments to obtain

$$\nu = \frac{2(\mathbb{E} \sum a_i Y_i)^2}{\text{Var}(\sum a_i Y_i)}.$$

Finally, use the fact that  $Y_1, \dots, Y_k$  are independent chi squared random variables to write

$$\begin{aligned} \text{Var} \left( \sum a_i Y_i \right) &= \sum a_i^2 \text{Var} Y_i \\ &= 2 \sum \frac{a_i^2 (\mathbb{E} Y_i)^2}{r_i}. \quad (\text{Var } Y_i = 2(\mathbb{E} Y_i)^2 / r_i) \end{aligned}$$

Substituting this expression for the variance, and removing the expectations, we obtain Satterthwaite's estimator

$$\hat{\nu} = \frac{(\sum a_i Y_i)^2}{\sum \frac{a_i^2}{r_i} Y_i^2}.$$

This approximation is quite good, and is still widely used today. Notice that Satterthwaite succeeded in obtaining an estimator that is always positive, thus alleviating the obvious problems with the straightforward method of moments estimator.

### 7.2.2 Maximum Likelihood Estimators

The method of maximum likelihood is, by far, the most popular technique for deriving estimators. Recall that if  $X_1, \dots, X_n$  are an iid sample from a population with pdf or pmf  $f(x|\theta_1, \dots, \theta_k)$ , the likelihood function is defined by

$$(7.2.3) \quad L(\theta|x) = L(\theta_1, \dots, \theta_k|x_1, \dots, x_n) = \prod_{i=1}^n f(x_i|\theta_1, \dots, \theta_k).$$

**DEFINITION 7.2.1:** For each sample point  $x$ , let  $\hat{\theta}(x)$  be a parameter value at which  $L(\theta|x)$  attains its maximum as a function of  $\theta$ , with  $x$  held fixed. A *maximum likelihood estimator* (MLE) of the parameter  $\theta$  based on a sample  $X$  is  $\hat{\theta}(X)$ .

Notice that, by its construction, the range of the MLE coincides with the range of the parameter. We also use the abbreviation MLE to stand for maximum likelihood *estimate*, when we are talking of the realized value of the estimator.

Intuitively, the MLE is a reasonable choice for an estimator. The MLE is the parameter point for which the observed sample is most likely. In general, the MLE is a good point estimator, possessing some of the optimality properties discussed later.

There are two inherent drawbacks associated with the general problem of finding the maximum of a function, and hence of maximum likelihood estimation. The first problem is that of actually finding the global maximum and verifying that, indeed, a global maximum has been found. In many cases this problem reduces to a simple differential calculus exercise but, sometimes even for common densities, difficulties do arise. The second problem is that of numerical sensitivity. That is, how sensitive is the estimate to small changes in the data? (Strictly speaking, this is a mathematical rather than statistical problem associated with any maximization procedure. Since an MLE is found through a maximization procedure, however, it is a problem that we must deal with.) Unfortunately, it is sometimes the case that a slightly different sample will produce a vastly different MLE, making its use suspect. We consider first the problem of finding MLEs.

If the likelihood function is differentiable (in  $\theta_i$ ), possible candidates for the MLE are the values of  $(\theta_1, \dots, \theta_k)$  that solve

$$(7.2.4) \quad \frac{\partial}{\partial \theta_i} L(\theta|x) = 0, \quad i = 1, \dots, k.$$

Note that the solutions to (7.2.4) are only *possible candidates* for the MLE since the first derivative being zero is only a necessary condition for a maximum, not a sufficient condition. Furthermore, the zeros of the first derivative only locate extreme points in the interior of the domain of a function. If the extrema occur on the

boundary the first derivative may not be zero. Thus, the boundary must be checked separately for extrema.

Points at which the first derivatives are zero may be local or global minima, local or global maxima, or inflection points. Our job is to find a global maximum.

**Example 7.2.4:** Let  $X_1, \dots, X_n$  be iid  $n(\theta, 1)$ , and let  $L(\theta|x)$  denote the likelihood function. Then

$$L(\theta|x) = \prod_{i=1}^n \frac{1}{(2\pi)^{1/2}} e^{-(1/2)(x_i - \theta)^2} = \frac{1}{(2\pi)^{n/2}} e^{(-1/2)\sum_{i=1}^n (x_i - \theta)^2}.$$

The equation  $(d/d\theta)L(\theta|x) = 0$  reduces to

$$\sum_{i=1}^n (x_i - \theta) = 0,$$

which has the solution  $\hat{\theta} = \bar{x}$ . Hence,  $\bar{x}$  is a candidate for the MLE. To verify that  $\bar{x}$  is, in fact, a global maximum of the likelihood function, we can use the following argument. First note that  $\hat{\theta} = \bar{x}$  is the only solution to  $\sum(x_i - \theta) = 0$ , hence  $\bar{x}$  is the only zero of the first derivative. Second, verify that

$$\frac{d^2}{d\theta^2} L(\theta|x)|_{\theta=\bar{x}} < 0.$$

Thus,  $\bar{x}$  is the only extreme point in the interior and it is a maximum. To finally verify that  $\bar{x}$  is a global maximum, we must check the boundaries,  $\pm\infty$ . By taking limits it is easy to establish that the likelihood is zero at  $\pm\infty$ . So  $\hat{\theta} = \bar{x}$  is a global maximum and hence  $\bar{X}$  is the MLE. (Actually, we can be a bit more clever and avoid checking  $\pm\infty$ . Since we established that  $\bar{x}$  is a *unique* interior extremum, and is a maximum, there can be no maximum at  $\pm\infty$ . If there were, then there would have to be an interior minimum, which contradicts uniqueness.) ||

Another way to find an MLE is to abandon differentiation and proceed with a direct maximization. This method is usually simpler algebraically, especially if the derivatives tend to get messy, but is sometimes harder to implement because there are no set rules to follow. One general technique is to find a global upper bound on the likelihood function and then establish that there is a unique point for which the upper bound is attained.

**Example 7.2.4 (Continued):** Recall (Theorem 5.2.1) that for any number  $a$ ,

$$\sum_{i=1}^n (x_i - a)^2 \geq \sum_{i=1}^n (x_i - \bar{x})^2$$

with equality if, and only if,  $a = \bar{x}$ . This implies that for any  $\theta$ ,

$$e^{-(1/2)\sum(x_i-\theta)^2} \leq e^{-(1/2)\sum(x_i-\bar{x})^2}$$

with equality if, and only if,  $\theta = \bar{x}$ . Hence  $\bar{X}$  is the MLE. ||

In most cases, especially when differentiation is to be used, it is easier to work with the natural logarithm of  $L(\theta|x)$ ,  $\log L(\theta|x)$  (known as the *log likelihood*), than it is to work with  $L(\theta|x)$  directly. This is possible because the log function is strictly increasing on  $(0, \infty)$ , which implies that the extrema of  $L(\theta|x)$  and  $\log L(\theta|x)$  coincide (see Exercise 7.3).

**Example 7.2.5:** Let  $X_1, \dots, X_n$  be iid Bernoulli( $p$ ). Then the likelihood function is

$$L(p|x) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^y (1-p)^{n-y}$$

where  $y = \sum x_i$ . While this function is not all that hard to differentiate, it is much easier to differentiate the log likelihood

$$\log L(p|x) = y \log p + (n-y) \log(1-p).$$

If  $0 < y < n$ , differentiating  $\log L(p|x)$  and setting the result equal to zero gives the solution,  $\hat{p} = y/n$ . It is also straightforward to verify that  $y/n$  is the global maximum in this case. If  $y = 0$  or  $y = n$ , then

$$\log L(p|x) = \begin{cases} n \log(1-p) & \text{if } y = 0 \\ n \log p & \text{if } y = n \end{cases}$$

In either case  $\log L(p|x)$  is a monotone function of  $p$ , and it is again straightforward to verify that  $\hat{p} = y/n$  in each case. Thus, we have shown that  $\sum X_i/n$  is the MLE of  $p$ . ||

One other point to be aware of when finding a maximum likelihood estimator is that the maximization takes place only over the range of parameter values. In some cases this point plays an important part.

**Example 7.2.6:** Let  $X_1, \dots, X_n$  be iid  $n(\theta, 1)$ , where it is known that  $\theta$  must be nonnegative. With no restrictions on  $\theta$ , we saw that the MLE of  $\theta$  is  $\bar{X}$ ; however, if  $\bar{X}$  is negative it will be outside the range of the parameter.

If  $\bar{x}$  is negative, it is easy to check (see Exercise 7.4) that the likelihood function  $L(\theta|x)$  is decreasing in  $\theta$  for  $\theta \geq 0$  and is maximized at  $\hat{\theta} = 0$ . Hence, in this case, the MLE of  $\theta$  is

$$\hat{\theta} = \bar{X} \text{ if } \bar{X} \geq 0 \quad \text{and} \quad \hat{\theta} = 0 \text{ if } \bar{X} < 0. \quad ||$$

If  $L(\theta|\mathbf{x})$  cannot be maximized analytically, it may be possible to use a computer and maximize  $L(\theta|\mathbf{x})$  numerically. In fact, this is one of the most important features of MLEs. If a model (likelihood) can be written down, then there is some hope of maximizing it numerically and, hence, finding MLEs of the parameters. When this is done, there is still always the question of whether a local or global maximum has been found. Thus, it is always important to analyze the likelihood function as much as possible, to find the number and nature of its local maxima, before using numeric maximization.

**Example 7.2.7:** Let  $X_1, \dots, X_n$  be a random sample from a binomial( $k, p$ ) population where  $p$  is known and  $k$  is unknown. For example, we flip a coin we know to be fair and observe  $x_i$  heads but we do not know how many times the coin was flipped. The likelihood function is

$$L(k|\mathbf{x}) = \prod_{i=1}^n \binom{k}{x_i} p^{x_i} (1-p)^{k-x_i}.$$

Maximizing  $L(k|\mathbf{x})$  by differentiation is difficult because of the factorials and because  $k$  must be an integer. Thus we try a different approach.

Of course,  $L(k|\mathbf{x}) = 0$  if  $k < \max_i x_i$ . Thus the MLE is an integer  $\hat{k} \geq \max_i x_i$  that satisfies  $L(\hat{k}|\mathbf{x})/L(\hat{k}-1|\mathbf{x}) \geq 1$  and  $L(\hat{k}+1|\mathbf{x})/L(\hat{k}|\mathbf{x}) < 1$ . We will show that there is only one such  $\hat{k}$ . The ratio of likelihoods is

$$\frac{L(k|\mathbf{x})}{L(k-1|\mathbf{x})} = \frac{(k(1-p))^n}{\prod_{i=1}^n (k-x_i)}.$$

Thus the condition for a maximum is

$$(k(1-p))^n \geq \prod_{i=1}^n (k-x_i) \quad \text{and} \quad ((k+1)(1-p))^n < \prod_{i=1}^n (k+1-x_i).$$

Dividing by  $k^n$  and letting  $z = 1/k$ , we want to solve

$$(1-p)^n = \prod_{i=1}^n (1-x_i z)$$

for  $0 \leq z \leq 1/\max_i x_i$ . The right-hand side is clearly a strictly decreasing function of  $z$  for  $z$  in this range with a value of 1 at  $z = 0$  and a value of 0 at  $z = 1/\max_i x_i$ . Thus there is a unique  $z$  (call it  $\hat{z}$ ) that solves the equation. The quantity  $1/\hat{z}$  may not be an integer. But the integer  $\hat{k}$  that satisfies the inequalities, and is the MLE, is the largest integer less than or equal to  $1/\hat{z}$  (see Exercise 7.5). Thus, this analysis shows that there is a unique maximum for the likelihood function and it can be found by numerically solving an  $n$ th-degree polynomial equality. This description of the MLE

for  $k$  was found by Feldman and Fox (1968). See Example 7.2.2 (Continued) (page 297) for more about estimating  $k$ . ||

Perhaps one of the most useful properties of maximum likelihood estimators is what has come to be known as the *invariance property of maximum likelihood estimators* (not to be confused with the type of invariance discussed in Chapter 6). Suppose that a distribution is indexed by a parameter  $\theta$ , but the interest is in finding an estimator for some function of  $\theta$ , say  $\tau(\theta)$ . Informally speaking, the invariance property of MLEs says that if  $\hat{\theta}$  is the MLE of  $\theta$ , then  $\tau(\hat{\theta})$  is the MLE of  $\tau(\theta)$ . For example, if  $\theta$  is the mean of a normal distribution, the MLE of  $\sin(\theta)$  is  $\sin(\bar{X})$ . This is discussed in Zehna (1966).

There are, of course, some technical problems to be overcome before we can formalize this notion of invariance of MLEs, and they mostly focus on the function  $\tau(\theta)$  that we are trying to estimate. If the mapping  $\theta \rightarrow \tau(\theta)$  is one-to-one (that is, for each value of  $\theta$  there is a unique value of  $\tau(\theta)$ , and vice versa), then there is no problem. In this case, it is easy to see that it makes no difference whether we maximize the likelihood as a function of  $\theta$  or as a function of  $\tau(\theta)$  — in each case we get the same answer. If we let  $\eta = \tau(\theta)$ , then the inverse function  $\tau^{-1}(\eta) = \theta$  is well defined and the likelihood function of  $\tau(\theta)$ , written as a function of  $\eta$ , is given by

$$L^*(\eta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\tau^{-1}(\eta)) = L(\tau^{-1}(\eta)|\mathbf{x})$$

and

$$\sup_{\eta} L^*(\eta|\mathbf{x}) = \sup_{\eta} L(\tau^{-1}(\eta)|\mathbf{x}) = \sup_{\theta} L(\theta|\mathbf{x}).$$

Thus, the maximum of  $L^*(\eta|\mathbf{x})$  is attained at  $\eta = \tau(\theta) = \tau(\hat{\theta})$ , showing that the MLE of  $\tau(\theta)$  is  $\tau(\hat{\theta})$ .

In many cases, this simple version of the invariance of MLEs is not useful because many of the functions we are interested in are not one-to-one. For example, to estimate  $\theta^2$ , the square of a normal mean, the mapping  $\theta \rightarrow \theta^2$  is not one-to-one. Thus, we need a more general theorem and, in fact, a more general definition of the likelihood function of  $\tau(\theta)$ .

If  $\tau(\theta)$  is not one-to-one, then for a given value  $\eta$  there may be more than one value of  $\theta$  that satisfy  $\tau(\theta) = \eta$ . In such cases, the correspondence between the maximization over  $\eta$  and that over  $\theta$  can break down. For example, if  $\hat{\theta}$  is the MLE of  $\theta$ , there may be another value of  $\theta$ , say  $\theta_0$ , for which  $\tau(\hat{\theta}) = \tau(\theta_0)$ . We need to avoid such difficulties.

We proceed by defining for  $\tau(\theta)$  the *induced likelihood function*  $L^*$ , given by

$$(7.2.5) \quad L^*(\eta|\mathbf{x}) = \sup_{\{\theta: \tau(\theta)=\eta\}} L(\theta|\mathbf{x}).$$

The value  $\hat{\eta}$  which maximizes  $L^*(\eta|\mathbf{x})$  will be called the MLE of  $\eta = \tau(\theta)$ , and it can be seen from (7.2.5) that the maxima of  $L^*$  and  $L$  coincide. We are now ready to state the invariance theorem for MLEs.

**THEOREM 7.2.1 (Invariance Property of Maximum Likelihood Estimators):** If  $\hat{\theta}$  is the MLE of  $\theta$ , then for any function  $\tau(\theta)$ , the MLE of  $\tau(\theta)$  is  $\tau(\hat{\theta})$ .

*Proof:* Let  $\hat{\eta}$  denote the value that maximizes  $L^*(\eta|\mathbf{x})$ . We must show that  $L^*(\hat{\eta}|\mathbf{x}) = L^*[\tau(\hat{\theta})|\mathbf{x}]$ . Now, as stated above, the maxima of  $L$  and  $L^*$  coincide, so we have

$$\begin{aligned} L^*(\hat{\eta}|\mathbf{x}) &= \sup_{\eta} \sup_{\{\theta: \tau(\theta)=\eta\}} L(\theta|\mathbf{x}) && \text{(definition of } L^*) \\ &= \sup_{\theta} L(\theta|\mathbf{x}) \\ &= L(\hat{\theta}|\mathbf{x}), && \text{(definition of } \hat{\theta}) \end{aligned}$$

where the second equality follows because the iterated maximization is equal to the unconditional maximization over  $\theta$ , which is attained at  $\hat{\theta}$ . Furthermore

$$\begin{aligned} L(\hat{\theta}|\mathbf{x}) &= \sup_{\{\theta: \tau(\theta)=\tau(\hat{\theta})\}} L(\theta|\mathbf{x}) && (\hat{\theta} \text{ is the MLE}) \\ &= L^*[\tau(\hat{\theta})|\mathbf{x}]. && \text{(definition of } L^*) \end{aligned}$$

Hence, the string of equalities shows that  $L^*(\hat{\eta}|\mathbf{x}) = L^*[\tau(\hat{\theta})|\mathbf{x}]$ , and that  $\tau(\hat{\theta})$  is the MLE of  $\tau(\theta)$ .  $\square$

Using this theorem, we now see that the MLE of  $\theta^2$ , the square of a normal mean, is  $\bar{X}^2$ . We can also apply Theorem 7.2.1 to more complicated functions to see that, for example, the MLE of  $\sqrt{p(1-p)}$ , where  $p$  is a binomial probability, is given by  $\sqrt{\hat{p}(1-\hat{p})}$ .

Before leaving the subject of finding maximum likelihood estimators, there are a few more points to be mentioned.

The invariance property of MLEs also holds in the multivariate case. There is nothing in the proof of Theorem 7.2.1 that precludes  $\theta$  from being a vector. If the MLE of  $(\theta_1, \dots, \theta_k)$  is  $(\hat{\theta}_1, \dots, \hat{\theta}_k)$ , and if  $\tau(\theta_1, \dots, \theta_k)$  is any function of the parameters, the MLE of  $\tau(\theta_1, \dots, \theta_k)$  is  $\tau(\hat{\theta}_1, \dots, \hat{\theta}_k)$ .

If  $\theta = (\theta_1, \dots, \theta_k)$  is multidimensional, then the problem of finding an MLE is that of maximizing a function of several variables. If the likelihood function is differentiable, setting the first partial derivatives equal to zero provides a necessary condition for an extremum in the interior. However, in the multidimensional case, using a second derivative condition to check for a maximum is a tedious task, and other methods might be tried first. We first illustrate a technique that usually proves simpler, that of successive maximizations.

**Example 7.2.8:** Let  $X_1, \dots, X_n$  be iid  $n(\theta, \sigma^2)$ , with both  $\theta$  and  $\sigma^2$  unknown. Then

$$L(\theta, \sigma^2 | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_{i=1}^n (x_i - \theta)^2 / \sigma^2}$$

and

$$\log L(\theta, \sigma^2 | \mathbf{x}) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2 / \sigma^2.$$

The partial derivatives, with respect to  $\theta$  and  $\sigma^2$ , are

$$\frac{\partial}{\partial \theta} \log L(\theta, \sigma^2 | \mathbf{x}) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \theta)$$

and

$$\frac{\partial}{\partial \sigma^2} \log L(\theta, \sigma^2 | \mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \theta)^2.$$

Setting these partial derivatives equal to zero and solving yields the solution  $\hat{\theta} = \bar{x}$ ,  $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$ . To verify that this solution is, in fact, a global maximum, recall first that if  $\theta \neq \bar{x}$ , then  $\sum (x_i - \theta)^2 > \sum (x_i - \bar{x})^2$ . Hence, for any value of  $\sigma^2$ ,

$$\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_{i=1}^n (x_i - \bar{x})^2 / \sigma^2} \geq \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_{i=1}^n (x_i - \theta)^2 / \sigma^2}.$$

Therefore, verifying that we have found the maximum likelihood estimators is reduced to a one-dimensional problem, verifying that  $(\sigma^2)^{-n/2} \exp(-\frac{1}{2} \sum (x_i - \bar{x})^2 / \sigma^2)$  achieves its global maximum at  $\sigma^2 = n^{-1} \sum (x_i - \bar{x})^2$ . This is straightforward to do using univariate calculus and, in fact, the estimators  $(\bar{X}, n^{-1} \sum (X_i - \bar{X})^2)$  are the MLEs. ||

Now consider the solution to the same problem using two-variate calculus.

**Example 7.2.8 (Continued):** To use two-variate calculus to verify that a function  $H(\theta_1, \theta_2)$  has a maximum at  $(\hat{\theta}_1, \hat{\theta}_2)$ , it must be shown that the following three conditions hold.

- a. The first-order partial derivatives are zero,

$$\frac{\partial}{\partial \theta_1} H(\theta_1, \theta_2) \Big|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} = 0 \quad \text{and} \quad \frac{\partial}{\partial \theta_2} H(\theta_1, \theta_2) \Big|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} = 0.$$

b. At least one second-order partial derivative is negative,

$$\frac{\partial^2}{\partial \theta_1^2} H(\theta_1, \theta_2) \Big|_{\theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2} < 0 \quad \text{or} \quad \frac{\partial^2}{\partial \theta_2^2} H(\theta_1, \theta_2) \Big|_{\theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2} < 0.$$

c. The Jacobian of the second-order partial derivatives is positive,

$$\begin{aligned} & \left| \begin{array}{cc} \frac{\partial^2}{\partial \theta_1^2} H(\theta_1, \theta_2) & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} H(\theta_1, \theta_2) \\ \frac{\partial^2}{\partial \theta_2 \partial \theta_1} H(\theta_1, \theta_2) & \frac{\partial^2}{\partial \theta_2^2} H(\theta_1, \theta_2) \end{array} \right|_{\theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2} \\ &= \frac{\partial^2}{\partial \theta_1^2} H(\theta_1, \theta_2) \frac{\partial^2}{\partial \theta_2^2} H(\theta_1, \theta_2) - \left( \frac{\partial^2}{\partial \theta_1 \partial \theta_2} H(\theta_1, \theta_2) \right)^2 \Big|_{\theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2} > 0. \end{aligned}$$

For the normal log likelihood, the second-order partial derivatives are

$$\begin{aligned} \frac{\partial^2}{\partial \theta^2} \log L(\theta, \sigma^2 | \mathbf{x}) &= \frac{-n}{\sigma^2}, \\ \frac{\partial^2}{\partial (\sigma^2)^2} \log L(\theta, \sigma^2 | \mathbf{x}) &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \theta)^2, \\ \frac{\partial^2}{\partial \theta \partial \sigma^2} \log L(\theta, \sigma^2 | \mathbf{x}) &= -\frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \theta). \end{aligned}$$

Properties (a) and (b) are easily seen to hold, and the Jacobian is

$$\begin{aligned} & \left| \begin{array}{cc} \frac{-n}{\sigma^2} & -\frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \theta) \\ -\frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \theta) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \theta)^2 \end{array} \right|_{\theta = \bar{x}, \sigma^2 = \hat{\sigma}^2} \\ &= \frac{1}{\sigma^6} \left[ \frac{-n^2}{2} + \frac{n}{\sigma^2} \sum_{i=1}^n (x_i - \theta)^2 - \frac{1}{\sigma^2} \left( \sum_{i=1}^n (x_i - \theta) \right)^2 \right] \Big|_{\theta = \bar{x}, \sigma^2 = \hat{\sigma}^2} \\ &= \frac{1}{\hat{\sigma}^6} \left[ \frac{-n^2}{2} + \frac{n^2}{\hat{\sigma}^2} \hat{\sigma}^2 - \frac{1}{\hat{\sigma}^2} \left( \sum_{i=1}^n (x_i - \bar{x}) \right)^2 \right] \\ &= \frac{1}{\hat{\sigma}^6} \frac{n^2}{2} > 0. \end{aligned}$$

Thus, the calculus conditions are satisfied and we have indeed found a maximum. (Of course, to be really formal, we have verified that  $(\bar{x}, \hat{\sigma}^2)$  is an interior maximum. We still have to check that it is unique, and that there is no maximum at infinity.) The amount of calculation, even in this simple problem, is formidable, and things will only get worse. (Think of what we would have to do for three parameters.) Thus, the moral is that, while we always have to verify that we have, indeed,

found a maximum, we should look for ways to do it other than using second derivative conditions.

Finally, it was mentioned earlier that, since MLEs are found by a maximization process, they are susceptible to the problems associated with that process, among them that of numerical instability. We now look at this problem in more detail.

Recall that the likelihood function is a function of the parameter,  $\theta$ , with the data,  $\mathbf{x}$ , held constant. However, since the data are measured with error, we might ask how small changes in the data might affect the MLE. That is, we calculate  $\hat{\theta}$  based on  $L(\theta|\mathbf{x})$ , but we might inquire what value we would get for the MLE if we based our calculations on  $L(\theta|\mathbf{x} + \epsilon)$ , for small  $\epsilon$ . Intuitively, this new MLE, say  $\hat{\theta}_1$ , should be close to  $\hat{\theta}$  if  $\epsilon$  is small. But this is not always the case.

**Example 7.2.2 (Continued):** Olkin et al. (1981) demonstrate that the MLEs of  $k$  and  $p$  in binomial sampling can be highly unstable. They illustrate their case with the following example. Five realizations of a binomial( $k, p$ ) experiment are observed, where both  $k$  and  $p$  are unknown. The first data set is (16, 18, 22, 25, 27). (These are the observed numbers of successes from an unknown number of binomial trials.) For this data set, the MLE of  $k$  is  $\hat{k} = 99$ . If a second data set is (16, 18, 22, 25, 28), where the only difference is that the 27 is replaced with 28, then the MLE of  $k$  is  $\hat{k} = 190$ , demonstrating a large amount of variability.

Such occurrences happen when the likelihood function is very flat in the neighborhood of its maximum, or when there is no finite maximum. When the MLEs can be found explicitly, as will often be the case in our examples, this is usually not a problem. However, in many instances, such as in the above example, the MLE cannot be solved for explicitly, and must be found by numeric methods. When faced with such a problem, it is often wise to spend a little extra time investigating the stability of the solution.

### 7.2.3 Bayes Estimators

The Bayesian approach to statistics is fundamentally different from the classical approach that we have been taking. Nevertheless, some aspects of the Bayesian approach can be quite helpful to other statistical approaches. Before going into the methods for finding Bayes estimators, we first discuss the Bayesian approach to statistics.

In the classical approach the parameter,  $\theta$ , is thought to be an unknown, but fixed, quantity. A random sample  $X_1, \dots, X_n$  is drawn from a population indexed by  $\theta$  and, based on the observed values in the sample, knowledge about the value of  $\theta$  is obtained. In the Bayesian approach  $\theta$  is considered to be a quantity whose variation can be described by a probability distribution (called the *prior distribution*). This is a subjective distribution, based on the experimenter's belief, and is formulated before the data are seen (hence the name prior distribution). A sample is then taken from a population indexed by  $\theta$  and the prior distribution is updated with this sample information. The updated prior is called the *posterior distribution*. This updating

is done with the use of Bayes' Rule (seen in Chapter 1), hence the name Bayesian statistics.

If we denote the prior distribution by  $\pi(\theta)$ , and the sampling distribution by  $f(x|\theta)$ , then the posterior distribution, the conditional distribution of  $\theta$  given the sample,  $x$ , is

$$(7.2.6) \quad \pi(\theta|x) = f(x|\theta)\pi(\theta)/m(x), \quad (f(x|\theta)\pi(\theta) = f(x,\theta))$$

where  $m(x)$  is the marginal distribution of  $x$ , that is,

$$(7.2.7) \quad m(x) = \int f(x|\theta)\pi(\theta)d\theta.$$

Notice that the posterior distribution is a conditional distribution, conditional upon observing the sample. The posterior distribution is now used to make statements about  $\theta$ , which is still considered a random quantity. For instance, the mean of the posterior distribution can be used as a point estimate of  $\theta$ .

*A note on notation:* When dealing with distributions on a parameter,  $\theta$ , we will break our notation convention of using uppercase letters for random variables and lowercase letters for arguments. Thus, we may speak of the random quantity  $\theta$  with distribution  $\pi(\theta)$ . This is more in line with common usage and should not cause confusion.

**Example 7.2.9:** Let  $X_1, \dots, X_n$  be iid Bernoulli( $p$ ). Then  $Y = \sum X_i$  is binomial( $n, p$ ). We assume the prior distribution on  $p$  is beta( $\alpha, \beta$ ). The joint distribution of  $Y$  and  $p$  is

$$\begin{aligned} f(y,p) &= \left[ \binom{n}{y} p^y (1-p)^{n-y} \right] \left[ \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \right] \left( \begin{array}{c} \text{conditional} \times \text{marginal} \\ f(y|p) \times \pi(p) \end{array} \right) \\ &= \binom{n}{y} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1} (1-p)^{n-y+\beta-1}. \end{aligned}$$

The marginal pdf of  $Y$  is

$$(7.2.8) \quad f(y) = \int_0^1 f(y,p) dp = \binom{n}{y} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y + \alpha)\Gamma(n - y + \beta)}{\Gamma(n + \alpha + \beta)},$$

a distribution known as the beta-binomial (see Exercise 4.36 and Example 4.4.3). The posterior distribution, the distribution of  $p$  given  $y$ , is

$$f(p|y) = \frac{f(y,p)}{f(y)} = \frac{\Gamma(n + \alpha + \beta)}{\Gamma(y + \alpha)\Gamma(n - y + \beta)} p^{y+\alpha-1} (1-p)^{n-y+\beta-1},$$

which is beta( $y + \alpha, n - y + \beta$ ). (Remember that  $p$  is the variable,  $y$  is treated as fixed.) A natural estimate for  $p$  is the mean of the posterior distribution, which would give us as the Bayes estimator of  $p$ ,

$$\hat{p}_B = \frac{y + \alpha}{\alpha + \beta + n}.$$

||

Consider how the Bayes estimate of  $p$  is formed. The prior distribution has mean  $\alpha/(\alpha + \beta)$ , which would be our best estimate of  $p$  without having seen the data. Ignoring the prior information, we would probably use  $p = y/n$  as our estimate of  $p$ . The Bayes estimate of  $p$  combines all of this information. The manner in which this information is combined is made clear if we write  $\hat{p}_B$  as

$$\hat{p}_B = \left( \frac{n}{\alpha + \beta + n} \right) \left( \frac{y}{n} \right) + \left( \frac{\alpha + \beta}{\alpha + \beta + n} \right) \left( \frac{\alpha}{\alpha + \beta} \right).$$

Thus  $p_B$  is a linear combination of the prior mean and the sample mean, with the weights being determined by  $\alpha$ ,  $\beta$ , and  $n$ .

When estimating a binomial parameter, it is not necessary to choose a prior distribution from the beta family. However, there was a certain advantage to choosing the beta family, not the least of which being that we obtained a closed-form expression for the estimator. In general, for any sampling distribution, there is a natural family of prior distributions, called the conjugate family.

**DEFINITION 7.2.2:** Let  $\mathcal{F}$  denote the class of pdfs or pmfs  $f(x|\theta)$  (indexed by  $\theta$ ). A class  $\prod$  of prior distributions is a *conjugate family* for  $\mathcal{F}$  if the posterior distribution is in the class  $\prod$  for all  $f \in \mathcal{F}$ , all priors in  $\prod$ , and all  $x \in \mathcal{X}$ .

The beta family is conjugate for the binomial family. Thus, if we start with a beta prior, we will end up with a beta posterior. The updating of the prior takes the form of updating its parameters. Mathematically, this is very convenient, for it usually makes calculation quite easy. Whether or not a conjugate family is a reasonable choice for a particular problem, however, is a question to be left to the experimenter.

We end this section with one more example.

**Example 7.2.10:** Let  $X \sim n(\theta, \sigma^2)$ , and suppose that the prior distribution on  $\theta$  is  $n(\mu, \tau^2)$ . (Here we assume that  $\sigma^2$ ,  $\mu$ , and  $\tau^2$  are all known.) The posterior distribution of  $\theta$  is also normal, with mean and variance given by

$$\begin{aligned} E(\theta|x) &= \frac{\tau^2}{\tau^2 + \sigma^2}x + \frac{\sigma^2}{\tau^2 + \sigma^2}\mu, \\ \text{Var}(\theta|x) &= \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}. \end{aligned}$$

(See Exercise 7.23 for details.) Notice that the normal family is its own conjugate family. Again using the posterior mean we have the Bayes estimator of  $\theta$  is  $E(\theta|X)$ .

The Bayes estimator is, again, a linear combination of the prior and sample means. Notice also that as  $\tau^2$ , the prior variance, is allowed to tend to infinity, the Bayes estimator tends toward the sample mean. We can interpret this as saying that, as the prior information becomes more vague, the Bayes estimator tends to give more weight to the sample information. On the other hand, if the prior information is good, so that  $\sigma^2 > \tau^2$ , then more weight is given to the prior mean. ||

### 7.2.4 Invariant Estimators

The concept of invariance has already been discussed in Chapter 6, where it was seen that invariance principles are data reduction methods which can simplify problems. Furthermore, in Section 6.3, we saw invariance principles applied to point estimation. In this section we will go into a little more detail, and show how invariance can help in finding point estimators. A more complete treatment of invariant point estimators is given in Section 10.6.

While invariance principles help in finding estimators, we will not, in general, be able to use the invariance reduction to arrive at one estimator (as we did, for example, with maximum likelihood) but rather invariance will lead us to a class of invariant estimators. This class will be smaller than the class of all estimators, but we will still need to invoke some criterion for choosing one estimator out of the class.

We begin with a definition of an invariant estimator. Recall Definition 6.3.2, where the definition of invariance of a family of distributions was given. We defined a family of distributions  $\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$  to be invariant under the group  $\mathcal{G}$  if for every  $\theta \in \Theta$  and  $g \in \mathcal{G}$  there is a unique  $\theta' \in \Theta$  such that  $X \sim f(x|\theta) \Rightarrow Y = g(X) \sim f(y|\theta')$ . For a fixed  $g \in \mathcal{G}$ , the correspondence that takes  $\theta \rightarrow \theta'$  defines a function, which we denote by  $\bar{g}(\theta)$ , that is,  $\bar{g}(\theta) = \theta'$ . (Section 10.6 extends these notions further.) With this notation, we are ready for the definition.

**DEFINITION 7.2.3:** Let  $\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$  be invariant under the group  $\mathcal{G}$ . A point estimator  $W(x)$  of  $\theta$  is *invariant under the group  $\mathcal{G}$*  if, for every  $x \in \mathcal{X}$ ,  $\theta \in \Theta$ , and  $g \in \mathcal{G}$ ,  $W(g(x)) = \bar{g}(W(x))$ .

This definition is actually not very mysterious, as it follows quite naturally from the requirements of Measurement Invariance and Formal Invariance of Section 6.3. Using these principles, we have

*Measurement Invariance:*  $W(x)$  estimates  $\theta \Rightarrow \bar{g}(W(x))$  estimates  $\bar{g}(\theta) = \theta'$ .  
*Formal Invariance:*  $W(x)$  estimates  $\theta \Rightarrow W(g(x))$  estimates  $\bar{g}(\theta) = \theta'$ .

Putting these two requirements together gives  $W(g(x)) = \bar{g}(W(x))$ .

We now look at some examples of finding point estimators using invariance.

**Example 7.2.11:** Let  $X_1, \dots, X_n$  be iid  $f(x - \theta)$  and consider the group of transformations defined by  $\mathcal{G} = \{g_a(\mathbf{x}) : -\infty < a < \infty\}$ , where  $g_a(x_1, \dots, x_n) = (x_1 + a, \dots, x_n + a)$ . For this group we have  $\bar{g}_a(\theta) = \theta + a$ , since  $\mathbf{X} \sim f(\mathbf{x}|\theta) \Rightarrow \mathbf{Y} = g(\mathbf{X}) \sim f(\mathbf{y}|\theta + a)$ . Thus, an estimator is invariant with respect to this group if

$$W(g_a(\mathbf{x})) = \bar{g}_a(W(\mathbf{x})),$$

that is,

$$\begin{aligned} W(g_a(x_1, \dots, x_n)) &= W(x_1 + a, \dots, x_n + a) \\ (7.2.9) \quad &= \bar{g}_a(W(x_1, \dots, x_n)) \\ &= W(x_1, \dots, x_n) + a, \end{aligned}$$

which we have seen already in (6.3.4). Without further restrictions we cannot reduce the class further. Suppose that we reduce the class further, however, by requiring both invariance and unbiasedness. (Unbiasedness, which would require that  $E_\theta W = \theta$ , has been seen in Theorem 5.2.2, and is discussed in detail in Section 7.3.) We have

$$\begin{aligned} E_\theta(W(X_1, \dots, X_n)) &= E_\theta(W(X_1 + a, \dots, X_n + a)) - a \quad (\text{invariance}) \\ &= E_\theta(W(X_1 - \theta, \dots, X_n - \theta)) + \theta. \quad (\text{take } a = -\theta) \end{aligned}$$

Thus,  $W$  is unbiased if  $E_\theta[W(X_1 - \theta, \dots, X_n - \theta)] = 0$ . Making the transformation  $t_i = x_i - \theta$ , we have that

$$E_\theta[W(X_1 - \theta, \dots, X_n - \theta)] = E[W(T_1, \dots, T_n)] \quad (\text{independent of } \theta)$$

which must be equal to zero for  $W$  to be unbiased. Thus, we have further reduced the class. For example, note that as long as  $E_0 X_1 = 0$ ,  $W(X_1, \dots, X_n) = \bar{X}$  is unbiased and invariant (see Exercise 7.28). ||

In the above example the function  $\bar{g}$  involved the parameter of interest, and dictated how the estimator should change. In some instances the function  $\bar{g}$  does not involve the parameter of interest, and the defining equation leaves the estimator unchanged. This can occur, for example, when there are *nuisance parameters*, parameters that are not of direct concern.

**Example 7.2.12:** Let  $X_1, \dots, X_n$  be iid  $\frac{1}{\sigma}f((x-\theta)/\sigma)$  and again consider the group of transformations defined by  $\mathcal{G} = \{g_a(\mathbf{x}) : -\infty < a < \infty\}$ , where  $g_a(x_1, \dots, x_n) = (x_1 + a, \dots, x_n + a)$ . Now, however, suppose we are interested in estimating  $\sigma^2$  using an invariant estimator. Since the group does not affect  $\sigma^2$ , we have  $\bar{g}_a(\sigma^2) = \sigma^2$ , that is,  $\bar{g}$  is the identity. Thus, an estimator of  $\sigma^2$  is invariant with respect to this group if

$$W(g_a(\mathbf{x})) = \bar{g}_a(W(\mathbf{x})) = W(\mathbf{x}),$$

that is,

$$W(x_1, \dots, x_n) = W(x_1 + a, \dots, x_n + a),$$

which has been seen in (6.3.5). Note that this requirement is satisfied by any estimator that is a function of  $x_1, \dots, x_n$  only through the  $n - 1$  differences  $x_1 - x_n, \dots, x_{n-1} - x_n$ . ||

Thus, depending on how the function  $\bar{g}$  acts on the parameter of interest, the defining equation for an invariant estimator can either prescribe a certain kind of movement, or prescribe no movement at all.

Different groups lead to different amounts of reduction and, loosely speaking, the bigger the group the smaller the class of invariant estimators.

**Example 7.2.13:** Once again consider the location-scale case, and let  $X_1, \dots, X_n$  be iid  $\frac{1}{\sigma}f((x - \theta)/\sigma)$ . First, suppose that we want to estimate  $\theta$ , and we use the group of transformations defined by  $\mathcal{G} = \{g_{a,c}(x) : -\infty < a < \infty, c > 0\}$ , where  $g_{a,c}(x_1, \dots, x_n) = (cx_1 + a, \dots, cx_n + a)$ . Invariant estimators must satisfy

$$(7.2.10) \quad cW(x_1, \dots, x_n) + a = W(cx_1 + a, \dots, cx_n + a),$$

a more stringent requirement than in Example 7.2.11, where we considered  $g_a(x_1, \dots, x_n) = (x_1 + a, \dots, x_n + a)$ . Thus, we achieve a greater reduction here, and the class of invariant estimators is smaller. In particular, estimators of the form

$$W(x_1, \dots, x_n) = \bar{x} + k,$$

where  $k$  is a nonzero constant, are invariant with respect to the group  $g_a(x)$ , but are not invariant with respect to the group  $g_{a,c}(x)$ . (See Exercise 7.29.)

Similar reductions can be done for invariant estimators of  $\sigma^2$ , some of which we have already seen in Example 7.2.12. For example, estimators of the form  $kS^2$ , where  $k$  is a positive constant and  $S^2$  is the sample variance, are invariant with respect to  $g_{a,c}(x)$ . However, if we consider the *scale group*

$$\mathcal{G}_c = \{g_c(x) : g_c(x) = cx, c > 0\}$$

then we get less of a reduction and, hence, a larger class of estimators. In particular, estimators of the form

$$W(X_1, \dots, X_n) = \phi\left(\frac{\bar{X}}{S}\right)S^2,$$

where  $\phi(x)$  is a function, are invariant with respect to  $g_c$ , but not with respect to either  $g_a$  or  $g_{a,c}$ , unless  $\phi(x)$  is a constant (see Exercise 7.30). Consideration of estimators of this form led Stein (1964) and Brewster and Zidek (1974) to find improved estimators of variance. ||

## 7.3 Methods of Evaluating Estimators

The methods discussed in the previous section have outlined reasonable techniques for finding point estimators of parameters. A difficulty that arises, however, is that since we can usually apply more than one of these methods in a particular situation, we are often faced with the task of choosing between estimators. Of course, it is possible that different methods of finding estimators will yield the same answer, which makes evaluation a bit easier, but, in many cases, different methods will lead to different estimators.

The general topic of evaluating statistical procedures is part of the branch of statistics known as decision theory, which will be treated in some detail in Chapter 10. However, no procedure should be considered until some clues about its performance have been gathered. In this section we will introduce some basic criteria for evaluating estimators, and examine several estimators against these criteria.

### 7.3.1 Mean Squared Error

We first investigate finite-sample measures of the quality of an estimator, beginning with its mean squared error.

**DEFINITION 7.3.1:** The *mean squared error* (MSE) of an estimator  $W$  of a parameter  $\theta$  is the function of  $\theta$  defined by  $E_\theta(W - \theta)^2$ .

Notice that the MSE measures the average squared difference between the estimator  $W$  and the parameter  $\theta$ , a somewhat reasonable measure of performance for a point estimator. In general, any increasing function of the absolute distance  $|W - \theta|$  would serve to measure the goodness of an estimator (mean absolute error,  $E_\theta(|W - \theta|)$  is a reasonable alternative), but MSE has at least two advantages over other distance measures: First, it is quite tractable analytically and, second, it has the interpretation

$$(7.3.1) \quad E_\theta(W - \theta)^2 = \text{Var}_\theta W + (E_\theta W - \theta)^2 = \text{Var}_\theta W + (\text{Bias}_\theta W)^2,$$

where we define the bias of an estimator as follows.

**DEFINITION 7.3.2:** The *bias* of a point estimator  $W$ , of a parameter  $\theta$ , is the difference between the expected value of  $W$  and  $\theta$ . That is,  $\text{Bias}_\theta W = E_\theta W - \theta$ . An estimator whose bias is identically (in  $\theta$ ) equal to zero is called *unbiased* and satisfies  $E_\theta W = \theta$  for all  $\theta$ .

Thus, MSE incorporates two components, one measuring the variability of the estimator (precision), and the other measuring its bias (accuracy). An estimator that has good MSE properties has small combined variance and bias. To find an estimator with good MSE properties, we need to find estimators that control both variance and bias. Clearly, unbiased estimators do a good job of controlling bias.

For an unbiased estimator we have

$$\mathbb{E}_\theta(W - \theta)^2 = \text{Var}_\theta W,$$

and so, if an estimator is unbiased, its MSE is equal to its variance.

**Example 7.3.1:** Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$ . The statistics  $\bar{X}$  and  $S^2$  are both unbiased estimators since

$$\mathbb{E}\bar{X} = \mu, \quad \mathbb{E}S^2 = \sigma^2, \quad \text{for all } \mu \text{ and } \sigma^2.$$

(This is true without the normality assumption; see Theorem 5.2.2.) The MSEs of these estimators are given by

$$\begin{aligned} \mathbb{E}(\bar{X} - \mu)^2 &= \text{Var } \bar{X} = \frac{\sigma^2}{n}, \\ \mathbb{E}(S^2 - \sigma^2)^2 &= \text{Var } S^2 = \frac{2\sigma^4}{n-1}. \end{aligned}$$

Although many unbiased estimators are also reasonable from the standpoint of MSE, be aware that controlling bias does not guarantee that MSE is controlled. In particular, it is sometimes the case that a trade-off occurs between variance and bias in such a way that a small increase in bias can be traded for a larger decrease in variance, resulting in an improvement in MSE.

**Example 7.3.1 (Continued):** An alternative estimator for  $\sigma^2$  is the maximum likelihood estimator  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S^2$ . It is straightforward to calculate

$$\mathbb{E}\hat{\sigma}^2 = \mathbb{E}\left(\frac{n-1}{n} S^2\right) = \frac{n-1}{n} \sigma^2,$$

so  $\hat{\sigma}^2$  is a biased estimator of  $\sigma^2$ . The variance of  $\hat{\sigma}^2$  can also be calculated as

$$\text{Var } \hat{\sigma}^2 = \text{Var} \left( \frac{n-1}{n} S^2 \right) = \left( \frac{n-1}{n} \right)^2 \text{Var } S^2 = \frac{2(n-1)\sigma^4}{n^2},$$

and, hence, its MSE is given by

$$\mathbb{E}(\hat{\sigma}^2 - \sigma^2)^2 = \frac{2(n-1)\sigma^4}{n^2} + \left( \frac{n-1}{n} \sigma^2 - \sigma^2 \right)^2 = \left( \frac{2n-1}{n^2} \right) \sigma^4.$$

We thus have

$$\mathbb{E}(\hat{\sigma}^2 - \sigma^2)^2 = \left( \frac{2n-1}{n^2} \right) \sigma^4 < \left( \frac{2}{n-1} \right) \sigma^4 = \mathbb{E}(S^2 - \sigma^2)^2,$$

showing that  $\hat{\sigma}^2$  has smaller MSE than  $S^2$ . Thus, by trading off variance for bias, the MSE is improved. ||

We hasten to point out that the above example does not imply that  $S^2$  should be abandoned as an estimator of  $\sigma^2$ . The above argument shows that, on the average,  $\hat{\sigma}^2$  will be closer to  $\sigma^2$  than  $S^2$  if MSE is used as a measure. However,  $\hat{\sigma}^2$  is biased and will, on the average, underestimate  $\sigma^2$ . This fact alone may make us uncomfortable about using  $\hat{\sigma}^2$  as an estimator of  $\sigma^2$ . Furthermore, it can be argued that MSE, while a reasonable criterion for location parameters, is not reasonable for scale parameters, so the above comparison should not even be made. (One problem is that MSE penalizes equally for overestimation and underestimation, which is fine in the location case. In the scale case, however, zero is a natural lower bound, so the estimation problem is not symmetric. Use of MSE in this case tends to be forgiving of underestimation.) The end result of this is that no absolute answer is obtained but rather more information is gathered about the estimators in the hope that, for a particular situation, a good estimator is chosen.

In general, since MSE is a function of the parameter, there will not be one "best" estimator. Often, the MSEs of two estimators will cross each other, showing that each estimator is better (with respect to the other) only in a portion of the parameter space. However, even this partial information can sometimes provide guidelines for choosing between estimators.

**Example 7.3.2:** Let  $X_1, \dots, X_n$  be iid Bernoulli( $p$ ). The MSE of  $\bar{X}$ , as an estimator of  $p$ , is

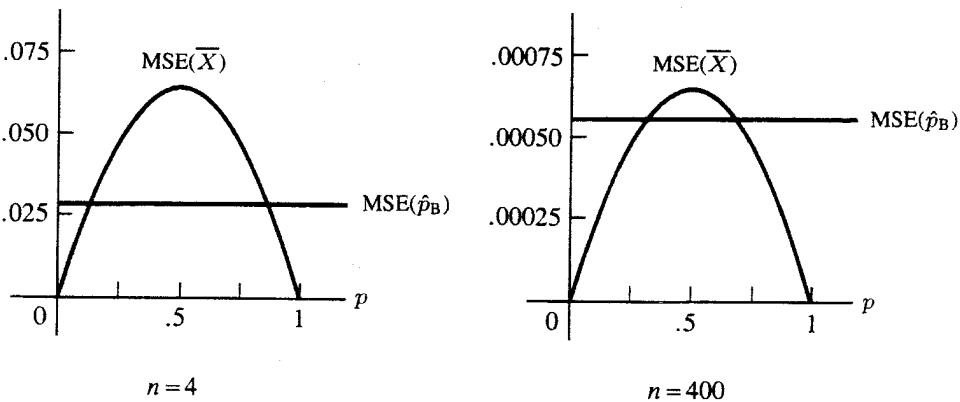
$$E_p(\bar{X} - p)^2 = \text{Var}_p \bar{X} = \frac{p(1-p)}{n}.$$

Let  $Y = \sum X_i$  and recall the Bayes estimator derived in Example 7.2.9,  $\hat{p}_B = \frac{Y+\alpha}{\alpha+\beta+n}$ . The MSE of this Bayes estimator of  $p$  is

$$\begin{aligned} E_p(\hat{p}_B - p)^2 &= \text{Var}_p \hat{p}_B + (\text{Bias}_p \hat{p}_B)^2 \\ &= \text{Var}_p \left( \frac{Y+\alpha}{\alpha+\beta+n} \right) + \left( E_p \left( \frac{Y+\alpha}{\alpha+\beta+n} \right) - p \right)^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left( \frac{np+\alpha}{\alpha+\beta+n} - p \right)^2. \end{aligned}$$

In the absence of good prior information about  $p$ , we might try to choose  $\alpha$  and  $\beta$  to make the MSE of  $\hat{p}_B$  constant. The details are not too difficult to work out (see Exercise 7.31), and the choice  $\alpha = \beta = \sqrt{n/4}$  yields

$$\hat{p}_B = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}} \quad \text{and} \quad E(\hat{p}_B - p)^2 = \frac{n}{4(n + \sqrt{n})^2}.$$



**FIGURE 7.3.1** Comparison of MSE of  $\bar{X}$  and  $\hat{p}_B$  for sample sizes  $n = 4$  and  $n = 400$  in Example 7.3.2.

If we want to choose between  $\hat{p}_B$  and  $\bar{X}$  on the basis of MSE, Figure 7.3.1 is helpful. For small  $n$ ,  $\hat{p}_B$  is the better choice (unless there is a strong belief that  $p$  is near 0 or 1). For large  $n$ ,  $\bar{X}$  is the better choice (unless there is a strong belief that  $p$  is close to  $\frac{1}{2}$ ). Even though the MSE criterion does not show one estimator to be uniformly better than the other, useful information is provided. This information, combined with the knowledge of the problem at hand, can lead to choosing the better estimator for the situation. ||

In certain situations, particularly in location parameter estimation, MSE can be a helpful criterion for finding the best estimator in a class of invariant estimators.

**Example 7.3.3:** Let  $X_1, \dots, X_n$  be iid  $f(x - \theta)$ . In Example 7.2.11 we saw that estimators  $W(X_1, \dots, X_n)$  that satisfy

$$(7.3.2) \quad W(x_1, \dots, x_n) + a = W(x_1 + a, \dots, x_n + a)$$

are invariant with respect to the group of transformations defined by  $\mathcal{G} = \{g_a(\mathbf{x}) : -\infty < a < \infty\}$ , where  $g_a(x_1, \dots, x_n) = (x_1 + a, \dots, x_n + a)$ . For these estimators we have

$$\begin{aligned}
 & E_\theta(W(X_1, \dots, X_n) - \theta)^2 \\
 &= E_\theta(W(X_1 + a, \dots, X_n + a) - a - \theta)^2 \\
 &= E_\theta(W(X_1 - \theta, \dots, X_n - \theta))^2 \quad (a = -\theta) \\
 &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (W(x_1 - \theta, \dots, x_n - \theta))^2 \prod_{i=1}^n f(x_i - \theta) dx_i \\
 (7.3.3) \quad &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (W(u_1, \dots, u_n))^2 \prod_{i=1}^n f(u_i) du_i. \quad (u_i = x_i - \theta)
 \end{aligned}$$

This last expression does not depend on  $\theta$ , hence the MSEs of these invariant estimators are not functions of  $\theta$ . The MSE can therefore be used to order the invariant

estimators, and an invariant estimator with smallest MSE can be found. In fact, this estimator is the solution to the mathematical problem of finding the function  $W$  that minimizes (7.3.3) subject to (7.3.2). (See Exercises 7.33 and 7.34.) ||

### 7.3.2 Best Unbiased Estimators

As noted in the previous section, a comparison of estimators based on MSE considerations may not yield a clear favorite. Indeed, there is no one “best MSE” estimator. Many find this troublesome or annoying, and rather than doing MSE comparisons of candidate estimators, they would rather have a “recommended” one.

The reason that there is no one “best MSE” estimator is that the class of all estimators is too large a class. (For example, the estimator  $\hat{\theta} = 17$  cannot be beaten in MSE at  $\theta = 17$ , but is a terrible estimator otherwise.) One way to make the problem of finding a “best” estimator tractable is to limit the class of estimators. A popular way of restricting the class of estimators, the one we consider in this section, is to consider only unbiased estimators.

If  $W_1$  and  $W_2$  are both unbiased estimators of a parameter  $\theta$ , that is,  $E_\theta W_1 = E_\theta W_2 = \theta$ , then their mean squared errors are equal to their variances, so we should choose the estimator with the smaller variance. If we can find an unbiased estimator with uniformly smallest variance—a best unbiased estimator—then our task is done.

Before proceeding we note that, although we will be dealing with unbiased estimators, the results here and in the next section are actually more general. Suppose that there is an estimator  $W^*$  of  $\theta$  with  $E_\theta W^* = \tau(\theta) \neq \theta$ , and we are interested in investigating the value of  $W^*$ . Consider the class of estimators

$$\mathcal{C}_\tau = \{W : E_\theta W = \tau(\theta)\}.$$

For any  $W_1, W_2 \in \mathcal{C}_\tau$ ,  $\text{Bias}_\theta W_1 = \text{Bias}_\theta W_2$ , so

$$E_\theta(W_1 - \theta)^2 - E_\theta(W_2 - \theta)^2 = \text{Var}_\theta W_1 - \text{Var}_\theta W_2,$$

and MSE comparisons, within the class  $\mathcal{C}_\tau$ , can be based on variance alone. Thus, although we speak in terms of unbiased estimators, we really are comparing estimators that have the same expected value,  $\tau(\theta)$ .

The goal of this section is to investigate a method for finding a “best” unbiased estimator, which we define in the following way.

**DEFINITION 7.3.3:** An estimator  $W^*$  is a *best unbiased estimator* of  $\tau(\theta)$  if it satisfies  $E_\theta W^* = \tau(\theta)$  for all  $\theta$  and, for any other estimator  $W$  with  $E_\theta W = \tau(\theta)$ , we have  $\text{Var}_\theta W^* \leq \text{Var}_\theta W$  for all  $\theta$ .  $W^*$  is also called a *uniform minimum variance unbiased estimator* (UMVUE) of  $\tau(\theta)$ .

Finding a best unbiased estimator (if one exists!) is not an easy task for a variety of reasons, two of which are illustrated in the following example.

**Example 7.3.4:** Let  $X_1, \dots, X_n$  be iid Poisson( $\lambda$ ), and let  $\bar{X}$  and  $S^2$  be the sample mean and variance, respectively. Recall that for the Poisson pmf both the mean and variance are equal to  $\lambda$ . Therefore, applying Theorem 5.2.2, we have

$$E_\lambda \bar{X} = \lambda, \quad \text{for all } \lambda,$$

and

$$E_\lambda S^2 = \lambda, \quad \text{for all } \lambda,$$

so both  $\bar{X}$  and  $S^2$  are unbiased estimators of  $\lambda$ .

To determine the better estimator,  $\bar{X}$  or  $S^2$ , we should now compare variances. Again from Theorem 5.2.2, we have  $\text{Var}_\lambda \bar{X} = \lambda/n$ , but  $\text{Var}_\lambda S^2$  is quite a lengthy calculation (resembling that in Exercise 5.10b). This is one of the first problems in finding a best unbiased estimator. Not only may the calculations be long and involved, but they may be for nought (as in this case), for we will see that  $\text{Var}_\lambda \bar{X} \leq \text{Var}_\lambda S^2$  for all  $\lambda$ .

Even if we can establish that  $\bar{X}$  is better than  $S^2$ , consider the class of estimators

$$W_a(\bar{X}, S^2) = a\bar{X} + (1 - a)S^2.$$

For every constant  $a$ ,  $E_\lambda W_a(\bar{X}, S^2) = \lambda$ , so we now have infinitely many unbiased estimators of  $\lambda$ . Even if  $\bar{X}$  is better than  $S^2$ , is it better than every  $W_a(\bar{X}, S^2)$ ? Furthermore, how can we be sure that there are not other, better, unbiased estimators lurking about? ||

This example shows some of the problems that might be encountered in trying to find a best unbiased estimator, and perhaps that a more comprehensive approach is desirable. Suppose that, for estimating a parameter  $\tau(\theta)$  of a distribution  $f(x|\theta)$ , we can specify a lower bound, say  $B(\theta)$ , on the variance of *any* unbiased estimator of  $\tau(\theta)$ . If we can then find an unbiased estimator  $W^*$  satisfying  $\text{Var}_\theta W^* = B(\theta)$ , we have found a best unbiased estimator. This is the approach taken with the use of the Cramér–Rao Lower Bound.

**THEOREM 7.3.1 (Cramér–Rao):** Let  $X_1, \dots, X_n$  be a sample with pdf  $f(x|\theta)$ , and let  $W(\mathbf{X}) = W(X_1, \dots, X_n)$  be any estimator where  $E_\theta W(\mathbf{X})$  is a differentiable function of  $\theta$ . Suppose the joint pdf  $f(\mathbf{x}|\theta) = f(x_1, \dots, x_n|\theta)$  satisfies

$$(7.3.4) \quad \frac{d}{d\theta} \int \cdots \int h(\mathbf{x}) f(\mathbf{x}|\theta) dx_1 \cdots dx_n = \int \cdots \int h(\mathbf{x}) \frac{\partial}{\partial \theta} f(\mathbf{x}|\theta) dx_1 \cdots dx_n,$$

for any function  $h(\mathbf{x})$  with  $E_\theta |h(\mathbf{X})| < \infty$ . Then

$$(7.3.5) \quad \text{Var}_\theta (W(\mathbf{X})) \geq \frac{\left( \frac{d}{d\theta} E_\theta W(\mathbf{X}) \right)^2}{E_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right)^2 \right)}.$$

*Proof:* The proof of this theorem is elegantly simple, and is a clever application of the Cauchy–Schwarz Inequality, or stated statistically, the fact that for any two random variables  $X$  and  $Y$ ,

$$(7.3.6) \quad [\text{Cov}(X, Y)]^2 \leq (\text{Var } X)(\text{Var } Y).$$

If we rearrange (7.3.6) we can get a lower bound on the variance of  $X$ ,

$$\text{Var } X \geq \frac{[\text{Cov}(X, Y)]^2}{\text{Var } Y}.$$

The cleverness in this theorem follows from choosing  $X$  to be the estimator  $W(\mathbf{X})$  and  $Y$  to be the quantity  $\frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta)$ , and applying the Cauchy–Schwarz Inequality. To begin we need to calculate the covariance of  $W(\mathbf{X})$  and  $\frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta)$ , but first we calculate the expected value of  $\frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta)$ . We have

$$\begin{aligned}
 (7.3.7) \quad \mathbb{E}_\theta \left( \frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right) &= \mathbb{E}_\theta \left[ \frac{\frac{\partial}{\partial \theta} f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)} \right] && \text{(property of logs)} \\
 &= \int \cdots \int \frac{\frac{\partial}{\partial \theta} f(\mathbf{x}|\theta)}{f(\mathbf{x}|\theta)} f(\mathbf{x}|\theta) dx_1 \cdots dx_n && \text{(definition of expectation)} \\
 &= \int \cdots \int \frac{\partial}{\partial \theta} f(\mathbf{x}|\theta) dx_1 \cdots dx_n && \text{(cancel } f(\mathbf{x}|\theta)) \\
 &= \frac{d}{d\theta} \int \cdots \int f(\mathbf{x}|\theta) dx_1 \cdots dx_n && \text{(apply (7.3.4) with } h(\mathbf{x}) = 1) \\
 &= \frac{d}{d\theta} \mathbb{E}_\theta(1) \\
 &= 0.
 \end{aligned}$$

Therefore  $\text{Cov}_\theta(W(\mathbf{X}), \frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta))$  is equal to the expectation of the product, and continuing our evaluation we obtain

$$\begin{aligned}
 (7.3.8) \quad \text{Cov}_\theta(W(\mathbf{X}), \frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta)) &= \mathbb{E}_\theta \left( W(\mathbf{X}) \frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right) \\
 &= \mathbb{E}_\theta \left[ \frac{W(\mathbf{X}) (\frac{\partial}{\partial \theta} f(\mathbf{X}|\theta))}{f(\mathbf{X}|\theta)} \right] && \text{(property of logs)} \\
 &= \int \cdots \int W(\mathbf{x}) \frac{\frac{\partial}{\partial \theta} f(\mathbf{x}|\theta)}{f(\mathbf{x}|\theta)} f(\mathbf{x}|\theta) dx_1 \cdots dx_n && \text{(definition of expectation)}
 \end{aligned}$$

$$\begin{aligned}
&= \int \cdots \int W(\mathbf{x}) \frac{\partial}{\partial \theta} f(\mathbf{x}|\theta) dx_1 \cdots dx_n && \text{(cancel } f(\mathbf{x}|\theta)\text{)} \\
&= \frac{d}{d\theta} \int \cdots \int W(\mathbf{x}) f(\mathbf{x}|\theta) dx_1 \cdots dx_n && \left( \begin{array}{l} \text{apply (7.3.4)} \\ \text{with } h(\mathbf{x}) = W(\mathbf{x}) \end{array} \right) \\
&= \frac{d}{d\theta} E_\theta W(\mathbf{X}).
\end{aligned}$$

Also, since  $E_\theta(\frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta)) = 0$  it follows that

$$(7.3.9) \quad \text{Var}_\theta \left( \frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right) = E_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right)^2 \right).$$

Using the Cauchy–Schwarz Inequality, together with (7.3.8) and (7.3.9), we obtain

$$\text{Var}_\theta(W(\mathbf{X})) \geq \frac{\left( \frac{d}{d\theta} E_\theta W(\mathbf{X}) \right)^2}{E_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right)^2 \right)},$$

proving the theorem.  $\square$

If we add the assumption of independent samples then the calculation of the lower bound is simplified. The expectation in the denominator becomes a univariate calculation, as the following corollary shows.

**COROLLARY 7.3.1** (Cramér–Rao, iid case): Let  $X_1, \dots, X_n$  be iid with pdf  $f(x|\theta)$ , and let  $W(\mathbf{X}) = W(X_1, \dots, X_n)$  be any estimator where  $E_\theta W(\mathbf{X})$  is a differentiable function of  $\theta$ . If the joint pdf  $f(\mathbf{x}|\theta) = \prod f(x_i|\theta)$  satisfies (7.3.4), then

$$\text{Var}_\theta W(\mathbf{X}) \geq \frac{\left( \frac{d}{d\theta} E_\theta W(\mathbf{X}) \right)^2}{n E_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right)^2 \right)}.$$

*Proof:* We only need to show that

$$E_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right)^2 \right) = n E_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right).$$

Since  $X_1, \dots, X_n$  are independent

$$\begin{aligned}
E_\theta \left( \frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right)^2 &= E_\theta \left( \left( \frac{\partial}{\partial \theta} \log \prod_{i=1}^n f(X_i|\theta) \right)^2 \right) \\
&= E_\theta \left( \left( \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(X_i|\theta) \right)^2 \right) && \text{(property of logs)}
\end{aligned}$$

(7.3.10)

$$\begin{aligned}
 &= \sum_{i=1}^n E_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(X_i | \theta) \right)^2 \right) \quad (\text{expand the square}) \\
 &\quad + \sum_{i \neq j} E_\theta \left( \frac{\partial}{\partial \theta} \log f(X_i | \theta) \frac{\partial}{\partial \theta} \log f(X_j | \theta) \right).
 \end{aligned}$$

For  $i \neq j$  we have

$$\begin{aligned}
 &E_\theta \left( \frac{\partial}{\partial \theta} \log f(X_i | \theta) \frac{\partial}{\partial \theta} \log f(X_j | \theta) \right) \\
 &= E_\theta \left( \frac{\partial}{\partial \theta} \log f(X_i | \theta) \right) E_\theta \left( \frac{\partial}{\partial \theta} \log f(X_j | \theta) \right) \quad (\text{independence}) \\
 &= 0. \quad (\text{from 7.3.7})
 \end{aligned}$$

Therefore the second sum in (7.3.10) is zero, and the first term is

$$\sum_{i=1}^n E_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(X_i | \theta) \right)^2 \right) = n E_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(X | \theta) \right)^2 \right), \quad (\text{identical distributions})$$

which establishes the corollary.  $\square$

Before going on we note that although the Cramér–Rao Lower Bound is stated for continuous random variables, it also applies to discrete random variables. The key condition, (7.3.4), which allows interchange of integration and differentiation, undergoes the obvious modification. If  $f(x|\theta)$  is a pmf then we must be able to interchange differentiation and summation. (Of course, this assumes that even though  $f(x|\theta)$  is a pmf and *not* differentiable in  $x$ , it *is* differentiable in  $\theta$ . This is the case for most common pmfs.)

The quantity  $E_\theta \left( \left( \frac{\partial}{\partial \theta} \log \prod_{i=1}^n f(X_i | \theta) \right)^2 \right)$  is called the *information number*, or *Fisher information* of the sample. This terminology reflects the fact that the information number gives a bound on the variance of the best unbiased estimator of  $\theta$ . As the information number gets bigger and we have more information about  $\theta$ , we have a smaller bound on the variance of the best unbiased estimator.

For any differentiable function  $\tau(\theta)$  we now have a lower bound on the variance of any estimator  $W$  satisfying  $E_\theta W = \tau(\theta)$ . The bound depends only on  $\tau(\theta)$  and  $f(x|\theta)$ , and is independent of the estimator under consideration. Hence, it is a uniform lower bound on the variance, and any estimator satisfying  $E_\theta W = \tau(\theta)$  and attaining this lower bound is a best unbiased estimator of  $\tau(\theta)$ .

Before looking at some examples, we present a computational result that aids in the application of this theorem.

LEMMA 7.3.1: If  $f(x|\theta)$  satisfies

$$\frac{d}{d\theta} E_\theta \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right) = \int \frac{\partial}{\partial \theta} \left[ \left( \frac{\partial}{\partial \theta} \log f(x|\theta) \right) f(x|\theta) \right] dx$$

(true for an exponential family), then

$$E_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right) = -E_\theta \left( \frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right).$$

*Proof:* Exercise 7.37. □

Using the tools just developed, we return to, and settle, the Poisson example.

**Example 7.3.4 (Continued):** Here  $\tau(\lambda) = \lambda$ , so  $\tau'(\lambda) = 1$ . Also, since we have an exponential family, using Lemma 7.3.1 gives us

$$\begin{aligned} E_\lambda \left( \left( \frac{\partial}{\partial \lambda} \log \prod_{i=1}^n f(X_i|\lambda) \right)^2 \right) &= -n E_\lambda \left( \frac{\partial^2}{\partial \lambda^2} \log f(X|\lambda) \right) \\ &= -n E_\lambda \left( \frac{\partial^2}{\partial \lambda^2} \log \left( \frac{e^{-\lambda} \lambda^X}{X!} \right) \right) \\ &= -n E_\lambda \left( \frac{\partial^2}{\partial \lambda^2} (-\lambda + X \log \lambda - \log X!) \right) \\ &= -n E_\lambda \left( -\frac{X}{\lambda^2} \right) \\ &= \frac{n}{\lambda}. \end{aligned}$$

Hence for any unbiased estimator,  $W$ , of  $\lambda$ , we must have

$$\text{Var}_\lambda W \geq \frac{\lambda}{n}.$$

Since  $\text{Var}_\lambda \bar{X} = \lambda/n$ ,  $\bar{X}$  is a best unbiased estimator of  $\lambda$ . ||

It is important to remember that a key assumption in the Cramér–Rao Theorem is the ability to differentiate under the integral sign which, of course, is somewhat restrictive. As we have seen, densities in the exponential class will satisfy the assumptions but, in general, such assumptions need to be checked, or contradictions such as the following will arise.

**Example 7.3.5:** Let  $X_1, \dots, X_n$  be iid with pdf  $f(x|\theta) = 1/\theta, 0 < x < \theta$ . Since  $\frac{\partial}{\partial \theta} \log f(x|\theta) = -1/\theta$ , we have

$$\mathbb{E}_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right) = \frac{1}{\theta^2}.$$

The Cramér–Rao Theorem would seem to indicate that if  $W$  is any unbiased estimator of  $\theta$ ,

$$\text{Var}_\theta W \geq \frac{\theta^2}{n}.$$

We would now like to find an unbiased estimator with small variance. As a first guess, consider the sufficient statistic  $Y = \max(X_1, \dots, X_n)$ , the largest order statistic. The pdf of  $Y$  is  $f_Y(y|\theta) = ny^{n-1}/\theta^n$ ,  $0 < y < \theta$ , so

$$\mathbb{E}_\theta Y = \int_0^\theta \frac{ny^n}{\theta^n} dy = \frac{n}{n+1}\theta,$$

showing that  $\frac{n+1}{n}Y$  is an unbiased estimator of  $\theta$ . We next calculate

$$\begin{aligned} \text{Var}_\theta \left( \frac{n+1}{n}Y \right) &= \left( \frac{n+1}{n} \right)^2 \text{Var}_\theta Y \\ &= \left( \frac{n+1}{n} \right)^2 \left[ \mathbb{E}_\theta Y^2 - \left( \frac{n}{n+1}\theta \right)^2 \right] \\ &= \left( \frac{n+1}{n} \right)^2 \left[ \frac{n}{n+2}\theta^2 - \left( \frac{n}{n+1}\theta \right)^2 \right] \\ &= \frac{1}{n(n+2)}\theta^2, \end{aligned}$$

which is uniformly smaller than  $\theta^2/n$ . This indicates that the Cramér–Rao Theorem is not applicable to this pdf. To see that this is so, we can use Leibnitz's Rule (Section 2.4) to calculate

$$\begin{aligned} \frac{d}{d\theta} \int_0^\theta h(x)f(x|\theta) dx &= \frac{d}{d\theta} \int_0^\theta h(x)\frac{1}{\theta} dx \\ &= \frac{h(\theta)}{\theta} + \int_0^\theta h(x)\frac{\partial}{\partial\theta} \left( \frac{1}{\theta} \right) dx \\ &\neq \int_0^\theta h(x)\frac{\partial}{\partial\theta} f(x|\theta) dx, \end{aligned}$$

unless  $h(\theta)/\theta = 0$  for all  $\theta$ . Hence, the Cramér–Rao Theorem does not apply. In

general, if the range of the pdf depends on the parameter, the theorem will not be applicable.

A shortcoming of this approach to finding best unbiased estimators is that, even if the Cramér–Rao Theorem is applicable, there is no guarantee that the bound is sharp. That is to say, the value of the Cramér–Rao Lower Bound may be *strictly smaller* than the variance of *any* unbiased estimator. In fact, in the usually favorable case of  $f(x|\theta)$  being a one-parameter exponential family, the most that we can say is that there exists a parameter  $\tau(\theta)$  with an unbiased estimator that achieves the Cramér–Rao Lower Bound. However, in other typical situations, the bound may not be attainable. These situations cause concern because, if we cannot find an estimator that attains the lower bound, we have to decide whether no estimator can attain it or if we must look at more estimators.

**Example 7.3.6:** Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$ , and consider estimation of  $\sigma^2$ , where  $\mu$  is unknown. The normal pdf satisfies the assumptions of the Cramér–Rao Theorem and Lemma 7.3.1, so we have

$$\frac{\partial^2}{\partial(\sigma^2)^2} \log \left( \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(1/2)(x-\mu)^2/\sigma^2} \right) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}$$

and

$$\begin{aligned} -E \left( \frac{\partial^2}{\partial(\sigma^2)^2} \log f(X|\mu, \sigma^2) \middle| \mu, \sigma^2 \right) &= -E \left( \frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6} \middle| \mu, \sigma^2 \right) \\ &= \frac{1}{2\sigma^4}. \end{aligned}$$

Thus, any unbiased estimator,  $W$ , of  $\sigma^2$  must satisfy

$$\text{Var}(W|\mu, \sigma^2) \geq \frac{2\sigma^4}{n}.$$

In Example 7.3.1 we saw

$$\text{Var}(S^2|\mu, \sigma^2) = \frac{2\sigma^4}{n-1},$$

so  $S^2$  does not attain the Cramér–Rao Lower Bound.

In the above example we are left with an incomplete answer; that is, is there a better unbiased estimator of  $\sigma^2$  than  $S^2$ , or is the Cramér–Rao Lower Bound unattainable?

The conditions for attainment of the Cramér–Rao Lower Bound are actually quite simple. Recall that the bound follows from an application of the Cauchy–Schwarz Inequality, so conditions for attainment of the bound are the conditions for equality in

the Cauchy–Schwarz Inequality (see Section 4.7.1). Note also that Corollary 7.3.2 is a useful tool because it implicitly gives us a way of finding a best unbiased estimator.

**COROLLARY 7.3.2:** Let  $X_1, \dots, X_n$  be iid  $f(x|\theta)$ , where  $f(x|\theta)$  satisfies the conditions of the Cramér–Rao Theorem. Let  $L(\theta|x) = \prod_{i=1}^n f(x_i|\theta)$  denote the likelihood function. If  $W(\mathbf{X}) = W(X_1, \dots, X_n)$  is any unbiased estimator of  $\tau(\theta)$ , then  $W(\mathbf{X})$  attains the Cramér–Rao Lower Bound if and only if

$$(7.3.11) \quad a(\theta)[W(\mathbf{x}) - \tau(\theta)] = \frac{\partial}{\partial \theta} \log L(\theta|\mathbf{x}),$$

for some function  $a(\theta)$ .

*Proof:* The Cramér–Rao Inequality, as given in (7.3.6), can be written as

$$\left[ \text{Cov}_\theta \left( W(\mathbf{X}), \frac{\partial}{\partial \theta} \log \prod_{i=1}^n f(X_i|\theta) \right) \right]^2 \leq \text{Var}_\theta W(\mathbf{X}) \text{Var}_\theta \left( \frac{\partial}{\partial \theta} \log \prod_{i=1}^n f(X_i|\theta) \right),$$

and, recalling that  $E_\theta W = \tau(\theta)$ ,  $E_\theta(\frac{\partial}{\partial \theta} \log \prod_{i=1}^n f(X_i|\theta)) = 0$ , and using the results of Theorem 4.5.4, we can have equality if and only if  $W(\mathbf{x}) - \tau(\theta)$  is proportional to  $\frac{\partial}{\partial \theta} \log \prod_{i=1}^n f(x_i|\theta)$ . That is exactly what is expressed in (7.3.11).  $\square$

**Example 7.3.6 (Continued):** Here we have

$$L(\mu, \sigma^2 | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_{i=1}^n (x_i - \mu)^2 / \sigma^2},$$

and hence

$$\frac{\partial}{\partial \sigma^2} \log L(\mu, \sigma^2 | \mathbf{x}) = \frac{n}{2\sigma^4} \left( \sum_{i=1}^n \frac{(x_i - \mu)^2}{n} - \sigma^2 \right).$$

Thus, taking  $a(\sigma^2) = n/(2\sigma^4)$  shows that the best unbiased estimator of  $\sigma^2$  is  $\sum_{i=1}^n (x_i - \mu)^2/n$ , which is calculable only if  $\mu$  is known. If  $\mu$  is unknown the bound *cannot be attained*.  $\parallel$

The theory developed in this section still leaves some questions unanswered. First, what can we do if  $f(x|\theta)$  does not satisfy the assumptions of the Cramér–Rao Theorem? (In Example 7.3.5, we still do not know if  $\frac{n+1}{n}Y$  is a best unbiased estimator.) Second, what if the bound is unattainable by allowable estimators, as in Example 7.3.6? There, we still do not know if  $S^2$  is a best unbiased estimator.

One way of answering these questions is to search for methods that are more widely applicable, and yield sharper (that is, greater) lower bounds. Much research has been done on this topic, with perhaps the most well known bound being that of Chapman and Robbins (1951). Kendall and Stuart (1979, Chapter 17) have a good

treatment of this subject. Rather than take this approach, however, we will continue the study of best unbiased estimators from another view, using the concept of sufficiency.

### 7.3.3 Sufficiency and Unbiasedness

In the previous section, the concept of sufficiency was not used in our search for unbiased estimates. We will now see that consideration of sufficiency is a powerful tool, indeed.

The main theorem of this section, which relates sufficient statistics to unbiased estimates is, as in the case of the Cramér–Rao Theorem, another clever application of some well-known theorems. Recall from Chapter 4 that if  $X$  and  $Y$  are any two random variables then, provided the expectations exist, we have

$$\begin{aligned} EX &= E[E(X|Y)], \\ (7.3.12) \quad \text{Var } X &= \text{Var}[E(X|Y)] + E[\text{Var}(X|Y)]. \end{aligned}$$

Using these tools we can prove the following theorem.

**THEOREM 7.3.2 (Rao–Blackwell):** Let  $W$  be any unbiased estimator of  $\tau(\theta)$ , and let  $T$  be a sufficient statistic for  $\theta$ . Define  $\phi(T) = E(W|T)$ . Then  $E_\theta \phi(T) = \tau(\theta)$  and  $\text{Var}_\theta \phi(T) \leq \text{Var}_\theta W$  for all  $\theta$ , that is,  $\phi(T)$  is a uniformly better unbiased estimator of  $\tau(\theta)$ .

*Proof:* From (7.3.12) we have

$$\tau(\theta) = E_\theta W = E_\theta [E(W|T)] = E_\theta \phi(T),$$

so  $\phi(T)$  is unbiased for  $\tau(\theta)$ . Also,

$$\begin{aligned} \text{Var}_\theta W &= \text{Var}_\theta [E(W|T)] + E_\theta [\text{Var}(W|T)] \\ &= \text{Var}_\theta \phi(T) + E_\theta [\text{Var}(W|T)] \\ &\geq \text{Var}_\theta \phi(T). \end{aligned} \quad (\text{Var}(W|T) \geq 0)$$

Hence  $\phi(T)$  is uniformly better than  $W$ , and it only remains to show that  $\phi(T)$  is indeed an estimator. That is, we must show that  $\phi(T) = E(W|T)$  is a function of only the sample and, in particular, is independent of  $\theta$ . But it follows from the definition of sufficiency, and the fact that  $W$  is a function only of the sample, that the distribution of  $W|T$  is independent of  $\theta$ . Hence  $\phi(T)$  is a uniformly better unbiased estimator of  $\tau(\theta)$ .  $\square$

Therefore, conditioning any unbiased estimator on a sufficient statistic will result in a uniform improvement, so we need consider only statistics that are functions of a sufficient statistic in our search for best unbiased estimators.

The identities in (7.3.12) make no mention of sufficiency, so it might at first seem that conditioning on anything will result in an improvement. This is, in effect, true, but the problem is that the resulting quantity will probably depend on  $\theta$  and not be an estimator.

**Example 7.3.7:** Let  $X_1, X_2$  be iid  $n(\theta, 1)$ . The statistic  $\bar{X} = \frac{1}{2}(X_1 + X_2)$  has

$$E_\theta \bar{X} = \theta \quad \text{and} \quad \text{Var}_\theta \bar{X} = \frac{1}{2}.$$

Consider conditioning on  $X_1$ , which is not sufficient. Let  $\phi(X_1) = E_\theta(\bar{X}|X_1)$ . It follows from (7.3.12) that  $E_\theta \phi(X_1) = \theta$  and  $\text{Var}_\theta \phi(X_1) \leq \text{Var}_\theta \bar{X}$ , so  $\phi(X_1)$  is better than  $\bar{X}$ . However,

$$\begin{aligned} \phi(X_1) &= E_\theta(\bar{X}|X_1) \\ &= \frac{1}{2}E_\theta(X_1|X_1) + \frac{1}{2}E_\theta(X_2|X_1) \\ &= \frac{1}{2}X_1 + \frac{1}{2}\theta, \end{aligned}$$

since  $E_\theta(X_2|X_1) = E_\theta X_2$  by independence. Hence,  $\phi(X_1)$  is not an estimator. ||

We now know that, in looking for a best unbiased estimator of  $\tau(\theta)$ , we need consider only estimators based on a sufficient statistic. The question now arises that if we have  $E_\theta \phi = \tau(\theta)$ , and  $\phi$  is based on a sufficient statistic, that is,  $E(\phi|T) = \phi$ , how do we know that  $\phi$  is best unbiased? Of course, if  $\phi$  attains the Cramér–Rao Lower Bound then it is best unbiased, but if it does not, have we gained anything? For example, if  $\phi^*$  is another unbiased estimator of  $\tau(\theta)$ , how does  $E(\phi^*|T)$  compare to  $\phi$ ? The next theorem answers this question in part, by showing that a best unbiased estimator is unique.

**THEOREM 7.3.3:** If  $W$  is a best unbiased estimator of  $\tau(\theta)$ , then  $W$  is unique.

*Proof:* Suppose  $W'$  is another best unbiased estimator, and consider the estimator  $W^* = \frac{1}{2}(W + W')$ . Note that  $E_\theta W^* = \tau(\theta)$  and

$$\begin{aligned} \text{Var}_\theta W^* &= \text{Var}_\theta \left( \frac{1}{2}W + \frac{1}{2}W' \right) \\ &= \frac{1}{4} \text{Var}_\theta W + \frac{1}{4} \text{Var}_\theta W' + \frac{1}{2} \text{Cov}_\theta(W, W') \quad (\text{Exercise 5.8}) \\ (7.3.13) \quad &\leq \frac{1}{4} \text{Var}_\theta W + \frac{1}{4} \text{Var}_\theta W' + \frac{1}{2} [\text{Var}_\theta W]^{1/2} [\text{Var}_\theta W']^{1/2} \quad (\text{Cauchy–Schwarz}) \\ &= \text{Var}_\theta W. \quad (\text{since } \text{Var}_\theta W = \text{Var}_\theta W') \end{aligned}$$

But if the above inequality is strict then the best unbiasedness of  $W$  is contradicted, so we must have equality for all  $\theta$ . Since the inequality is an application of Cauchy-Schwarz, we can have equality only if  $W' = a(\theta)W + b(\theta)$ . Now using properties of covariance we have

$$\begin{aligned}\text{Cov}_\theta(W, W') &= \text{Cov}_\theta[W, a(\theta)W + b(\theta)] \\ &= \text{Cov}_\theta[W, a(\theta)W] \\ &= a(\theta)\text{Var}_\theta W,\end{aligned}$$

but  $\text{Cov}_\theta(W, W') = \text{Var}_\theta W$  since we had equality in (7.3.13). Hence  $a(\theta) = 1$  and, since  $E_\theta W' = \tau(\theta)$ , we must have  $b(\theta) = 0$  and  $W = W'$ , showing that  $W$  is unique.  $\square$

To see when an unbiased estimator is best unbiased, we might ask how could we improve upon a given unbiased estimator? Suppose that  $W$  satisfies  $E_\theta W = \tau(\theta)$ , and we have another estimator,  $U$ , that satisfies  $E_\theta U = 0$  for all  $\theta$ , that is,  $U$  is an *unbiased estimator of zero*. The estimator

$$\phi_a = W + aU,$$

where  $a$  is a constant, satisfies  $E_\theta \phi_a = \tau(\theta)$ , and hence is also an unbiased estimator of  $\tau(\theta)$ . Can  $\phi_a$  be better than  $W$ ? The variance of  $\phi_a$  is

$$\text{Var}_\theta \phi_a = \text{Var}_\theta (W + aU) = \text{Var}_\theta W + 2a\text{Cov}_\theta(W, U) + a^2\text{Var}_\theta U.$$

Now, if for some  $\theta = \theta_0$ ,  $\text{Cov}_{\theta_0}(W, U) < 0$ , then we can make  $2a\text{Cov}_{\theta_0}(W, U) + a^2\text{Var}_{\theta_0} U < 0$  by choosing  $a \in (0, -2\text{Cov}_{\theta_0}(W, U)/\text{Var}_{\theta_0} U)$ . Hence,  $\phi_a$  will be better than  $W$  at  $\theta = \theta_0$  and  $W$  cannot be best unbiased. A similar argument will show that if  $\text{Cov}_{\theta_0}(W, U) > 0$  for any  $\theta_0$ ,  $W$  also cannot be best unbiased. (See Exercise 7.49.) Thus, the relationship of  $W$  with unbiased estimators of zero is crucial in evaluating whether  $W$  is best unbiased. This relationship, in fact, characterizes best unbiasedness.

**THEOREM 7.3.4:** If  $E_\theta W = \tau(\theta)$ ,  $W$  is the best unbiased estimator of  $\tau(\theta)$  if and only if  $W$  is uncorrelated with all unbiased estimators of zero.

*Proof:* If  $W$  is best unbiased, the above argument shows that  $W$  must satisfy  $\text{Cov}_\theta(W, U) = 0$  for all  $\theta$ , for any  $U$  satisfying  $E_\theta U = 0$ . Hence the necessity is established.

Suppose now that we have an unbiased estimator  $W$  that is uncorrelated with all unbiased estimators of zero. Let  $W'$  be any other estimator satisfying  $E_\theta W' = E_\theta W = \tau(\theta)$ . We will show that  $W$  is better than  $W'$ . Write

$$W' = W + (W' - W),$$

and calculate

$$(7.3.14) \quad \begin{aligned} \text{Var}_\theta W' &= \text{Var}_\theta W + \text{Var}_\theta (W' - W) + 2\text{Cov}_\theta(W, W' - W) \\ &= \text{Var}_\theta W + \text{Var}_\theta (W' - W), \end{aligned}$$

where the last equality is true because  $W' - W$  is an unbiased estimator of zero and is uncorrelated with  $W$  by assumption. Since  $\text{Var}_\theta (W' - W) \geq 0$ , (7.3.14) implies that  $\text{Var}_\theta W' \geq \text{Var}_\theta W$ . Since  $W'$  is arbitrary, it follows that  $W$  is the best unbiased estimator of  $\tau(\theta)$ .  $\square$

Note that an unbiased estimator of zero is nothing more than *random noise*, that is, there is no information in an estimator of zero. (It makes sense that the most sensible way to estimate zero is with zero, not with random noise.) Therefore, if an estimator could be improved by adding random noise to it, the estimator probably is defective. (Alternatively, we could question the criterion used to evaluate the estimator, but in this case the criterion seems above suspicion.) This intuition is what is formalized in Theorem 7.3.4.

Although we now have an interesting characterization of best unbiased estimators, its usefulness is limited in application. It is often a difficult task to verify that an estimator is uncorrelated with *all* unbiased estimators of zero because it is usually difficult to describe all unbiased estimators of zero. However, it is sometimes useful in determining that an estimator is not best unbiased.

**Example 7.3.8:** Let  $X$  be an observation from a uniform( $\theta, \theta+1$ ) distribution. Then

$$\text{E}_\theta X = \int_\theta^{\theta+1} x dx = \theta + \frac{1}{2},$$

and so  $X - \frac{1}{2}$  is an unbiased estimator of  $\theta$ , and it is easy to check that  $\text{Var}_\theta X = \frac{1}{12}$ .

For this pdf, unbiased estimators of zero are periodic functions with period 1. This follows from the fact that if  $h(x)$  satisfies

$$\int_\theta^{\theta+1} h(x) dx = 0, \quad \text{for all } \theta,$$

then

$$0 = \frac{d}{d\theta} \int_\theta^{\theta+1} h(x) dx = h(\theta+1) - h(\theta), \quad \text{for all } \theta.$$

Such a function is  $h(x) = \sin(2\pi x)$ . Now

$$\text{Cov}_\theta(X - \frac{1}{2}, \sin(2\pi X)) = \text{Cov}_\theta(X, \sin(2\pi X))$$

$$= \int_\theta^{\theta+1} x \sin(2\pi x) dx$$

$$\begin{aligned}
 &= -\frac{x \cos(2\pi x)}{2\pi} \Big|_{\theta}^{\theta+1} + \int_{\theta}^{\theta+1} \frac{\cos(2\pi x)}{2\pi} dx \\
 &= -\frac{\cos(2\pi\theta)}{2\pi},
 \end{aligned}$$

(integration by parts)

where we used  $\cos(2\pi(\theta+1)) = \cos(2\pi\theta)$  and  $\sin(2\pi(\theta+1)) = \sin(2\pi\theta)$ .

Hence  $X - \frac{1}{2}$  is correlated with an unbiased estimator of zero, and cannot be a best unbiased estimator of  $\theta$ . In fact, it is straightforward to check that the estimator  $X - \frac{1}{2} + \sin(2\pi X)/(2\pi)$  is unbiased for  $\theta$  and has variance  $.071 < \frac{1}{12}$ . ||

To answer the question about existence of a best unbiased estimator, what is needed is some characterization of all unbiased estimators of zero. Given such a characterization, we could then see if our candidate for best unbiased estimator is, in fact, optimal.

Characterizing the unbiased estimators of zero is not an easy task and requires conditions on the pdf (or pmf) with which we are working. Note that, thus far in this section, we have not specified conditions on pdfs (as were needed, for example, in the Cramér–Rao Lower Bound). The price we have paid for this generality is the difficulty in verifying the existence of the best unbiased estimator.

If a family of pdfs or pmfs  $f(x|\theta)$  has the property that there are *no* unbiased estimators of zero (other than zero itself), then our search would be ended, since any statistic  $W$  satisfies  $\text{Cov}_{\theta}(W, 0) = 0$ . Recall that the property of *completeness*, defined in Definition 6.1.4, guarantees such a situation.

**Example 7.3.5 (Continued):** For  $X_1, \dots, X_n$  iid uniform  $(0, \theta)$ , we saw that  $\frac{n+1}{n}Y$  is an unbiased estimator of  $\theta$ , where  $Y = \max\{X_1, \dots, X_n\}$ . The conditions of the Cramér–Rao Theorem are not satisfied, and we have not yet established whether this estimator is best unbiased. In Example 6.1.13, however, it was shown that  $Y$  is a *complete* sufficient statistic. This means that the family of pdfs of  $Y$  is complete, and there are no unbiased estimators of zero that are based on  $Y$ . (By sufficiency, in the form of the Rao–Blackwell Theorem, we need consider only unbiased estimators of zero based on  $Y$ .) Therefore,  $\frac{n+1}{n}Y$  is uncorrelated with all unbiased estimators of zero (since the only one is zero itself) and thus  $\frac{n+1}{n}Y$  is the best unbiased estimator of  $\theta$ . ||

It is worthwhile to note once again that what is important is the completeness of the family of distributions of the sufficient statistic. Completeness of the original family is of no consequence. This follows from the Rao–Blackwell Theorem, which says that we can restrict attention to functions of a sufficient statistic, so all expectations will be taken with respect to its distribution.

We sum up the relationship between completeness and best unbiasedness in the following theorem.

**THEOREM 7.3.5:** Let  $T$  be a complete sufficient statistic for a parameter  $\theta$ , and let  $\phi(T)$  be any estimator based only on  $T$ . Then  $\phi(T)$  is the unique best unbiased estimator of its expected value. □

We close this section with an interesting and useful application of the theory developed here. In many situations, there will be no obvious candidate for an unbiased estimator of a function  $\tau(\theta)$ , much less a candidate for best unbiased estimator. However, in the presence of completeness, the theory of this section tells us that if we can find any unbiased estimator, we can find the best unbiased estimator. If  $T$  is a complete sufficient statistic for a parameter  $\theta$ , and  $h(X_1, \dots, X_n)$  is any unbiased estimator of  $\tau(\theta)$ , then  $\phi(T) = E(h(X_1, \dots, X_n)|T)$  is the best unbiased estimator of  $\tau(\theta)$  (see Exercise 7.53).

**Example 7.3.9:** Let  $X_1, \dots, X_n$  be iid binomial  $(k, \theta)$ . The problem is to estimate the probability of exactly one success from a binomial  $(k, \theta)$ , that is, estimate

$$\tau(\theta) = P_\theta(X = 1) = k\theta(1 - \theta)^{k-1}.$$

Now  $\sum_{i=1}^n X_i \sim \text{binomial}(kn, \theta)$  is a complete sufficient statistic, but no unbiased estimator based on it is immediately evident. When in this situation, try for the simplest solution. The simple-minded estimator

$$h(X_1) = \begin{cases} 1 & \text{if } X_1 = 1 \\ 0 & \text{otherwise} \end{cases}$$

satisfies

$$\begin{aligned} E_\theta h(X_1) &= \sum_{x_1=0}^k h(x_1) \binom{k}{x_1} \theta^{x_1} (1 - \theta)^{k-x_1} \\ &= k\theta(1 - \theta)^{k-1}, \end{aligned}$$

and hence is an unbiased estimator of  $k\theta(1 - \theta)^{k-1}$ . Our theory now tells us that the estimator

$$\phi\left(\sum_{i=1}^n X_i\right) = E\left(h(X_1) \mid \sum_{i=1}^n X_i\right)$$

is the best unbiased estimator of  $k\theta(1 - \theta)^{k-1}$ . (Notice that we do not need to actually calculate the expectation of  $\phi(\sum_{i=1}^n X_i)$ ; we know that it has the correct expected value by properties of iterated expectations.) We must, however, be able to evaluate  $\phi$ . Suppose that we observe  $\sum_{i=1}^n X_i = t$ . Then

$$\begin{aligned} \phi(t) &= E\left(h(X_1) \mid \sum_{i=1}^n X_i = t\right) && (\text{the expectation does not depend on } \theta) \\ &= P\left(X_1 = 1 \mid \sum_{i=1}^n X_i = t\right) && (h \text{ is 0 or 1}) \end{aligned}$$

$$\begin{aligned}
 &= \frac{P_\theta(X_1 = 1, \sum_{i=1}^n X_i = t)}{P_\theta(\sum_{i=1}^n X_i = t)} \quad (\text{definition of conditional probability}) \\
 &= \frac{P_\theta(X_1 = 1, \sum_{i=2}^n X_i = t - 1)}{P_\theta(\sum_{i=1}^n X_i = t)} \quad (X_1 = 1 \text{ is redundant}) \\
 &= \frac{P_\theta(X_1 = 1) P_\theta(\sum_{i=2}^n X_i = t - 1)}{P_\theta(\sum_{i=1}^n X_i = t)}. \quad (X_1 \text{ is independent of } X_2, \dots, X_n)
 \end{aligned}$$

Now  $X_1 \sim \text{binomial}(k, \theta)$ ,  $\sum_{i=2}^n X_i \sim \text{binomial}(k(n-1), \theta)$ , and  $\sum_{i=1}^n X_i \sim \text{binomial}(kn, \theta)$ . Using these facts we have

$$\begin{aligned}
 \phi(t) &= \frac{[k\theta(1-\theta)^{k-1}] \left[ \binom{k(n-1)}{t-1} \theta^{t-1} (1-\theta)^{k(n-1)-(t-1)} \right]}{\binom{kn}{t} \theta^t (1-\theta)^{kn-t}} \\
 &= k \frac{\binom{k(n-1)}{t-1}}{\binom{kn}{t}}.
 \end{aligned}$$

Note that all of the  $\theta$ s cancel, as they must since  $\sum_{i=1}^n X_i$  is sufficient. Hence, the best unbiased estimator of  $k\theta(1-\theta)^{k-1}$  is

$$\phi\left(\sum_{i=1}^n X_i\right) = k \frac{\binom{k(n-1)}{\sum X_i - 1}}{\binom{kn}{\sum X_i}}.$$

We can assert unbiasedness without performing the difficult evaluation of  $E_\theta[\phi(\sum_{i=1}^n X_i)]$ .

### 7.3.4 Consistency

All of the criteria we have considered thus far have been finite-sample criteria. In contrast, the property of consistency is an asymptotic one, describing the behavior of an estimator as the sample size becomes infinite. It is one of the few asymptotic criteria we will examine, and is important in that the worth of an inconsistent estimator should be seriously questioned.

Consistency is a property of a sequence of estimators rather than of a single estimator, although it is common to speak of a "consistent estimator." If we observe  $X_1, X_2, \dots$  according to a distribution  $f(x|\theta)$ , we can construct a sequence of estimators  $W_n = W_n(X_1, \dots, X_n)$  merely by performing the same estimation procedure for each sample size  $n$ . For example,  $\bar{X}_1 = X_1$ ,  $\bar{X}_2 = (X_1 + X_2)/2$ ,  $\bar{X}_3 = (X_1 + X_2 + X_3)/3$ , etc. We can now define a consistent sequence.

**DEFINITION 7.3.4:** A sequence of estimators  $W_n = W_n(X_1, \dots, X_n)$  is a *consistent sequence of estimators* of the parameter  $\theta$  if, for every  $\epsilon > 0$  and every  $\theta \in \Theta$ ,

$$(7.3.15) \quad \lim_{n \rightarrow \infty} P_\theta(|W_n - \theta| < \epsilon) = 1.$$

Informally, (7.3.15) says that as the sample size becomes infinite (and the sample information becomes better and better), the estimator will be arbitrarily close to the parameter with high probability, an eminently desirable property. Or, turning things around, we can say that the probability that a consistent sequence of estimators misses the true parameter is small. An equivalent statement to (7.3.15) is this. For every  $\epsilon > 0$  and every  $\theta \in \Theta$ , a consistent sequence  $W_n$  will satisfy

$$(7.3.16) \quad \lim_{n \rightarrow \infty} P_\theta(|W_n - \theta| \geq \epsilon) = 0.$$

Definition 7.3.4 should be compared to Definition 5.3.1, the definition of convergence in probability. Definition 7.3.4 says that a consistent sequence of estimators converges in probability to the parameter  $\theta$  it is estimating. Whereas Definition 5.3.1 dealt with one sequence of random variables with one probability structure, Definition 7.3.4 deals with an entire family of probability structures, indexed by  $\theta$ . For each different value of  $\theta$ , the probability structure associated with the sequence  $W_n$  is different. And the definition says that for each value of  $\theta$ , the probability structure is such that the sequence converges in probability to the true  $\theta$ . This is the usual difference between a probability definition and a statistics definition. The probability definition deals with one probability structure but the statistics definition deals with an entire family.

**Example 7.3.10:** Let  $X_1, X_2, \dots$  be iid  $n(\theta, 1)$ , and consider the sequence

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Recall that  $\bar{X}_n \sim n(\theta, 1/n)$ , so

$$\begin{aligned} P_\theta(|\bar{X}_n - \theta| < \epsilon) &= \int_{\theta-\epsilon}^{\theta+\epsilon} \left( \frac{n}{2\pi} \right)^{\frac{1}{2}} e^{-(n/2)(\bar{x}_n - \theta)^2} d\bar{x}_n && \text{(definition)} \\ &= \int_{-\epsilon}^{\epsilon} \left( \frac{n}{2\pi} \right)^{\frac{1}{2}} e^{-(n/2)y^2} dy && \text{(substitute } y = \bar{x}_n - \theta) \\ &= \int_{-\epsilon\sqrt{n}}^{\epsilon\sqrt{n}} \left( \frac{1}{2\pi} \right)^{\frac{1}{2}} e^{-(1/2)t^2} dt && \text{(substitute } t = y\sqrt{n}) \\ &= P(-\epsilon\sqrt{n} < Z < \epsilon\sqrt{n}) && (Z \sim n(0, 1)) \\ &\rightarrow 1 \quad \text{as } n \rightarrow \infty \end{aligned}$$

and, hence,  $\bar{X}_n$  is a consistent sequence of estimators of  $\theta$ . ||

In general, a detailed calculation, such as the above, is not necessary to verify consistency. Recall that, for an estimator  $W_n$ , Chebychev's Inequality states

$$P_\theta(|W_n - \theta| \geq \epsilon) \leq \frac{E_\theta((W_n - \theta)^2)}{\epsilon^2},$$

so it follows that

$$\lim_{n \rightarrow \infty} E_\theta((W_n - \theta)^2) = 0$$

is a sufficient condition for a sequence to be consistent. Furthermore, we can write

$$\begin{aligned} E_\theta((W_n - \theta)^2) &= E_\theta((W_n - E_\theta W_n + E_\theta W_n - \theta)^2) \\ (7.3.17) \quad &= E_\theta(W_n - E_\theta W_n)^2 + (E_\theta W_n - \theta)^2 \quad (\text{cross-term is zero}) \\ &= \text{Var}_\theta W_n + [\text{Bias}_\theta W_n]^2 \end{aligned}$$

and, putting this all together, we can state the following theorem.

**THEOREM 7.3.6:** If  $W_n$  is a sequence of estimators of a parameter  $\theta$  satisfying

- a.  $\lim_{n \rightarrow \infty} \text{Var}_\theta W_n = 0$ ,
- b.  $\lim_{n \rightarrow \infty} \text{Bias}_\theta W_n = 0$ ,

then  $W_n$  is a consistent sequence of estimators of  $\theta$ . □

**Example 7.3.10 (Continued):** Since

$$E_\theta \bar{X}_n = \theta \quad \text{and} \quad \text{Var}_\theta \bar{X}_n = \frac{1}{n},$$

the conditions of Theorem 7.3.6 are satisfied and the sequence  $\bar{X}_n$  is consistent. Furthermore, from Theorem 5.2.2, if there is iid sampling from any population with mean  $\theta$  then  $\bar{X}_n$  is consistent for  $\theta$  as long as the population has a finite variance. ||

At the beginning of this section we commented that the worth of an inconsistent sequence of estimators should be questioned. Part of the basis for this comment is the fact that there are so many consistent sequences, as the next theorem shows.

**THEOREM 7.3.7:** Let  $W_n$  be a consistent sequence of estimators of a parameter  $\theta$ . Let  $a_1, a_2, \dots$  and  $b_1, b_2, \dots$  be sequences of constants satisfying

- a.  $\lim_{n \rightarrow \infty} a_n = 1$
- b.  $\lim_{n \rightarrow \infty} b_n = 0$ .

Then the sequence  $U_n = a_n W_n + b_n$  is a consistent sequence of estimators of  $\theta$ .

*Proof:* Exercise 7.60. □

We close this section with the outline of a more general result concerning the consistency of maximum likelihood estimators. This result shows that MLEs are consistent estimators of their parameters, and is the first case we have seen in which a method of finding an estimator guarantees an optimality property. The result presented here holds in greater generality. For more details see Wald (1949), Kendall and Stuart (1979), or Huber (1967).

**THEOREM 7.3.8 (Consistency of MLEs):** Let  $X_1, \dots, X_n$  be iid  $f(x|\theta)$ , and let  $L(\theta|x) = \prod_{i=1}^n f(x_i|\theta)$  be the likelihood function. Let  $\hat{\theta}$  denote the MLE of  $\theta$ . Let  $\tau(\theta)$  be a continuous function of  $\theta$ . Under certain regularity conditions on  $f(x|\theta)$  and, hence,  $L(\theta|x)$ , for every  $\epsilon > 0$  and every  $\theta \in \Theta$ ,

$$\lim_{n \rightarrow \infty} P_\theta(|\tau(\hat{\theta}) - \tau(\theta)| \geq \epsilon) = 0.$$

That is,  $\tau(\hat{\theta})$  is a consistent estimator of  $\tau(\theta)$ .

*Proof:* The proof proceeds by showing that  $\frac{1}{n} \log L(\hat{\theta}|x)$  converges almost surely to  $E_\theta(\log f(X|\theta))$  for every  $\theta \in \Theta$ . Under some conditions on  $f(x|\theta)$ , this implies that  $\hat{\theta}$  converges to  $\theta$  in probability and, hence,  $\tau(\hat{\theta})$  converges to  $\tau(\theta)$  in probability. For details see Kendall and Stuart (1979, Chapter 18).  $\square$

## 7.4 Other Considerations

In this section we consider some related ideas, not necessarily concerned with obtaining optimal answers, but more concerned with obtaining *some* answers.

### 7.4.1 Asymptotic Variance of Maximum Likelihood Estimators

A variance approximation that is useful for maximum likelihood estimators is based on a general property of MLEs, that of asymptotic efficiency.

**DEFINITION 7.4.1:** A sequence of estimators  $W_n$  is *asymptotically efficient* for a parameter  $\tau(\theta)$  if

$$\lim_{n \rightarrow \infty} \frac{\text{Var}_\theta W_n}{\left[ \frac{[\tau'(\theta)]^2}{n E_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right)} \right]} = 1,$$

that is,  $W_n$  achieves the Cramér–Rao Lower Bound as  $n \rightarrow \infty$ .

Recall that Theorem 7.3.8 stated that, under general conditions, MLEs are consistent. The same type of theorem holds with respect to asymptotic efficiency so, in general, we can consider MLEs to be consistent and asymptotically efficient. Again, we have to be concerned with regularity conditions, but these are quite general, and almost always satisfied in common circumstances. One condition deserves special mention, however, whose violation can lead to complications, as we have already

seen in Example 7.3.5. For the following approximations to be valid, it must be the case that the support of the pdf or pmf, hence likelihood function, must be independent of the parameter.

Assuming that an MLE is asymptotically efficient, we can use the Cramér–Rao Lower Bound as an approximation to the true variance of the MLE. Furthermore, it has been shown (Efron and Hinkley, 1978) that use of the *observed information number* is superior to the *expected information number*, the information number as it appears in the Cramér–Rao Lower Bound. Following the steps of the proof of Theorem 7.3.1, the expected information number is

$$\begin{aligned} nE_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right) &= E_\theta \left( \left( \frac{\partial}{\partial \theta} \log \prod_{i=1}^n f(X_i|\theta) \right)^2 \right) \\ &= E_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta) \right)^2 \right) \\ &= E_\theta \left( \left( \frac{\partial}{\partial \theta} \log L(\theta|\mathbf{X}) \right)^2 \right). \quad \left( \begin{array}{l} L(\theta|\mathbf{X}) \text{ is the} \\ \text{likelihood function} \end{array} \right) \end{aligned}$$

We now estimate the expected information number to get an estimate for the variance of a function  $h(\hat{\theta})$  like this. If  $X_1, \dots, X_n$  are iid  $f(x|\theta)$ , and  $\hat{\theta}$  is the MLE of  $\theta$ , then using the asymptotic efficiency of MLEs, the variance of  $h(\hat{\theta})$  can be approximated by

$$\begin{aligned} \text{Var}(h(\hat{\theta})|\theta) &\approx \frac{[h'(\theta)]^2}{E_\theta \left( -\frac{\partial^2}{\partial \theta^2} \log L(\theta|\mathbf{X}) \right)} \quad \left( \begin{array}{l} \text{using the identity} \\ \text{of Lemma 7.3.1} \end{array} \right) \\ (7.4.1) \quad &\approx \frac{[h'(\theta)]^2|_{\theta=\hat{\theta}}}{-\frac{\partial^2}{\partial \theta^2} \log L(\theta|\mathbf{X})|_{\theta=\hat{\theta}}}. \quad \left( \begin{array}{l} \text{the denominator is the} \\ \text{observed information number} \end{array} \right) \end{aligned}$$

Notice that the variance estimation process is a two-step procedure, a fact that is somewhat masked by (7.4.1). To estimate  $\text{Var}_\theta h(\hat{\theta})$  first we *approximate*  $\text{Var}_\theta h(\hat{\theta})$ , then we *estimate* the resulting approximation, usually by substituting  $\hat{\theta}$  for  $\theta$ . The resulting estimate can be denoted by  $\text{Var}_{\hat{\theta}} h(\hat{\theta})$  or  $\widehat{\text{Var}}_\theta h(\hat{\theta})$ .

An argument similar to that in Theorem 7.3.8 shows that  $-\frac{\partial^2}{\partial \theta^2} \log L(\theta|\mathbf{X})|_{\theta=\hat{\theta}}$  is a reasonable estimator of  $nE_\theta(-\frac{\partial^2}{\partial \theta^2} \log f(X|\theta))$ .

**Example 7.4.1:** In Example 7.2.5 we saw that  $\hat{p} = \sum X_i/n$  is the MLE of  $p$ , where we have a random sample  $X_1, \dots, X_n$  from a  $\text{Bernoulli}(p)$  population. We also know by direct calculation that

$$\text{Var}_p \hat{p} = \frac{p(1-p)}{n},$$

and a reasonable estimate of  $\text{Var}_p \hat{p}$  is

$$(7.4.2) \quad \widehat{\text{Var}}_p \hat{p} = \frac{\hat{p}(1 - \hat{p})}{n}.$$

If we apply the approximation in (7.4.1), with  $h(p) = p$ , we get as an estimate of  $\text{Var}_p \hat{p}$

$$\widehat{\text{Var}}_p \hat{p} \approx \frac{1}{-\frac{\partial^2}{\partial p^2} \log L(p|\mathbf{x})|_{p=\hat{p}}}.$$

Recall that

$$\log L(p|\mathbf{x}) = n\hat{p} \log(p) + n(1 - \hat{p}) \log(1 - p),$$

and so

$$\frac{\partial^2}{\partial p^2} \log L(p|\mathbf{x}) = -\frac{n\hat{p}}{p^2} - \frac{n(1 - \hat{p})}{(1 - p)^2}.$$

Evaluating the second derivative at  $p = \hat{p}$  yields

$$\frac{\partial^2}{\partial p^2} \log L(p|\mathbf{x})|_{p=\hat{p}} = -\frac{n\hat{p}}{\hat{p}^2} - \frac{n(1 - \hat{p})}{(1 - \hat{p})^2} = -\frac{n}{\hat{p}(1 - \hat{p})},$$

which gives a variance approximation identical to (7.4.2).

Estimating the variance of  $\hat{p}$  is not really that difficult and it is not necessary to bring in all of the machinery of these approximations. If we move to a slightly more complicated function, however, things can get a bit tricky. A quantity that is sometimes of interest in binomial problems is the *odds ratio*, which is given by  $p/(1 - p)$ . The MLE of the odds ratio is simply  $\hat{p}/(1 - \hat{p})$ , but how might we estimate the variance of  $\hat{p}/(1 - \hat{p})$ ? Intuition abandons us, and exact calculation is relatively hopeless, so we have to rely on an approximation. Using (7.4.1), we get the approximation

$$\begin{aligned} \widehat{\text{Var}} \left( \frac{\hat{p}}{1 - \hat{p}} \right) &\approx \frac{\left[ \frac{\partial}{\partial p} \left( \frac{p}{1-p} \right) \right]^2 \Big|_{p=\hat{p}}}{-\frac{\partial^2}{\partial p^2} \log L(p|\mathbf{x}) \Big|_{p=\hat{p}}} \\ &= \frac{\left[ \frac{(1-p)+p}{(1-p)^2} \right]^2 \Big|_{p=\hat{p}}}{\frac{n}{p(1-p)} \Big|_{p=\hat{p}}} \\ &= \frac{\hat{p}}{n(1 - \hat{p})^3}. \end{aligned}$$

||

The MLE variance approximation works well in many cases, but it is not infallible. In particular, we must be careful when the function  $h(\hat{\theta})$  is not monotone. In such cases, the derivative  $h'$  will have a sign change, and that may lead to an underestimated variance approximation. Realize that, since the approximation is based on the Cramér–Rao Lower Bound, it is probably an underestimate. However, nonmonotone functions can make this problem worse.

**Example 7.4.1 (Continued):** Suppose now that we want to estimate the variance of the Bernoulli distribution,  $p(1 - p)$ . The MLE of this variance is given by  $\hat{p}(1 - \hat{p})$ , and an estimate of the variance of this estimator can be obtained by applying the approximation of (7.4.1). We have

$$\begin{aligned}\widehat{\text{Var}}(\hat{p}(1 - \hat{p})) &\approx \frac{\left[ \frac{\partial}{\partial p} (p(1 - p)) \right]^2 \Big|_{p=\hat{p}}}{-\frac{\partial^2}{\partial p^2} \log L(p|\mathbf{x}) \Big|_{p=\hat{p}}} \\ &= \frac{(1 - 2p)^2 \Big|_{p=\hat{p}}}{\frac{n}{p(1-p)} \Big|_{p=\hat{p}}} \\ &= \frac{\hat{p}(1 - \hat{p})(1 - 2\hat{p})^2}{n},\end{aligned}$$

which can be zero if  $\hat{p} = \frac{1}{2}$ , a clear underestimate of the variance of  $\hat{p}(1 - \hat{p})$ . The fact that the function  $p(1 - p)$  is not monotone is a cause of this problem (see Exercise 7.65). ||

## 7.4.2 Taylor Series Approximations

Another method of approximation is based on a mathematical, rather than statistical, argument. Taylor series provide a means of approximating a function through polynomials and give a natural method for obtaining variance approximations.

**DEFINITION 7.4.2:** If a function  $g(x)$  has derivatives of order  $r$ , that is,  $g^{(r)}(x) = \frac{d^r}{dx^r} g(x)$  exists, then for any constant  $a$ , the *Taylor polynomial of order r about a* is

$$T_r(x) = \sum_{i=0}^r \frac{g^{(i)}(a)}{i!} (x - a)^i.$$

Taylor's major theorem, which we will not prove here, is that the *remainder* from the approximation,  $g(x) - T_r(x)$  always tends to zero faster than the highest-order explicit term.

**THEOREM 7.4.1 (Taylor):** If  $g^{(r)}(a) = \frac{d^r}{dx^r} g(x)|_{x=a}$  exists, then

$$\lim_{x \rightarrow a} \frac{g(x) - T_r(x)}{(x - a)^r} = 0$$
□

In general, we will not be concerned with the explicit form of the remainder. Since we are interested in approximations, we are just going to ignore the remainder. There are, however, many explicit forms, one useful one being

$$g(x) - T_r(x) = \int_a^x \frac{g^{(r+1)}(t)}{r!} (x-t)^r dt.$$

For the statistical application of Taylor's Theorem, we are most concerned with the *first-order* Taylor series, that is, an approximation using just the first derivative (taking  $r = 1$  in the above formulas). Furthermore, we will also find use for a multivariate Taylor series. Since the above detail is univariate, some of the following will have to be accepted on faith.

Let  $X_1, \dots, X_k$  be random variables with means  $\theta_1, \dots, \theta_k$ , and define  $\mathbf{X} = (X_1, \dots, X_k)$  and  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ . Suppose there is a differentiable function  $g(\mathbf{X})$  (an estimator of some parameter) for which we want an approximate estimate of variance. Define

$$g'_i(\boldsymbol{\theta}) = \frac{\partial}{\partial x_i} g(\mathbf{x})|_{x_1=\theta_1, \dots, x_k=\theta_k}.$$

The first-order Taylor series expansion of  $g$  about  $\boldsymbol{\theta}$  is

$$g(\mathbf{x}) = g(\boldsymbol{\theta}) + \sum_{i=1}^k g'_i(\boldsymbol{\theta})(x_i - \theta_i) + \text{Remainder.}$$

For our statistical approximation we forget about the remainder and write

$$(7.4.3) \quad g(\mathbf{x}) \approx g(\boldsymbol{\theta}) + \sum_{i=1}^k g'_i(\boldsymbol{\theta})(x_i - \theta_i).$$

Now, take expectations on both sides of (7.4.3) to get

$$(7.4.4) \quad \begin{aligned} E_\theta g(\mathbf{X}) &\approx g(\boldsymbol{\theta}) + \sum_{i=1}^k g'_i(\boldsymbol{\theta}) E_\theta (X_i - \theta_i) \\ &= g(\boldsymbol{\theta}). \end{aligned} \quad (X_i \text{ has mean } \theta_i)$$

We can now approximate the variance of  $g(\mathbf{X})$  by

$$\text{Var}_\theta g(\mathbf{X}) \approx E_\theta ([g(\mathbf{X}) - g(\boldsymbol{\theta})]^2) \quad (\text{using 7.4.4})$$

$$(7.4.5) \quad \begin{aligned} &\approx E_\theta \left( \left( \sum_{i=1}^k g'_i(\boldsymbol{\theta})(X_i - \theta_i) \right)^2 \right) \quad (\text{using 7.4.3}) \\ &= \sum_{i=1}^k [g'_i(\boldsymbol{\theta})]^2 \text{Var}_\theta X_i + 2 \sum_{i>j} g'_i(\boldsymbol{\theta}) g'_j(\boldsymbol{\theta}) \text{Cov}_\theta (X_i, X_j), \end{aligned}$$

where the last equality comes from expanding the square and using the definition of variance and covariance (similar to Exercise 5.8). Approximation (7.4.5) is very useful because it gives us a variance formula for a general function, using only simple variances and covariances. Remember, however, that one more practical step needs to be taken. We must now estimate this approximate variance, which we do by replacing parameters with estimates.

**Example 7.4.2:** Suppose  $\bar{X}$  is the mean of a random sample, where  $E_\mu \bar{X} = \mu$ . If we want to estimate a function  $g(\mu)$ , a first-order approximation would give us

$$g(\bar{X}) = g(\mu) + g'(\mu)(\bar{X} - \mu).$$

If we use  $g(\bar{X})$  as an estimator of  $g(\mu)$ , we can say that approximately

$$\begin{aligned} E_\mu g(\bar{X}) &\approx g(\mu), \\ \text{Var}_\mu g(\bar{X}) &\approx [g'(\mu)]^2 \text{Var}_\mu \bar{X}. \end{aligned}$$

For a specific example, take  $g(\mu) = 1/\mu$ . We estimate  $1/\mu$  with  $1/\bar{X}$ , and we can say

$$\begin{aligned} E_\mu \left( \frac{1}{\bar{X}} \right) &\approx \frac{1}{\mu}, \\ \text{Var}_\mu \left( \frac{1}{\bar{X}} \right) &\approx \left( \frac{1}{\mu} \right)^4 \text{Var}_\mu \bar{X}. \end{aligned}$$

If we now want to estimate the variance of  $1/\bar{X}$ , we would replace  $\mu$  with  $\bar{X}$  in the variance formula and get

$$\widehat{\text{Var}} \left( \frac{1}{\bar{X}} \right) \approx \left( \frac{1}{\bar{X}} \right)^4 \widehat{\text{Var}}(\bar{X}).$$

Approximation techniques are very useful when more than one parameter makes up the function to be estimated, and more than one random variable is used in the estimator. One common example is in growth studies, where a ratio of weight/height is a variable of interest. (Recall that in Chapter 3 we saw that a ratio of two *normal* random variables has a Cauchy distribution. The ratio problem, while being important to experimenters, is nasty in theory.) Such a circumstance can be handled with the following approximation.

**Example 7.4.3:** Suppose  $\bar{X}$  and  $\bar{Y}$  are random variables with means  $\mu_X$  and  $\mu_Y$ , respectively. The parametric function to be estimated is  $g(\mu_X, \mu_Y) = \mu_X/\mu_Y$ . It is straightforward to calculate

$$\frac{\partial}{\partial \mu_X} g(\mu_X, \mu_Y) = \frac{1}{\mu_Y}$$

and

$$\frac{\partial}{\partial \mu_Y} g(\mu_X, \mu_Y) = \frac{-\mu_X}{\mu_Y^2}.$$

The first-order Taylor approximations, (7.4.4) and (7.4.5), give

$$E\left(\frac{\bar{X}}{\bar{Y}}\right) \approx \frac{\mu_X}{\mu_Y},$$

and

$$\begin{aligned} \text{Var}\left(\frac{\bar{X}}{\bar{Y}}\right) &\approx \frac{1}{\mu_Y^2} \text{Var } \bar{X} + \frac{\mu_X^2}{\mu_Y^4} \text{Var } \bar{Y} - 2 \frac{\mu_X}{\mu_Y^3} \text{Cov}(\bar{X}, \bar{Y}) \\ &= \left(\frac{\mu_X}{\mu_Y}\right)^2 \left( \frac{\text{Var } \bar{X}}{\mu_X^2} + \frac{\text{Var } \bar{Y}}{\mu_Y^2} - 2 \frac{\text{Cov}(\bar{X}, \bar{Y})}{\mu_X \mu_Y} \right). \end{aligned}$$

Thus, we have an approximation for the mean and variance of the ratio estimator, and the approximations only use the means, variances, and covariance of  $\bar{X}$  and  $\bar{Y}$ . Exact calculations would be quite hopeless, with closed-form expressions being unattainable. ||

## EXERCISES

---

- 7.1** One observation is taken on a discrete random variable  $X$  with pmf  $f(x|\theta)$ , where  $\theta \in \{1, 2, 3\}$ . Find the MLE of  $\theta$ .

$x$	$f(x 1)$	$f(x 2)$	$f(x 3)$
0	$\frac{1}{3}$	$\frac{1}{4}$	0
1	$\frac{1}{3}$	$\frac{1}{4}$	0
2	0	$\frac{1}{4}$	$\frac{1}{4}$
3	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{2}$
4	$\frac{1}{6}$	0	$\frac{1}{4}$

- 7.2** Let  $X_1, \dots, X_n$  be a random sample from a gamma( $\alpha, \beta$ ) population.

a. Find the MLE of  $\beta$ , assuming  $\alpha$  is known.

b. If  $\alpha$  and  $\beta$  are both unknown, there is no explicit formula for the MLEs of  $\alpha$  and  $\beta$ , but the maximum can be found numerically. The result in part (a) can be used to reduce

the problem to the maximization of a univariate function. Find the MLEs for  $\alpha$  and  $\beta$  for the data in Exercise 7.10(c).

- 7.3 Given a random sample  $X_1, \dots, X_n$  from a population with pdf  $f(x|\theta)$ , show that maximizing the likelihood function,  $L(\theta|x)$ , as a function of  $\theta$ , is equivalent to maximizing  $\log L(\theta|x)$ .
- 7.4 Prove the assertion in Example 7.2.6. That is, prove that  $\hat{\theta}$ , given there, is the MLE when the range of  $\theta$  is restricted to the positive axis.
- 7.5 Consider estimating the binomial parameter  $k$  as in Example 7.2.7.
- Prove the assertion that the integer  $\hat{k}$ , that satisfies the inequalities and is the MLE, is the largest integer less than or equal to  $1/\hat{z}$ .
  - Let  $p = \frac{1}{2}$ ,  $n = 4$  and  $X_1 = 0$ ,  $X_2 = 20$ ,  $X_3 = 1$ , and  $X_4 = 19$ . What is  $\hat{k}$ ?
- 7.6 Let  $X_1, \dots, X_n$  be a random sample from the pdf

$$f(x|\theta) = \theta x^{-2}, \quad 0 < \theta \leq x < \infty.$$

- What is a sufficient statistic for  $\theta$ ?
- Find the MLE of  $\theta$ .
- Find the method of moments estimator of  $\theta$ .

- 7.7 Let  $X_1, \dots, X_n$  be iid with one of two pdfs. If  $\theta = 0$  then

$$f(x|\theta) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases},$$

while if  $\theta = 1$

$$f(x|\theta) = \begin{cases} 1/(2\sqrt{x}) & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}.$$

Find the MLE of  $\theta$ .

- 7.8 One observation,  $X$ , is taken from a  $n(0, \sigma^2)$  population.
- Find an unbiased estimator of  $\sigma^2$ .
  - Find the MLE of  $\sigma$ .
  - Discuss how the method of moments estimator of  $\sigma$  might be found.
- 7.9 Let  $X_1, \dots, X_n$  be iid with pdf

$$f(x|\theta) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta, \quad \theta > 0.$$

Estimate  $\theta$  using both the method of moments and maximum likelihood. Calculate the means and variances of the two estimators. Which one should be preferred and why?

- 7.10 The independent random variables  $X_1, \dots, X_n$  have the common distribution

$$P(X_i \leq x|\alpha, \beta) = \begin{cases} 0 & \text{if } x < 0 \\ (x/\beta)^\alpha & \text{if } 0 \leq x \leq \beta, \\ 1 & \text{if } x > \beta \end{cases}$$

where the parameters  $\alpha$  and  $\beta$  are positive.

- Find a two-dimensional sufficient statistic for  $(\alpha, \beta)$ .
- Find the MLEs of  $\alpha$  and  $\beta$ .
- The length (in millimeters) of cuckoos' eggs found in hedge sparrow nests can be modeled with this distribution. For the data

22.0, 23.9, 20.9, 23.8, 25.0, 24.0, 21.7, 23.8, 22.8, 23.1, 23.1, 23.5, 23.0, 23.0,

find the maximum likelihood estimates of  $\alpha$  and  $\beta$ .

- 7.11** Let  $X_1, \dots, X_n$  be iid with pdf

$$f(x|\theta) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, \quad 0 < \theta < \infty.$$

- a. Find the MLE of  $\theta$ , and show that its variance  $\rightarrow 0$  as  $n \rightarrow \infty$ .  
 b. Find the method of moments estimator of  $\theta$ .

- 7.12** Let  $X_1, \dots, X_n$  be a random sample from a population with pmf

$$P_\theta(X = x) = \theta^x(1-\theta)^{1-x}, \quad x = 0 \text{ or } 1, \quad 0 \leq \theta \leq \frac{1}{2}.$$

- a. Find the method of moments estimator and MLE of  $\theta$ .  
 b. Find the mean squared errors of each of the estimators.  
 c. Which estimator is preferred? Justify your choice.

- 7.13** Let  $X_1, \dots, X_n$  be a sample from a population with double exponential pdf

$$f(x|\theta) = \frac{1}{2}e^{-|x-\theta|}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

Find the MLE of  $\theta$ . [Hint: Consider the case of even  $n$  separate from that of odd  $n$ , and find the MLE in terms of the order statistics. A complete treatment of this problem is given in Norton (1984).]

- 7.14** Let  $X$  and  $Y$  be independent exponential random variables, with

$$f(x|\lambda) = \frac{1}{\lambda}e^{-x/\lambda}, \quad x > 0, \quad f(y|\mu) = \frac{1}{\mu}e^{-y/\mu}, \quad y > 0.$$

We observe  $Z$  and  $W$  with

$$Z = \min(X, Y) \quad \text{and} \quad W = \begin{cases} 1 & \text{if } Z = X \\ 0 & \text{if } Z = Y \end{cases}.$$

In Exercise 4.26 the joint distribution of  $Z$  and  $W$  was obtained. Now assume that  $(Z_i, W_i), i = 1, \dots, n$ , are  $n$  iid observations. Find the MLEs of  $\lambda$  and  $\mu$ .

- 7.15** Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be iid bivariate normal random variables (pairs) where all five parameters are unknown. Derive the MLEs of the unknown parameters. (One attack is to write the joint pdf as the product of a conditional and a marginal, that is, write

$$f(x, y|\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) = f(y|x, \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) \times f(x|\mu_X, \sigma_X^2),$$

and argue that the MLEs for  $\mu_X$  and  $\sigma_X^2$  are given by  $\bar{x}$  and  $\hat{\sigma}_X^2$ . Then, turn things around to get the MLEs for  $\mu_Y$  and  $\sigma_Y^2$ . Finally, work with the “partially maximized” likelihood function  $L(\bar{x}, \bar{y}, \hat{\sigma}_X^2, \hat{\sigma}_Y^2, \rho | \mathbf{x}, \mathbf{y})$  to get the MLE for  $\rho$ . As might be guessed, this is a difficult problem.)

- 7.16** For the bivariate normal problem of Exercise 7.15, show that the MLEs for  $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$  are the same as the method of moments estimators. That is, show that the MLEs can be obtained by equating parameters to analogous sample quantities:  $\hat{\mu}_X = \bar{x}, \hat{\mu}_Y = \bar{y}, \hat{\sigma}_X^2 = \frac{1}{n} \sum (x_i - \bar{x})^2, \hat{\sigma}_Y^2 = \frac{1}{n} \sum (y_i - \bar{y})^2, \hat{\rho} = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) / (\hat{\sigma}_X \hat{\sigma}_Y)$ .

- 7.17** Again consider a bivariate normal problem, but now we assume that the means and variances are known. That is, let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be iid bivariate normal random variables with  $\mu_X = 0, \mu_Y = 0, \sigma_X^2 = 1, \sigma_Y^2 = 1$ , and we want to estimate  $\rho$ . Find the MLE of  $\rho$ . (Note that this is a more difficult problem than finding the MLE of  $\rho$  when all parameters are unknown. The additional knowledge here results in additional constraints on the solution, and makes things more involved.)
- 7.18** In a large population of married men (none of whom have ever been widowed), a fraction,  $p$ , have been divorced at least once. The following procedure is followed:

A box contains 100 envelopes.  $100x$  of these contain the question:

“Have you ever been divorced?”

The other  $100(1 - x)$  contain the question:

“Is this your first marriage?”

(Note that if a person would answer “yes” to the first question he would necessarily answer “no” to the second. We assume that everyone answers truthfully.)  $N$  married men were selected at random from the population. Each of them in succession was asked to pick an envelope from the box, read its contents, and return it to the box. Each then answered, aloud, “yes” or “no” to the question he read. (Note that only the subject knows which question he read.) Let  $Y$  denote the total number of men who answered “yes.” Assume that  $x$  is known,  $0 < x < 1$ .

- Find the distribution of  $Y$ .
- Find the method of moments estimator of  $p$ .
- Find the mean and variance of the estimator of part (b), and comment on the value of  $x$  that should be used.
- Find the MLE of  $p$ . Discuss any advantages or disadvantages it has over the method of moments estimator.

(This exercise describes a sampling procedure called *randomized response*, a technique that is useful in gathering information about sensitive topics. For more information about this topic, see Campbell and Joiner (1973) or, for a more technical discussion, Moors (1981).)

- 7.19** A communications device transmits sequences of binary digits, 0 or 1. A transmission error occurs if a digit is sent as a zero but received as a one, or vice versa. Assume that the probability of error of each digit transmitted is  $p$ , and all transmission errors are independent (that is, the probability of an error occurring at a given digit is independent of the outcomes of all other digits). To ensure veracity of a message, the same sequence of  $n$  binary digits is transmitted twice. The receiver notes the number of digits that differ in the two received messages. For example, with  $n = 8$  the two received messages might be 00110000 and 00100001, so  $X = 2$ .
- Find the probability that a given digit differs in the two received messages, and the pmf of  $X$ .
  - Find the equation for determining the MLE of  $p$ , simplifying as much as possible. Is the MLE unique?
  - Estimate  $p$  for the example given above.
  - What would be the interpretation if  $X$  were substantially greater than  $n/2$ ?
- 7.20** Suppose that the random variables  $Y_1, \dots, Y_n$  satisfy

$$Y_i = \beta x_i + \epsilon_i, \quad i = 1, \dots, n,$$

where  $x_1, \dots, x_n$  are fixed constants, and  $\epsilon_1, \dots, \epsilon_n$  are iid  $n(0, \sigma^2)$ ,  $\sigma^2$  unknown.

- Find a two-dimensional sufficient statistic for  $(\beta, \sigma^2)$ .

- b. Find the MLE of  $\beta$ , and show that it is an unbiased estimator of  $\beta$ .  
c. Find the distribution of the MLE of  $\beta$ .
- 7.21** For  $Y_1, \dots, Y_n$  as defined in Exercise 7.20  
a. Show that  $\sum Y_i / \sum x_i$  is an unbiased estimator of  $\beta$ .  
b. Calculate the exact variance of  $\sum Y_i / \sum x_i$  and compare it to the variance of the MLE.
- 7.22** Again, let  $Y_1, \dots, Y_n$  be as defined in Exercise 7.20.  
a. Show that  $[\sum(Y_i/x_i)]/n$  is also an unbiased estimator of  $\beta$ .  
b. Calculate the exact variance of  $[\sum(Y_i/x_i)]/n$  and compare it to the variances of the estimators in the previous two exercises.
- 7.23** This exercise will prove the assertions in Example 7.2.10, and more. Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population, and suppose that the prior distribution on  $\theta$  is  $n(\mu, \tau^2)$ . Here we assume that  $\sigma^2$ ,  $\mu$ , and  $\tau^2$  are all known.  
a. Find the joint pdf of  $\bar{X}$  and  $\theta$ .  
b. Show that  $m(\bar{x}|\sigma^2, \mu, \tau^2)$ , the marginal distribution of  $\bar{X}$ , is  $n(\mu, (\sigma^2/n) + \tau^2)$ .  
c. Show that  $\pi(\theta|\bar{x}, \sigma^2, \mu, \tau^2)$ , the posterior distribution of  $\theta$ , is normal with mean and variance given by

$$\begin{aligned} E(\theta|\bar{x}) &= \frac{\tau^2}{\tau^2 + (\sigma^2/n)} \bar{x} + \frac{\sigma^2/n}{(\sigma^2/n) + \tau^2} \mu, \\ \text{Var}(\theta|\bar{x}) &= \frac{(\sigma^2/n)\tau^2}{(\sigma^2/n) + \tau^2}. \end{aligned}$$

- 7.24** If  $S^2$  is the sample variance based on a sample of size  $n$  from a normal population, we know that  $(n-1)S^2/\sigma^2$  has a  $\chi_{n-1}^2$  distribution. The conjugate prior for  $\sigma^2$  is the *inverted gamma* pdf,  $IG(\alpha, \beta)$ , given by

$$\pi(\sigma^2) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \frac{1}{(\sigma^2)^{\alpha+1}} e^{-1/(\beta\sigma^2)}, \quad 0 < \sigma^2 < \infty,$$

where  $\alpha$  and  $\beta$  are positive constants. Show that the posterior distribution of  $\sigma^2$  is  $IG(\alpha + \frac{n-1}{2}, [\frac{(n-1)S^2}{2} + \frac{1}{\beta}]^{-1})$ . Find the mean of this distribution, the Bayes estimator of  $\sigma^2$ .

- 7.25** Let  $X_1, \dots, X_n$  be iid  $Poisson(\lambda)$ , and let  $\lambda$  have a  $gamma(\alpha, \beta)$  distribution, the conjugate family for the Poisson.  
a. Find the posterior distribution of  $\lambda$ .  
b. Calculate the posterior mean and variance.
- 7.26** We examine a generalization of the hierarchical (Bayes) model considered in Example 7.2.10 and Exercise 7.23. Suppose that we observe  $X_1, \dots, X_n$ , where

$$\begin{aligned} X_i | \theta_i &\sim n(\theta_i, \sigma^2), \quad i = 1, \dots, n, \quad \text{independent,} \\ \theta_i &\sim n(\mu, \tau^2), \quad i = 1, \dots, n, \quad \text{independent.} \end{aligned}$$

- a. Show that the marginal distribution of  $X_i$  is  $n(\mu, \sigma^2 + \tau^2)$ , and that, marginally,  $X_1, \dots, X_n$  are iid. (*Empirical Bayes analysis* would use the marginal distribution of the  $X_i$ 's to estimate the prior parameters  $\mu$  and  $\tau^2$ . See Casella (1985) for an introduction to empirical Bayes methods.)  
b. Show, in general, that if

$$\begin{aligned} X_i | \theta_i &\sim f(x | \theta_i), & i = 1, \dots, n, & \text{independent,} \\ \theta_i &\sim \pi(\theta | \tau), & i = 1, \dots, n, & \text{independent,} \end{aligned}$$

then marginally,  $X_1, \dots, X_n$  are iid.

- 7.27 In Example 7.2.10 we saw that the normal distribution is its own conjugate family. It is sometimes the case, however, that a conjugate prior does not accurately reflect prior knowledge, and a different prior is sought. Let  $X_1, \dots, X_n$  be iid  $n(\theta, \sigma^2)$ , and let  $\theta$  have a double exponential distribution, that is,  $\pi(\theta) = e^{-|\theta|/a}/(2a)$ ,  $a$  known. Find the mean of the posterior distribution of  $\theta$ .
- 7.28 Let  $X_1, \dots, X_n$  be iid  $f(x - \theta)$  and consider the group of transformations defined by  $\mathcal{G} = \{g_a(\mathbf{x}): -\infty < a < \infty\}$ , where  $g_a(x_1, \dots, x_n) = (x_1 + a, \dots, x_n + a)$ . Show that, as long as  $E_0 X_1 = 0$ , the estimator  $W(X_1, \dots, X_n) = \bar{X}$  is both unbiased and invariant for estimating  $\theta$ .
- 7.29 Suppose we have a random sample  $X_1, \dots, X_n$  from  $\frac{1}{\sigma} f((x - \theta)/\sigma)$ , a location-scale pdf. We want to estimate  $\theta$ , and we have two groups of transformations under consideration:

$$\mathcal{G}_1 = \{g_{a,c}(\mathbf{x}): -\infty < a < \infty, c > 0\},$$

where  $g_{a,c}(x_1, \dots, x_n) = (cx_1 + a, \dots, cx_n + a)$ , and

$$\mathcal{G}_2 = \{g_a(\mathbf{x}): -\infty < a < \infty\},$$

where  $g_a(x_1, \dots, x_n) = (x_1 + a, \dots, x_n + a)$ .

a. Show that estimators of the form

$$W(x_1, \dots, x_n) = \bar{x} + k,$$

where  $k$  is a nonzero constant, are invariant with respect to the group  $\mathcal{G}_2$  but are not invariant with respect to the group  $\mathcal{G}_1$ .

b. For each group, under what conditions are invariant estimators unbiased for estimating  $\theta$ ?

- 7.30 Again, suppose we have a random sample  $X_1, \dots, X_n$  from  $\frac{1}{\sigma} f((x - \theta)/\sigma)$ , a location-scale pdf, but we are now interested in estimating  $\sigma^2$ . We can consider three groups of transformations:

$$\mathcal{G}_1 = \{g_{a,c}(\mathbf{x}): -\infty < a < \infty, c > 0\},$$

where  $g_{a,c}(x_1, \dots, x_n) = (cx_1 + a, \dots, cx_n + a)$ ,

$$\mathcal{G}_2 = \{g_a(\mathbf{x}): -\infty < a < \infty\},$$

where  $g_a(x_1, \dots, x_n) = (x_1 + a, \dots, x_n + a)$ , and

$$\mathcal{G}_3 = \{g_c(\mathbf{x}): c > 0\},$$

where  $g_c(x_1, \dots, x_n) = (cx_1, \dots, cx_n)$ .

- a. Show that estimators of  $\sigma^2$  of the form  $kS^2$ , where  $k$  is a positive constant and  $S^2$  is the sample variance, are invariant with respect to all three groups.
- b. Show that the larger class of estimators of  $\sigma^2$  of the form (Brewster and Zidek, 1974)

$$W(X_1, \dots, X_n) = \phi\left(\frac{\bar{X}}{S}\right) S^2,$$

where  $\phi(x)$  is a function, are invariant with respect to  $\mathcal{G}_3$ , but not with respect to either  $\mathcal{G}_1$  or  $\mathcal{G}_2$ , unless  $\phi(x)$  is a constant.

- 7.31 In Example 7.3.2 the MSE of the Bayes estimator,  $\hat{p}_B$ , of a success probability was calculated (the estimator was derived in Example 7.2.9). Show that the choice  $\alpha = \beta = \sqrt{n/4}$  yields a constant MSE for  $\hat{p}_B$ .
- 7.32 Let  $X_1, \dots, X_n$  be a random sample from a binomial( $n, p$ ). We want to find invariant point estimators of  $p$  using the group described in Example 6.3.1.
- Find the class of estimators that are invariant with respect to this group.
  - Within the class of Bayes estimators of Example 7.2.9, find the estimators that are invariant with respect to this group.
  - From the invariant Bayes estimators of part (b), find the one with the smallest MSE.
- 7.33 The *Pitman Estimator of Location* (see Lehmann (1983) or the original paper by Pitman (1939)) is given by

$$\delta_P(\mathbf{X}) = \frac{\int_{-\infty}^{\infty} t \prod_{i=1}^n f(x_i - t) dt}{\int_{-\infty}^{\infty} \prod_{i=1}^n f(x_i - t) dt},$$

where we observe a random sample  $X_1, \dots, X_n$  from  $f(x - \theta)$ . Pitman showed that this estimator is the location-invariant estimator with smallest mean squared error (that is, it minimizes (7.3.3)). The goals of this exercise are more modest.

- Show that  $\delta_P(\mathbf{X})$  is invariant with respect to the location group, that is, it satisfies Definition 7.2.3 using the group given in Example 7.2.11.
  - Show that if  $f(x - \theta)$  is  $n(\theta, 1)$ , then  $\delta_P(\mathbf{X}) = \bar{X}$ .
  - Show that if  $f(x - \theta)$  is uniform( $\theta - \frac{1}{2}, \theta + \frac{1}{2}$ ), then  $\delta_P(\mathbf{X}) = \frac{1}{2}(X_{(1)} + X_{(n)})$ .
- 7.34 The *Pitman Estimator of Scale* is given by

$$\delta_P^r(\mathbf{X}) = \frac{\int_0^{\infty} t^{n+r-1} \prod_{i=1}^n f(tx_i) dt}{\int_0^{\infty} t^{n+2r-1} \prod_{i=1}^n f(tx_i) dt},$$

where we observe a random sample  $X_1, \dots, X_n$  from  $\frac{1}{\sigma} f(x/\sigma)$ . Pitman showed that this estimator is the scale-invariant estimator of  $\sigma^r$  with smallest scaled mean squared error (that is, it minimizes  $E(\delta - \sigma^r)^2 / \sigma^{2r}$ ).

- Show that  $\delta_P^r(\mathbf{X})$  is invariant with respect to the scale group, that is, it satisfies

$$\delta_P^r(cx_1, \dots, cx_n) = c^r \delta_P^r(x_1, \dots, x_n),$$

for any constant  $c > 0$ .

- Find the Pitman scale-invariant estimator for  $\sigma^2$  if  $X_1, \dots, X_n$  are iid  $n(0, \sigma^2)$ .
- Find the Pitman scale-invariant estimator for  $\beta$  if  $X_1, \dots, X_n$  are iid exponential( $\beta$ ).
- Find the Pitman scale-invariant estimator for  $\theta$  if  $X_1, \dots, X_n$  are iid uniform( $0, \theta$ ).

- 7.35 Let  $X_1, \dots, X_n$  be a random sample from a population with pdf

$$f(x|\theta) = \frac{1}{2\theta}, \quad -\theta < x < \theta, \quad \theta > 0.$$

Find, if one exists, a best unbiased estimator of  $\theta$ .

- 7.36 For each of the following distributions, let  $X_1, \dots, X_n$  be a random sample. Is there a function of  $\theta$ , say  $g(\theta)$ , for which there exists an unbiased estimator whose variance attains the Cramér–Rao Lower Bound? Is so, find it. If not, show why not.

a.  $f(x|\theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad \theta > 0.$

b.  $f(x|\theta) = \frac{\log(\theta)}{\theta-1} \theta^x, \quad 0 < x < 1, \quad \theta > 1.$

- 7.37 Assume whatever is needed about  $f(x|\theta)$  to show that

$$E_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right) = -E_\theta \left( \frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right).$$

- 7.38 Let  $X_1, \dots, X_n$  be iid Bernoulli( $p$ ). Show that the variance of  $\bar{X}$  attains the Cramér–Rao Lower Bound, and hence  $\bar{X}$  is the best unbiased estimator of  $p$ .

- 7.39 Let  $X_1, \dots, X_n$  be a random sample from a population with mean  $\mu$  and variance  $\sigma^2$ .

a. Show that the estimator  $\sum_{i=1}^n a_i X_i$  is an unbiased estimator of  $\mu$  if  $\sum_{i=1}^n a_i = 1$ .

b. Among all unbiased estimators of this form (called *linear unbiased estimators*) find the one with minimum variance, and calculate the variance.

- 7.40 Let  $W_1, \dots, W_k$  be unbiased estimators of a parameter  $\theta$  with  $\text{Var } W_i = \sigma_i^2$  and  $\text{Cov}(W_i, W_j) = 0$  if  $i \neq j$ .

a. Show that, of all estimators of the form  $\sum a_i W_i$  where the  $a_i$ 's are constant and  $E_\theta(\sum a_i W_i) = \theta$ , the estimator

$$W^* = \frac{\sum W_i / \sigma_i^2}{\sum (1/\sigma_i^2)}$$

has minimum variance.

b. Show that

$$\text{Var } W^* = \frac{1}{\sum (1/\sigma_i^2)}.$$

- 7.41 Let  $X_1, \dots, X_n$  be iid  $n(\theta, 1)$ . Show that the best unbiased estimator of  $\theta^2$  is  $\bar{X}^2 - (1/n)$ . Calculate its variance (use Stein's Identity from Section 4.7), and show that it is greater than the Cramér–Rao Lower Bound, which is  $4\theta^2/n$ .

- 7.42 Let  $X_1, X_2$ , and  $X_3$  be a random sample of size three from a uniform( $\theta, 2\theta$ ) distribution, where  $\theta > 0$ .

a. Find the method of moments estimator of  $\theta$ .

b. Find the MLE,  $\hat{\theta}$ , and find a constant  $k$  such that  $E_\theta(k\hat{\theta}) = \theta$ .

c. Which of the two estimators can be improved by using sufficiency? How?

d. Find the method of moments estimate and the MLE of  $\theta$  based on the data

1.29, .86, 1.33,

three observations of average berry sizes (in centimeters) of wine grapes.

- 7.43 Suppose that when measuring the radius of a circle, an error is made that has a  $n(0, \sigma^2)$  distribution. If  $n$  independent measurements are made, find an unbiased estimator of the area of the circle. Is it best unbiased?

- 7.44 Suppose that  $X_i, i = 1, \dots, n$ , are iid Bernoulli( $p$ ).

a. Show that the variance of the MLE of  $p$  attains the Cramér–Rao Lower Bound.

- b. For  $n \geq 4$ , show that the product  $X_1 X_2 X_3 X_4$  is an unbiased estimator of  $p^4$ , and use this fact to find the best unbiased estimator of  $p^4$ .

**7.45** Let  $X_1, \dots, X_n$  be iid exponential( $\lambda$ ).

- Find an unbiased estimator of  $\lambda$  based only on  $Y = \min\{X_1, \dots, X_n\}$ .
- Find a better estimator than the one in part (a). Prove that it is better.
- The following data are high-stress failure times (in hours) of Kevlar/epoxy spherical vessels used in a sustained pressure environment on the space shuttle:

$$50.1, 70.1, 137.0, 166.9, 170.5, 152.8, 80.5, 123.5, 112.6, 148.5, 160.0, 125.4.$$

Failure times are often modeled with the exponential distribution. Estimate the mean failure time using the estimators from parts (a) and (b).

**7.46** Let  $X_1, \dots, X_n$  be iid  $n(\theta, \theta^2)$ ,  $\theta > 0$ . For this model both  $\bar{X}$  and  $cS$  are unbiased estimators of  $\theta$ , where  $c = \frac{\sqrt{n-1}\Gamma((n-1)/2)}{\sqrt{2}\Gamma(n/2)}$ .

- Prove that for any number  $a$  the estimator  $a\bar{X} + (1-a)(cS)$  is an unbiased estimator of  $\theta$ .
- Find the value of  $a$  that produces the estimator with minimum variance.
- Show that  $(\bar{X}, S^2)$  is a sufficient statistic for  $\theta$  but it is not a complete sufficient statistic.

**7.47** Gleser and Healy (1976) give a detailed treatment of the estimation problem in the  $n(\theta, a\theta^2)$  family, where  $a$  is a known constant (of which Exercise 7.46 is a special case). We explore a small part of their results here. Again let  $X_1, \dots, X_n$  be iid  $n(\theta, \theta^2)$ ,  $\theta > 0$ , and let  $\bar{X}$  and  $cS$  be as in Exercise 7.46. Define the class of estimators

$$\mathcal{T} = \{T: T = a_1 \bar{X} + a_2(cS)\},$$

where we do not assume that  $a_1 + a_2 = 1$ .

- Find the estimator  $T \in \mathcal{T}$  that minimizes  $E_\theta(\theta - T)^2$ , call it  $T^*$ .
- Show that the MSE of  $T^*$  is smaller than the MSE of the estimator derived in Exercise 7.46b.
- Show that the MSE of  $T^{*+} = \max\{0, T^*\}$  is smaller than the MSE of  $T^*$ .
- Would  $\theta$  be classified as a location parameter or a scale parameter? Explain.

**7.48** Let  $X_1, \dots, X_n$  be iid Poisson( $\lambda$ ), and let  $\bar{X}$  and  $S^2$  denote the sample mean and variance, respectively. We now complete Example 7.3.4 in a different way. There we used the Cramér–Rao Bound, now we use completeness.

- Prove that  $\bar{X}$  is the best unbiased estimator of  $\lambda$ , without using the Cramér–Rao Theorem.
- Prove the rather remarkable identity  $E(S^2|\bar{X}) = \bar{X}$ , and use it to explicitly demonstrate that  $\text{Var } S^2 > \text{Var } \bar{X}$ .
- Using completeness, can a general theorem be formulated for which the identity in part (b) is a special case?

**7.49** Finish some of the details left out of the proof of Theorem 7.3.4. Suppose  $W$  is an unbiased estimator of  $\tau(\theta)$ , and  $U$  is an unbiased estimator of zero. Show that if, for some  $\theta = \theta_0$ ,  $\text{Cov}_{\theta_0}(W, U) \neq 0$ , then  $W$  cannot be the best unbiased estimator of  $\tau(\theta)$ .

**7.50** Let  $X_1, \dots, X_n$  be iid  $n(0, \sigma^2)$ , and define two estimators  $U$  and  $V$  by

$$U = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1},$$

$$V = \frac{(n-2) \sum_{i=1}^n X_i^2 + (n\bar{X})^2}{n(n-1)}.$$

- a. Show that  $EU = EV = \sigma^2$ .
- b. Show that  $\text{Var } U = \text{Var } V$ .
- c. Consider all estimators of the form  $T_a(U, V) = aU + (1-a)V$ , where  $a$  is a constant. Show that  $E[T_a(U, V)] = \sigma^2$ . Find the value of  $a$  that minimizes  $\text{Var}[T_a(U, V)]$ , and show that this variance is smaller than that of  $U$  or  $V$ .
- d. Show that the minimizing value of  $a$  produces the best unbiased estimator of  $\sigma^2$ . [Hint: It will help to write  $V$  in terms of  $\sum(X_i - \bar{X})^2$  and  $\bar{X}^2$ , which are independent.]
- 7.51 Let  $X_1, \dots, X_n$  be iid with cdf  $F(x|\theta)$ , and let  $Y_1, \dots, Y_m$  be iid with cdf  $G(y|\mu)$ , where the  $X$ s and  $Y$ s are independent. Suppose that the statistics  $T = T(X_1, \dots, X_n)$  and  $W = W(Y_1, \dots, Y_m)$  are complete sufficient statistics for  $\theta$  and  $\mu$ , respectively, with

$$E_\theta T = \theta, \quad E_\theta T^2 < \infty,$$

$$E_\mu W = \mu, \quad E_\mu W^2 < \infty,$$

for all  $\theta$  and  $\mu$ . Prove or disprove that there exists a best unbiased estimator of  $\theta\mu$ .

- 7.52 For each of the following pdfs, let  $X_1, \dots, X_n$  be a sample from that distribution. In each case, find the best unbiased estimator of  $\theta^r$ . (See Guenther (1978) for a complete discussion of this problem.)
- a.  $f(x|\theta) = \frac{1}{\theta}, \quad 0 < x < \theta$ .
  - b.  $f(x|\theta) = e^{-(x-\theta)}, \quad x > \theta$ .
  - c.  $f(x|\theta) = \frac{e^{-x}}{e^{-\theta}-e^{-b}}, \quad \theta < x < b, \quad b \text{ known}$ .
- 7.53 Prove the assertion made in the text preceding Example 7.3.9: If  $T$  is a complete sufficient statistic for a parameter  $\theta$ , and  $h(X_1, \dots, X_n)$  is any unbiased estimator of  $\tau(\theta)$ , then  $\phi(T) = E(h(X_1, \dots, X_n)|T)$  is the best unbiased estimator of  $\tau(\theta)$ .
- 7.54 Let  $X_1, \dots, X_{n+1}$  be iid Bernoulli( $p$ ), and define the function  $h(p)$  by

$$h(p) = P\left(\sum_{i=1}^n X_i > X_{n+1} \mid p\right),$$

the probability that the first  $n$  observations exceed the  $(n+1)$ st.

- a. Show that

$$T(X_1, \dots, X_{n+1}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n X_i > X_{n+1} \\ 0 & \text{otherwise} \end{cases}$$

is an unbiased estimator of  $h(p)$ .

- b. Find the best unbiased estimator of  $h(p)$ .

- 7.55 Suppose that  $X_1, \dots, X_n$  are iid Poisson( $\lambda$ ). Find the best unbiased estimator of
- a.  $e^{-\lambda}$ , the probability that  $X = 0$ .
  - b.  $\lambda e^{-\lambda}$ , the probability that  $X = 1$ .
  - c. A preliminary test of a possible carcinogenic compound can be performed by measuring the mutation rate of microorganisms exposed to the compound. An experimenter

places the compound in 15 petri dishes and records the following number of mutant colonies:

$$10, 7, 8, 13, 8, 9, 5, 7, 6, 8, 3, 6, 6, 3, 5.$$

Estimate  $e^{-\lambda}$ , the probability that no mutant colonies emerge, and  $\lambda e^{-\lambda}$ , the probability that one mutant colony will emerge. Calculate both the best unbiased estimator and the MLE.

- 7.56 Let  $X$  be an observation from the pdf

$$f(x|\theta) = \left(\frac{\theta}{2}\right)^{|x|} (1-\theta)^{1-|x|}, \quad x = -1, 0, 1; \quad 0 \leq \theta \leq 1.$$

- a. Find the MLE of  $\theta$ .
- b. Define the estimator  $T(X)$  by

$$T(X) = \begin{cases} 2 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}.$$

Show that  $T(X)$  is an unbiased estimator of  $\theta$ .

c. Find a better estimator than  $T(X)$  and prove that it is better.

- 7.57 Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$ . Find the best unbiased estimator of  $\sigma^p$ , where  $p$  is a known positive constant, not necessarily an integer.

- 7.58 Let  $X_1, \dots, X_n$  be iid gamma( $\alpha, \beta$ ) with  $\alpha$  known. Find the best unbiased estimator of  $1/\beta$ .

- 7.59 The *jackknife* is a general technique for reducing bias in an estimator (Quenouille, 1956). A one-step jackknife estimator is defined as follows. Let  $X_1, \dots, X_n$  be a random sample, and let  $T_n = T_n(X_1, \dots, X_n)$  be some estimator of a parameter  $\theta$ . In order to "jackknife"  $T_n$  we calculate the  $n$  statistics  $T_n^{(i)}$ ,  $i = 1, \dots, n$ , where  $T_n^{(i)}$  is calculated just as  $T_n$  except that  $X_i$  is removed from the sample. The jackknife estimator of  $\theta$ , denoted by  $JK(T_n)$ , is given by

$$JK(T_n) = nT_n - \frac{n-1}{n} \sum_{i=1}^n T_n^{(i)}.$$

(In general,  $JK(T_n)$  will have a smaller bias than  $T_n$ . See Miller (1974) for a good review of the properties of the jackknife.)

Now, to be specific, let  $X_1, \dots, X_n$  be iid Bernoulli( $\theta$ ). The object is to estimate  $\theta^2$ .

- a. Show that the MLE of  $\theta^2$ ,  $(\sum_{i=1}^n X_i/n)^2$ , is a biased estimator of  $\theta^2$ .
- b. Derive the one-step jackknife estimator based on the MLE.
- c. Show that the one-step jackknife estimator is an unbiased estimator of  $\theta^2$ . (In general, jackknifing only reduces bias. In this special case, however, it removes it entirely.)
- d. Is this jackknife estimator the best unbiased estimator of  $\theta^2$ ? If so, prove it. If not, find the best unbiased estimator.

- 7.60 Prove Theorem 7.3.7.

- 7.61 Suppose  $X_1, \dots, X_n$  are iid and satisfy

$$E_\theta X_i = \theta + b, \quad \text{Var}_\theta X_i = \sigma^2 < \infty, \quad b \text{ known and } b \neq 0.$$

Show that  $\bar{X}$  is not a consistent estimator of  $\theta$ . Construct an unbiased estimator of  $\theta$  that is consistent.

- 7.62 A random sample  $X_1, \dots, X_n$  is drawn from a population with pdf

$$f(x|\theta) = \frac{1}{2}(1 + \theta x), \quad -1 < x < 1, \quad -1 < \theta < 1.$$

Find a consistent estimator of  $\theta$  and show that it is consistent.

- 7.63 A random sample  $X_1, \dots, X_n$  is drawn from a population that is  $n(\theta, \theta)$ , where  $\theta > 0$ .

- a. Show that the MLE of  $\theta, \hat{\theta}$ , is a root of the quadratic equation  $\theta^2 + \theta - W = 0$ , where

$W = (1/n) \sum_{i=1}^n X_i^2$ , and determine which root equals the MLE.

- b. Find the approximate variance of  $\hat{\theta}$  using the techniques of Section 7.4.1.

- c. Find the approximate variance of  $\hat{\theta}$  using a Taylor series approximation.

- 7.64 A variation of the model in Exercise 7.20 is to let the random variables  $Y_1, \dots, Y_n$  satisfy

$$Y_i = \beta X_i + \epsilon_i, \quad i = 1, \dots, n,$$

where  $X_1, \dots, X_n$  are independent  $n(\mu, \tau^2)$  random variables, and  $\epsilon_1, \dots, \epsilon_n$  are iid  $n(0, \sigma^2)$ , and the  $X$ s and  $\epsilon$ s are independent. Exact variance calculations become quite difficult, so we might resort to approximations. In terms of  $\mu, \tau^2$ , and  $\sigma^2$ , find approximate means and variances for

a.  $\sum X_i Y_i / \sum X_i^2$

b.  $\sum Y_i / \sum X_i$

c.  $\sum (Y_i / X_i) / n$

- 7.65 Continue the calculations of Example 7.4.1, where an expression for  $\text{Var}[\hat{p}(1 - \hat{p})]$  was sought.

- a. Using a Taylor series, find an approximation for  $\text{Var}[\hat{p}(1 - \hat{p})]$ .

- b. Show that the Taylor series approximation of part (a) still has problems, like those of the approximation in Example 7.4.1.

- c. Calculate the exact expression for  $\text{Var}[\hat{p}(1 - \hat{p})]$ . Is the reason for the failure of the approximations any clearer?

## Miscellanea

### Moment Estimators and MLEs

In general, method of moments estimators are not functions of sufficient statistics, hence, they can always be improved upon by conditioning on a sufficient statistic. In the case of exponential families, however, there can be a correspondence between a modified method of moments strategy and maximum likelihood estimation. This correspondence is discussed in detail by Davidson and Solomon (1974), who also relate some interesting history.

Suppose that we have a random sample  $\mathbf{X} = (X_1, \dots, X_n)$  from a pdf in the exponential family (see Theorem 5.2.5)

$$f(x|\theta) = h(x)c(\theta) \exp\left(\sum_{i=1}^k w_i(\theta)t_i(x)\right),$$

where the range of  $f(x|\theta)$  is independent of  $\theta$ . (Note that  $\theta$  may be a vector.) The likelihood function is of the form

$$L(\theta|x) = H(x)[c(\theta)]^n \exp\left(\sum_{i=1}^k w_i(\theta) \sum_{j=1}^n t_i(x_j)\right),$$

and a modified method of moments would estimate  $w_i(\theta)$ ,  $i = 1, \dots, k$ , by  $\hat{w}_i(\theta)$ , the solutions to the  $k$  equations

$$\sum_{j=1}^n t_i(x_j) = E_\theta\left(\sum_{j=1}^n t_i(X_j)\right), \quad i = 1, \dots, k.$$

Davidson and Solomon, extending work of Huzurbazar (1949), show that the estimators  $\hat{w}_i(\theta)$  are, in fact, the MLEs of  $w_i(\theta)$ . If we define  $\eta_i = w_i(\theta)$ ,  $i = 1, \dots, k$ , then the MLE of  $g(\eta_i)$  is equal to  $g(\hat{\eta}_i) = g(\hat{w}_i(\theta))$  for any one-to-one function  $g$ . Calculation of the above expectations may be simplified by using the facts (Lehmann, 1986, Section 2.7) that

$$E_\theta(t_i(X_j)) = \frac{\partial}{\partial w_i(\theta)} \log(c(\theta)), \quad i = 1, \dots, k, \quad j = 1, \dots, n;$$

$$\text{Cov}_\theta(t_i(X_j), t_{i'}(X_j)) = \frac{\partial^2}{\partial w_i(\theta) \partial w_{i'}(\theta)} \log(c(\theta)), \quad i, i' = 1, \dots, k, \quad j = 1, \dots, n.$$

### Unbiased Bayes Estimates

As was seen in Section 7.2.3, if a Bayesian calculation is done, the mean of the posterior distribution usually is taken as a point estimator. To be specific, if  $X$  has pdf  $f(x|\theta)$  with  $E_\theta(X) = \theta$ , and there is a prior distribution  $\pi(\theta)$ , then the posterior mean, a Bayesian point estimator of  $\theta$ , is given by

$$E(\theta|x) = \int \theta \pi(\theta|x) d\theta.$$

A question that could be asked is whether  $E(\theta|X)$  can be an unbiased estimator of  $\theta$ , and thus satisfy the equation

$$E_\theta[E(\theta|X)] = \int \left[ \int \theta \pi(\theta|x) d\theta \right] f(x|\theta) dx = \theta.$$

The answer is no. That is, posterior means are *never* unbiased estimators. If they were, then taking the expectation over the joint distribution of  $X$  and  $\theta$ , we could write

$$\begin{aligned} E[(X - \theta)^2] &= E[X^2 - 2X\theta + \theta^2] && \text{(expand the square)} \\ &= E(E((X^2 - 2X\theta + \theta^2)|\theta)) && \text{(iterate the expectation)} \\ &= E(E(X^2|\theta) - 2\theta^2 + \theta^2) && (E(X|\theta) = E_\theta X = \theta) \\ &= E(E(X^2|\theta) - \theta^2) \\ &= E(X^2) - E(\theta^2), && \text{(properties of expectations)} \end{aligned}$$

doing the conditioning one way, and conditioning on  $X$  we could similarly calculate

$$\begin{aligned} E[(X - \theta)^2] &= E(E[(X^2 - 2X\theta + \theta^2)|X]) \\ &= E(X^2 - 2X^2 + E(\theta^2|X)) && \left( \begin{array}{l} E(\theta|X) = X \\ \text{by assumption} \end{array} \right) \\ &= E(\theta^2) - E(X^2). \end{aligned}$$

Comparing the two calculations, we see that the only way that there is no contradiction is if  $E(X^2) = E(\theta^2)$ , which then implies that  $E(X - \theta)^2 = 0$ , so  $X = \theta$ . This occurs only if  $P(X = \theta) = 1$ , an impossibility, so we have argued to a contradiction. Thus, either  $E(X|\theta) \neq \theta$  or  $E(\theta|X) \neq X$ , showing that posterior means cannot be unbiased estimators. Notice that we have implicitly made the assumption that  $E(X^2) < \infty$ , but, in fact, this result holds under more general conditions. Bickel and Mallows (1988) have a more thorough development of this topic. At a more advanced level, this connection is characterized by Noorbaloochi and Meeden (1983).

### The Lehmann–Scheffé Theorem

The Lehmann–Scheffé Theorem represents a major achievement in mathematical statistics, tying together sufficiency, completeness, and uniqueness. The development in the text is somewhat complementary to the Lehmann–Scheffé Theorem, and thus we never stated it in its classical form (which is similar to Theorem 7.3.5). In fact, the Lehmann–Scheffé Theorem is contained in Theorems 7.3.3 and 7.3.5.

*Theorem (Lehmann–Scheffé):* Unbiased estimators based on complete sufficient statistics are unique.

*Proof:* Suppose  $T$  is a complete sufficient statistic, and  $\phi(T)$  is an estimator with  $E_\theta \phi(T) = \tau(\theta)$ . From Theorem 7.3.5 we know that  $\phi(T)$  is the best unbiased estimator of  $\tau(\theta)$ , and from Theorem 7.3.3, best unbiased estimators are unique.  $\square$

This theorem can also be proved without Theorem 7.3.3, using just the consequences of completeness, and provides a slightly different route to Theorem 7.3.5.

# 8 Hypothesis Testing

*"We all learn by experience, and your lesson this time is that you should never lose sight of the alternative."*

**Sherlock Holmes**

*The Adventure of Black Peter*

## 8.1 Introduction

In Chapter 7 we studied a method of inference called point estimation. Now we move to another inference method, hypothesis testing. Reflecting the need both to find and to evaluate hypothesis tests, this chapter is divided into two parts, as was Chapter 7. We begin with the definition of a statistical hypothesis.

**DEFINITION 8.1.1:** A *hypothesis* is a statement about a population parameter.

The definition of a hypothesis is rather general, but the important point is that a hypothesis makes a statement about the population. The goal of a hypothesis test is to decide, based on a sample from the population, which of two complementary hypotheses is true.

**DEFINITION 8.1.2:** The two complementary hypotheses in a hypothesis testing problem are called the *null hypothesis* and the *alternative hypothesis*. They are denoted by  $H_0$  and  $H_1$ , respectively.

If  $\theta$  denotes a population parameter, the general format of the null and alternative hypotheses is  $H_0: \theta \in \Theta_0$  and  $H_1: \theta \in \Theta_0^c$  where  $\Theta_0$  is some subset of the parameter space and  $\Theta_0^c$  is its complement. For example, if  $\theta$  denotes the average change in a patient's blood pressure after taking a drug, an experimenter might be interested in testing  $H_0: \theta = 0$  versus  $H_1: \theta \neq 0$ . The null hypothesis states that, on the average, the drug has no effect on blood pressure and the alternative hypothesis states that there is some effect. This common situation, in which  $H_0$  states that a treatment has no effect, has led to the term "null" hypothesis. As another example, a consumer might be interested in the proportion of defective items produced by a supplier. If  $\theta$  denotes the proportion of defective items, the consumer might wish to test  $H_0: \theta \geq \theta_0$  versus  $H_1: \theta < \theta_0$ . The value  $\theta_0$  is the maximum acceptable proportion of defective items and  $H_0$  states that the proportion of defectives is unacceptably high. Problems in which the hypotheses concern the quality of a product are called acceptance sampling problems.

In a hypothesis testing problem, after observing the sample the experimenter must decide either to accept  $H_0$  as true or to reject  $H_0$  as false and decide  $H_1$  is true.

**DEFINITION 8.1.3:** A *hypothesis testing procedure* or *hypothesis test* is a rule that specifies:

- i. For which sample values the decision is made to accept  $H_0$  as true.
- ii. For which sample values  $H_0$  is rejected and  $H_1$  is accepted as true.

The subset of the sample space for which  $H_0$  will be rejected is called the *rejection region* or *critical region*. The complement of the rejection region is called the *acceptance region*.

On a philosophical level, some people worry about the distinction between “rejecting  $H_0$ ” and “accepting  $H_1$ .” In the first case, there is nothing implied about what state the experimenter *is* accepting, only that the state defined by  $H_0$  is being rejected. Similarly, a distinction can be made between “accepting  $H_0$ ” and “not rejecting  $H_0$ .” The first phrase implies that the experimenter is willing to assert the state of nature specified by  $H_0$ , while the second phrase implies that the experimenter really does not believe  $H_0$ , but does not have the evidence to reject it. For the most part, we will not be concerned with these issues. We view a hypothesis testing problem as a problem in which one of two actions is going to be taken—the actions being the assertion of  $H_0$  or  $H_1$ .

Typically, a hypothesis test is specified in terms of a *test statistic*  $W(X_1, \dots, X_n) = W(\mathbf{X})$ , a function of the sample. For example, a test might specify that  $H_0$  is to be rejected if  $\bar{X}$ , the sample mean, is greater than 3. In this case  $W(\mathbf{X}) = \bar{X}$  is the test statistic and the rejection region is  $\{(x_1, \dots, x_n) : \bar{x} > 3\}$ . In Section 8.2, methods of choosing test statistics and rejection regions are discussed. Criteria for evaluating tests are introduced in Section 8.3. As with point estimators, the methods of finding tests carry no guarantees; the tests they yield must be evaluated before their worth is established.

## 8.2 Methods of Finding Tests

We will detail four methods of finding test procedures, procedures that are useful in different situations, and take advantage of different aspects of a problem. We start with a very general method, one that is almost always applicable, and is also optimal in some cases.

### 8.2.1 Likelihood Ratio Tests

The likelihood ratio method of hypothesis testing is related to the maximum likelihood estimators discussed in Section 7.2.2, and likelihood ratio tests are as widely applicable as maximum likelihood estimation. Recall that if  $X_1, \dots, X_n$  is a random sample from a population with pdf or pmf  $f(x|\theta)$  ( $\theta$  may be a vector), the likelihood function is defined as

$$L(\theta|x_1, \dots, x_n) = L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \prod_{i=1}^n f(x_i|\theta).$$

Let  $\Theta$  denote the entire parameter space. Likelihood ratio tests are defined as follows.

**DEFINITION 8.2.1:** The *likelihood ratio test statistic* for testing  $H_0: \theta \in \Theta_0$  versus  $H_1: \theta \in \Theta_0^c$  is

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})}.$$

A *likelihood ratio test* (LRT) is any test that has a rejection region of the form  $\{\mathbf{x}: \lambda(\mathbf{x}) \leq c\}$ , where  $c$  is any number satisfying  $0 \leq c \leq 1$ .

The rationale behind LRTs may best be understood in the situation in which  $f(x|\theta)$  is the pmf of a discrete random variable. In this case, the numerator of  $\lambda(\mathbf{x})$  is the maximum probability of the observed sample, the maximum being computed over parameters in the null hypothesis. (See Exercise 8.4.) The denominator of  $\lambda(\mathbf{x})$  is the maximum probability of the observed sample over all possible parameters. The ratio of these two maxima is small if there are parameter points in the alternative hypothesis for which the observed sample is much more likely than for any parameter point in the null hypothesis. In this situation, the LRT criterion says  $H_0$  should be rejected and  $H_1$  accepted as true. Methods for selecting the number  $c$  are discussed in Section 8.3.

If we think of doing the maximization over both the entire parameter space (unrestricted maximization) and a subset of the parameter space (restricted maximization) then the correspondence between LRTs and MLEs becomes more clear. Suppose  $\hat{\theta}$ , an MLE of  $\theta$ , exists;  $\hat{\theta}$  is obtained by doing an unrestricted maximization of  $L(\theta|\mathbf{x})$ . We can also consider the MLE of  $\theta$ , call it  $\hat{\theta}_0$ , obtained by doing a restricted maximization, assuming  $\Theta_0$  is the parameter space. That is,  $\hat{\theta}_0 = \hat{\theta}_0(\mathbf{x})$  is the value of  $\theta \in \Theta_0$  that maximizes  $L(\theta|\mathbf{x})$ . Then, the LRT statistic is

$$\lambda(\mathbf{x}) = \frac{L(\hat{\theta}_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})}.$$

**Example 8.2.1:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, 1)$  population. Consider testing  $H_0: \theta = \theta_0$  versus  $H_1: \theta \neq \theta_0$ . Here  $\theta_0$  is a number fixed by the experimenter prior to the experiment. Since there is only one value of  $\theta$  specified by  $H_0$ , the numerator of  $\lambda(\mathbf{x})$  is  $L(\theta_0|\mathbf{x})$ . In Example 7.2.4 the (unrestricted) MLE of  $\theta$  was found to be  $\bar{X}$ , the sample mean. Thus the denominator of  $\lambda(\mathbf{x})$  is  $L(\bar{X}|\mathbf{x})$ . So the LRT statistic is

$$(8.2.1) \quad \lambda(\mathbf{x}) = \frac{(2\pi)^{-n/2} \exp[-\sum_{i=1}^n (x_i - \theta_0)^2/2]}{(2\pi)^{-n/2} \exp[-\sum_{i=1}^n (x_i - \bar{x})^2/2]}$$

$$= \exp \left[ \left( - \sum_{i=1}^n (x_i - \theta_0)^2 + \sum_{i=1}^n (x_i - \bar{x})^2 \right) / 2 \right].$$

The expression for  $\lambda(\mathbf{x})$  can be simplified by noting that

$$\sum_{i=1}^n (x_i - \theta_0)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \theta_0)^2.$$

Thus the LRT statistic is

$$(8.2.2) \quad \lambda(\mathbf{x}) = \exp [-n(\bar{x} - \theta_0)^2 / 2].$$

An LRT is a test that rejects  $H_0$  for small values of  $\lambda(\mathbf{x})$ . Using (8.2.2), the rejection region,  $\{\mathbf{x}: \lambda(\mathbf{x}) \leq c\}$ , can be written as

$$\{\mathbf{x}: |\bar{x} - \theta_0| \geq \sqrt{-2(\log c)/n}\}.$$

As  $c$  ranges between 0 and 1,  $\sqrt{-2(\log c)/n}$  ranges between 0 and  $\infty$ . Thus the LRTs are just those tests that reject  $H_0: \theta = \theta_0$  if the sample mean differs from the hypothesized value  $\theta_0$  by more than a specified amount. ||

The analysis in Example 8.2.1 is typical in that first the expression for  $\lambda(\mathbf{X})$  from Definition 8.2.1 is found, as we did in (8.2.1). Then the description of the rejection region is simplified, if possible, to an expression involving a simpler statistic,  $|\bar{X} - \theta_0|$  in the example:

**Example 8.2.2:** Let  $X_1, \dots, X_n$  be a random sample from an exponential population with pdf

$$f(x|\theta) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & x < \theta \end{cases}$$

where  $-\infty < \theta < \infty$ . The likelihood function is

$$L(\theta|\mathbf{x}) = \begin{cases} e^{-\sum x_i + n\theta} & \theta \leq x_{(1)} \\ 0 & \theta > x_{(1)} \end{cases}. \quad (x_{(1)} = \min x_i)$$

Consider testing  $H_0: \theta \leq \theta_0$  versus  $H_1: \theta > \theta_0$  where  $\theta_0$  is a value specified by the experimenter. Clearly  $L(\theta|\mathbf{x})$  is an increasing function of  $\theta$  on  $-\infty < \theta \leq x_{(1)}$ . Thus, the denominator of  $\lambda(\mathbf{x})$ , the unrestricted maximum of  $L(\theta|\mathbf{x})$ , is

$$L(x_{(1)}|\mathbf{x}) = e^{-\sum x_i + nx_{(1)}}.$$

If  $x_{(1)} \leq \theta_0$ , the numerator of  $\lambda(\mathbf{x})$  is also  $L(x_{(1)}|\mathbf{x})$ . But since we are maximizing  $L(\theta|\mathbf{x})$  over  $\theta \leq \theta_0$ , the numerator of  $\lambda(\mathbf{x})$  is  $L(\theta_0|\mathbf{x})$  if  $x_{(1)} > \theta_0$ . Therefore, the likelihood ratio test statistic is

$$\lambda(\mathbf{x}) = \begin{cases} 1 & x_{(1)} \leq \theta_0 \\ e^{-n(x_{(1)} - \theta_0)} & x_{(1)} > \theta_0 \end{cases}$$

A graph of  $\lambda(\mathbf{x})$  is shown in Figure 8.2.1. An LRT, a test that rejects  $H_0$  if  $\lambda(\mathbf{X}) \leq c$ , is a test with rejection region  $\{\mathbf{x} : x_{(1)} \geq \theta_0 - \frac{\log c}{n}\}$ . Note that the rejection region depends on the sample only through the sufficient statistic  $X_{(1)}$ . That this is generally the case will be seen in Theorem 8.2.1.

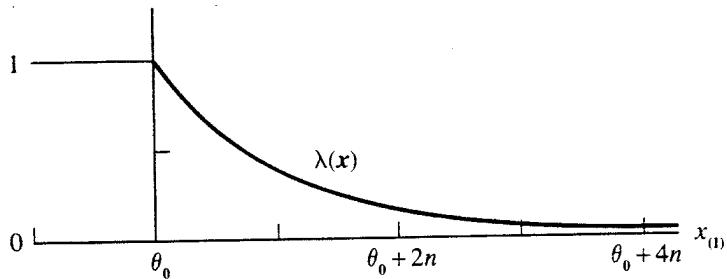


FIGURE 8.2.1  $\lambda(\mathbf{x})$ , a function only of  $x_{(1)}$ , for Example 8.2.2

Example 8.2.2 again illustrates the point, expressed in Section 7.2.2, that differentiation of the likelihood function is not the only method of finding an MLE. In Example 8.2.2,  $L(\theta|\mathbf{x})$  is not differentiable at  $\theta = x_{(1)}$ .

If  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$  with pdf or pmf  $g(t|\theta)$ , then we might consider constructing an LRT based on  $T$  and its likelihood function  $L^*(\theta|t) = g(t|\theta)$ , rather than on the sample  $\mathbf{X}$  and its likelihood function  $L(\theta|\mathbf{x})$ . Let  $\lambda^*(t)$  denote the likelihood ratio test statistic based on  $T$ . Given the intuitive notion that all the information about  $\theta$  in  $\mathbf{x}$  is contained in  $T(\mathbf{x})$ , the test based on  $T$  should be as good as the test based on the complete sample  $\mathbf{X}$ . In fact the tests are equivalent.

**THEOREM 8.2.1:** If  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$  and  $\lambda^*(t)$  and  $\lambda(\mathbf{x})$  are the LRT statistics based on  $T$  and  $\mathbf{X}$ , respectively, then  $\lambda^*(T(\mathbf{x})) = \lambda(\mathbf{x})$  for every  $\mathbf{x}$  in the sample space.

*Proof:* From the Factorization Theorem (Theorem 6.1.2), the pdf or pmf of  $\mathbf{X}$  can be written as  $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$  where  $g(t|\theta)$  is the pdf or pmf of  $T$  and  $h(\mathbf{x})$  does not depend on  $\theta$ . Thus

$$\begin{aligned} \lambda(\mathbf{x}) &= \frac{\sup_{\Theta_0} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})} \\ &= \frac{\sup_{\Theta_0} f(\mathbf{x}|\theta)}{\sup_{\Theta} f(\mathbf{x}|\theta)} \end{aligned}$$

$$\begin{aligned}
 &= \frac{\sup_{\Theta_0} g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sup_{\Theta} g(T(\mathbf{x})|\theta)h(\mathbf{x})} && (T \text{ is sufficient}) \\
 &= \frac{\sup_{\Theta_0} g(T(\mathbf{x})|\theta)}{\sup_{\Theta} g(T(\mathbf{x})|\theta)} && (h \text{ does not depend on } \theta) \\
 &= \frac{\sup_{\Theta_0} L^*(\theta|T(\mathbf{x}))}{\sup_{\Theta} L^*(\theta|T(\mathbf{x}))} && (g \text{ is the pdf or pmf of } T) \\
 &= \lambda^*(T(\mathbf{x})). \quad \square
 \end{aligned}$$

The comment after Example 8.2.1 was that, after finding an expression for  $\lambda(\mathbf{x})$ , we try to simplify that expression. In light of Theorem 8.2.1, one interpretation of this comment is that the simplified expression for  $\lambda(\mathbf{x})$  should depend on  $\mathbf{x}$  only through  $T(\mathbf{x})$  if  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$ .

**Example 8.2.3:** In Example 8.2.1, we can recognize that  $\bar{X}$  is a sufficient statistic for  $\theta$ . We could use the likelihood function associated with  $\bar{X}$  ( $\bar{X} \sim n(\theta, \frac{1}{n})$ ) to more easily reach the conclusion that a likelihood ratio test of  $H_0: \theta = \theta_0$  versus  $H_1: \theta \neq \theta_0$  rejects  $H_0$  for large values of  $|\bar{X} - \theta_0|$ .

Similarly in Example 8.2.2,  $X_{(1)} = \min X_i$  is a sufficient statistic for  $\theta$ . The likelihood function of  $X_{(1)}$  (the pdf of  $X_{(1)}$ ) is

$$L^*(\theta|x_{(1)}) = \begin{cases} ne^{-n(x_{(1)}-\theta)} & \theta \leq x_{(1)} \\ 0 & \theta > x_{(1)} \end{cases}.$$

This likelihood could also be used to derive the fact that a likelihood ratio test of  $H_0: \theta \leq \theta_0$  versus  $H_1: \theta > \theta_0$  rejects  $H_0$  for large values of  $X_{(1)}$ . ||

Likelihood ratio tests are also useful in situations where there are *nuisance* parameters, that is, parameters that are present in a model, but are not of direct inferential interest. The presence of such nuisance parameters does not affect the LRT construction method but, as might be expected, the presence of nuisance parameters might lead to a different test.

**Example 8.2.4:** Suppose  $X_1, \dots, X_n$  are a random sample from a  $n(\mu, \sigma^2)$ , and an experimenter is interested only in inferences about  $\mu$ , such as testing  $H_0: \mu \leq \mu_0$  versus  $H_1: \mu > \mu_0$ . Then the parameter  $\sigma^2$  is a nuisance parameter. The LRT statistic is

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{\max_{\{\mu, \sigma^2: \mu \leq \mu_0, \sigma^2 \geq 0\}} L(\mu, \sigma^2 | \mathbf{x})}{\max_{\{\mu, \sigma^2: -\infty < \mu < \infty, \sigma^2 \geq 0\}} L(\mu, \sigma^2 | \mathbf{x})} \\ &= \frac{\max_{\{\mu, \sigma^2: \mu \leq \mu_0, \sigma^2 \geq 0\}} L(\mu, \sigma^2 | \mathbf{x})}{L(\hat{\mu}, \hat{\sigma}^2 | \mathbf{x})},\end{aligned}$$

where  $\hat{\mu}$  and  $\hat{\sigma}^2$  are the MLEs of  $\mu$  and  $\sigma^2$  (Example 7.2.8). Furthermore, if  $\hat{\mu} \leq \mu_0$ , then the restricted maximum is the same as the unrestricted maximum, while if  $\hat{\mu} > \mu_0$ , the restricted maximum is  $L(\mu_0, \hat{\sigma}_0^2 | \mathbf{x})$  where  $\hat{\sigma}_0^2 = \sum (x_i - \mu_0)^2 / n$ . Thus

$$\lambda(\mathbf{x}) = \begin{cases} 1 & \text{if } \hat{\mu} \leq \mu_0 \\ \frac{L(\mu_0, \hat{\sigma}_0^2 | \mathbf{x})}{L(\hat{\mu}, \hat{\sigma}^2 | \mathbf{x})} & \text{if } \hat{\mu} > \mu_0 \end{cases}.$$

With some algebra, it can be shown that the test based on  $\lambda(\mathbf{x})$  is equivalent to a test based on Student's  $t$  distribution. Details are left to Exercise 8.47. (Exercises 8.48–8.52 also deal with nuisance parameter problems.) ||

### 8.2.2 Invariant Tests

The likelihood ratio method of Section 8.2.1 defines one test statistic which can be used in a hypothesis testing problem. In this section we use invariance considerations to develop different test statistics. We specify some invariance properties we would like a hypothesis testing procedure to possess, then see which tests (if any) possess these properties. The invariance properties state that the decisions reached by the hypothesis testing procedure should remain unchanged (invariant) under certain transformations of the data. The rationale behind these properties is essentially the same as that underlying the invariance properties of estimators we discussed in Sections 6.3 and 7.2.4. Our discussion of these properties will be simplified by use of a test function, which we now define in general.

**DEFINITION 8.2.2:** A *test function*,  $\phi(\mathbf{x})$ , for a hypothesis testing procedure is a function on the sample space whose value is one if  $\mathbf{x}$  is in the rejection region and zero if  $\mathbf{x}$  is in the acceptance region. That is,  $\phi(\mathbf{x})$  is the indicator function of the rejection region.

As described in Section 6.3, there are two types of invariance we consider, measurement invariance and formal invariance. We now discuss these concepts in the hypothesis testing framework.

Measurement invariance requires that if  $\mathbf{y} = (y_1, \dots, y_n) = g(x_1, \dots, x_n)$  is a one-to-one transformation of the data, then the decision made after observing  $\mathbf{y}$  should be the same as the decision made after observing  $\mathbf{x}$ . The function  $g$  simply denotes a change in measurement scale. For example,  $\mathbf{x}$  could be temperature measurements in degrees Fahrenheit and  $\mathbf{y}$  the measurements on the same samples in degrees Celsius. If  $\phi(\mathbf{x})$  is the proposed test function for the  $\mathbf{x}$  measurements and  $\phi^*(\mathbf{y})$  is the proposed

test function for the  $\mathbf{y}$  measurements, invariance with respect to measurement scale requires that  $\phi(\mathbf{x}) = \phi^*(g(\mathbf{x})) = \phi^*(\mathbf{y})$ .

Formal invariance requires that if two hypothesis testing problems have the same structure, that is, the same sample and parameter spaces, the same set of pdfs or pmfs, and the same hypotheses, then the same test should be used in both problems. Thus, if the transformed problem in terms of  $\mathbf{y}$  has the same structure as the original problem in terms of  $\mathbf{x}$ , then the tests should satisfy  $\phi^*(\mathbf{y}) = \phi(\mathbf{y}) = \phi(g(\mathbf{x}))$ . Putting this with the measurement invariance leads to the requirement that  $\phi(\mathbf{x}) = \phi(g(\mathbf{x}))$  for all  $\mathbf{x}$  in the sample space.

**DEFINITION 8.2.3:** An *invariant test* with respect to a function  $g(\mathbf{x})$  is any test whose test function satisfies  $\phi(\mathbf{x}) = \phi(g(\mathbf{x}))$  for all  $\mathbf{x}$  in the sample space.

The requirement that the transformed problem in terms of  $\mathbf{y}$  has the same structure as the original problem in terms of  $\mathbf{x}$  requires more comment. Suppose the original model is defined by the set of pdfs or pmfs for  $\mathbf{X}$ ,  $\{f(\mathbf{x}|\theta) : \theta \in \Theta\}$ . Denote the set of pdfs or pmfs for  $\mathbf{Y} = g(\mathbf{X})$  by  $\{h(\mathbf{y}|\theta) : \theta \in \Theta\}$ .

**DEFINITION 8.2.4:** The hypothesis testing problem  $H_0 : \theta \in \Theta_0$  versus  $H_1 : \theta \in \Theta_0^c$  is *invariant under the transformation*  $\mathbf{y} = g(\mathbf{x})$  if

- a.  $\{f(\mathbf{x}|\theta) : \theta \in \Theta_0\} = \{h(\mathbf{y}|\theta) : \theta \in \Theta_0\}$  and
- b.  $\{f(\mathbf{x}|\theta) : \theta \in \Theta_0^c\} = \{h(\mathbf{y}|\theta) : \theta \in \Theta_0^c\}$ .

Definition 8.2.4 implies that  $\mathbf{X}$  and  $\mathbf{Y}$  have the same sample space since their distributions are defined by the same pdfs or pmfs. Thus,  $g(\mathbf{x})$  must be a function from the sample space  $\mathcal{X}$  of  $\mathbf{X}$  onto itself. Furthermore, we are usually interested not just in one transformation, but in a *family* of transformations  $\{g(\mathbf{x}) : g \in \mathcal{G}\}$ . We say the testing problem is invariant under the *family* if it is invariant under each transformation  $g$ . In this case the invariance properties require that  $\phi(\mathbf{x}) = \phi(g(\mathbf{x}))$  for every  $g \in \mathcal{G}$  as well as every  $\mathbf{x} \in \mathcal{X}$ . These notions of invariance of the problem are the same as those in Section 6.3 with the addition that, according to Definition 8.2.4, the family of possible distributions of  $\mathbf{X}$  under  $H_0$  gets mapped onto the family of possible distributions of  $\mathbf{Y}$  under  $H_0$  and the family of possible distributions of  $\mathbf{X}$  under  $H_1$  gets mapped onto the family of possible distributions of  $\mathbf{Y}$  under  $H_1$ . These concepts are illustrated in the following examples.

**Example 8.2.5:** Let  $X_1, \dots, X_n$  be iid Bernoulli( $p$ ), and suppose that we are interested in an invariant test of

$$H_0 : p = \frac{1}{2} \quad \text{versus} \quad H_1 : p \neq \frac{1}{2},$$

using the two-element group described in Example 6.3.1. If  $Y = \sum X_i$  (sufficient statistic), then the two elements of the group prescribe

$$g_1(y) = n - y, \quad g_2(y) = y.$$

From Definition 8.2.3, an invariant test function  $\phi$  must satisfy  $\phi(y) = \phi(n - y)$ , that is, the same action is taken whether  $y$  or  $n - y$  is observed. Any test function  $\phi$  satisfying this requirement is invariant with respect to this group. We need to check the conditions in Definition 8.2.4 for both  $g_1(y)$  and  $g_2(y)$ . But since  $g_2(y)$  is the identity map, the conditions are obviously satisfied for  $g_2(y)$ . For  $g_1(y)$ ,  $g_1(Y) = n - Y \sim \text{binomial}(n, 1 - p)$ . Since  $H_0$  specifies  $p = \frac{1}{2}$ , the two sets in condition (a) both have just one element. Furthermore, the distribution of  $Y$  is  $\text{binomial}(n, \frac{1}{2})$  and the distribution of  $n - Y$  is  $\text{binomial}(n, \frac{1}{2})$ , the same distribution. In condition (b), the set of distributions of  $Y$  is all  $\text{binomial}(n, p)$  distributions,  $0 < p < 1, p \neq \frac{1}{2}$ . The set of distributions of  $n - Y$  is all  $\text{binomial}(n, 1 - p)$  distributions,  $0 < p < 1, p \neq \frac{1}{2}$ . But this is the same set of distributions. As  $p$  varies between 0 and 1 (skipping  $\frac{1}{2}$ ),  $1 - p$  varies between 0 and 1 (skipping  $\frac{1}{2}$ ).

For a specific example, if 10 Bernoulli trials are performed, an invariant test is

$$\phi(\Sigma x) = \begin{cases} 0 & \text{if } \Sigma x = 3, 4, 5, 6, 7 \\ 1 & \text{if } \Sigma x = 0, 1, 2, 8, 9, 10 \end{cases}.$$

(See Exercise 8.46.) ||

In the next example we consider a somewhat more complicated situation. The family of distributions is parameterized with two parameters, but the hypothesis test is concerned with only one parameter. Thus, we must deal with nuisance parameters.

**Example 8.2.6:** Let  $\bar{X}$  and  $S^2$  be independent with  $\bar{X} \sim n(\mu, \sigma^2/n)$  and  $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$ . (Think of  $(\bar{X}, S^2)$  as the sufficient statistic for a random sample  $X_1, \dots, X_n$  from a  $n(\mu, \sigma^2)$  population.) Suppose we want to test  $H_0: \mu \leq 0$  versus  $H_1: \mu > 0$ , using a test that is invariant with respect to the group of scale transformations

$$\{g_c(\bar{x}, s^2): c > 0\}, \quad \text{where } g_c(\bar{x}, s^2) = (c\bar{x}, c^2 s^2).$$

To check that the hypothesis testing problem is invariant under the group of transformations, note that

$$\begin{aligned} c\bar{X} \text{ and } c^2 S^2 \text{ are independent,} \\ c\bar{X} \sim n(c\mu, c^2 \sigma^2/n), \text{ and} \\ (n-1)c^2 S^2 / c^2 \sigma^2 \sim \chi_{n-1}^2. \end{aligned}$$

Therefore, the distribution of  $(c\bar{X}, c^2 S^2)$  when  $\theta = (\mu, \sigma^2)$  is the same as the distribution of  $(\bar{X}, S^2)$  when  $\theta = (c\mu, c^2 \sigma^2)$ . On the other hand, the distribution of  $(\bar{X}, S^2)$  when  $\theta = (\mu, \sigma^2)$  is the same as the distribution of  $(c\bar{X}, c^2 S^2)$  when  $\theta = (\mu/c, \sigma^2/c^2)$ . Furthermore, the hypothesis test is invariant because  $c\mu \leq 0$  if and only if  $\mu \leq 0$  since  $c > 0$ . Therefore, conditions (a) and (b) in Definition 8.2.4 are satisfied.

It may not be immediately clear from Definition 8.2.3 which tests are invariant. We will now examine the requirements of Definition 8.2.3 more closely to obtain an

alternative description of tests that are invariant in this problem. An invariant test is one that satisfies

$$\phi(\bar{x}, s^2) = \phi(c\bar{x}, c^2 s^2) \quad \text{for every } c > 0, s^2 > 0, \text{ and } -\infty < \bar{x} < \infty.$$

In particular, setting  $c = 1/\sqrt{s^2} = 1/s$  we obtain  $\phi(\bar{x}, s^2) = \phi(\bar{x}/s, 1)$ . Thus, invariant tests depend on  $(\bar{x}, s^2)$  only through the ratio  $\bar{x}/s$ . On the other hand, if  $\bar{x}_1/s_1 \neq \bar{x}_2/s_2$ , then there is no  $c > 0$  such that  $(\bar{x}_2, s_2^2) = (c\bar{x}_1, c^2 s_1^2)$ . So invariance places no restriction on  $\phi(\bar{x}_1, s_1^2)$  and  $\phi(\bar{x}_2, s_2^2)$  if  $\bar{x}_1/s_1 \neq \bar{x}_2/s_2$ . Thus, the invariant tests are exactly those tests for which  $\phi(\bar{x}, s^2)$  depends only on  $\bar{x}/s$ . Tests based on  $\bar{x}/s$  are equivalent to tests based on  $\bar{x}/\sqrt{s^2/n}$  since this is a one-to-one function. Recall that the statistic  $\bar{X}/\sqrt{S^2/n}$  has Student's  $t$  distribution with  $n - 1$  degrees of freedom if  $\mu = 0$ .

To make this example more concrete, if  $x_1, \dots, x_n$  are measurements of temperature in degrees Fahrenheit above freezing ( $32^\circ\text{F}$ ) (negative values being degrees below freezing) then  $g_c(\bar{x}, s^2)$  for  $c = \frac{5}{9}$  are the corresponding measurements in degrees Celsius and  $g_c(\bar{x}, s^2)$  for other values of  $c$  correspond to other unnamed temperature scales. The null hypothesis states that the average temperature is at most freezing. ||

In Example 8.2.6, restriction to invariant tests reduced consideration from all tests based on the bivariate statistic  $(\bar{X}, S^2)$  to all tests based on the univariate statistic  $\bar{X}/\sqrt{S^2/n}$ . Restriction to this smaller set of tests should make the choice of a good test easier.

Furthermore, we began Example 8.2.6 by first reducing the sample  $X_1, \dots, X_n$  to a sufficient statistic  $(\bar{X}, S^2)$ , and *then* applying an invariance argument. A natural question to ask is, "If we *first* applied the invariance argument, do we end up with the same answer as above?" The answer, in most cases, is yes. Whether we first reduce by invariance, then invoke sufficiency or first reduce by sufficiency, then invoke invariance, we end up with the same answer. The exact conditions under which the order of application of sufficiency and invariance is irrelevant are rather technical and are tied in with properties of groups. This topic was first investigated by Hall, Wijsman, and Ghosh (1965), and is treated in detail by Lehmann (1986, Chapter 6).

As mentioned before, Example 8.2.6 involved a nuisance parameter. The hypotheses concerned only  $\mu$  but the model also included the unknown parameter  $\sigma^2$ . The invariant tests are based on the  $t$  statistic whose distribution under  $H_0$ ,  $t$  with  $n - 1$  degrees of freedom, does not depend on the nuisance parameter  $\sigma^2$ . This is no accident. Invariance considerations are sometimes helpful in deriving a test statistic whose distribution under  $H_0$  does not depend on the nuisance parameter. The group of scale transformations in Example 8.2.6 achieved this goal.

### 8.2.3 Bayesian Tests

Hypothesis testing problems may also be formulated in a Bayesian model. Recall from Section 7.2.3 that a Bayesian model includes not only the sampling distribution

$f(\mathbf{x}|\theta)$  but also the prior distribution  $\pi(\theta)$ , with the prior distribution reflecting the experimenter's opinion about the parameter  $\theta$  prior to sampling.

The Bayesian paradigm prescribes that the sample information be combined with the prior information using Bayes' Theorem to obtain the posterior distribution  $\pi(\theta|\mathbf{x})$ . All inferences about  $\theta$  are now based on the posterior distribution.

In a hypothesis testing problem, the posterior distribution may be used to calculate the probabilities that  $H_0$  and  $H_1$  are true. Remember,  $\pi(\theta|\mathbf{x})$  is a probability distribution for a random variable. Hence, the posterior probabilities  $P(\theta \in \Theta_0|\mathbf{x}) = P(H_0 \text{ is true}|\mathbf{x})$  and  $P(\theta \in \Theta_0^c|\mathbf{x}) = P(H_1 \text{ is true}|\mathbf{x})$  may be computed.

The probabilities  $P(H_0 \text{ is true}|\mathbf{x})$  and  $P(H_1 \text{ is true}|\mathbf{x})$  are not meaningful to the classical statistician. The classical statistician considers  $\theta$  to be a fixed number. Consequently, a hypothesis is *either true or false*. If  $\theta \in \Theta_0$ ,  $P(H_0 \text{ is true}|\mathbf{x}) = 1$  and  $P(H_1 \text{ is true}|\mathbf{x}) = 0$  for all values of  $\mathbf{x}$ . If  $\theta \in \Theta_0^c$ , these values are reversed. Since these probabilities are unknown (since  $\theta$  is unknown) and do not depend on the sample  $\mathbf{x}$ , they are not used by the classical statistician. In a Bayesian formulation of a hypothesis testing problem, these probabilities depend on the sample  $\mathbf{x}$  and can give useful information about the veracity of  $H_0$  and  $H_1$ .

One way a Bayesian hypothesis tester may choose to use the posterior distribution is to decide to accept  $H_0$  as true if  $P(\theta \in \Theta_0|\mathbf{X}) \geq P(\theta \in \Theta_0^c|\mathbf{X})$  and to reject  $H_0$  otherwise. In the terminology of the previous sections, the test statistic, a function of the sample, is  $P(\theta \in \Theta_0^c|\mathbf{X})$  and the rejection region is  $\{\mathbf{x} : P(\theta \in \Theta_0^c|\mathbf{x}) > \frac{1}{2}\}$ . Alternatively, if the Bayesian hypothesis tester wishes to guard against falsely rejecting  $H_0$ , he may decide to reject  $H_0$  only if  $P(\theta \in \Theta_0^c|\mathbf{X})$  is greater than some large number, .99 for example.

**Example 8.2.7:** Let  $X_1, \dots, X_n$  be iid  $n(\theta, \sigma^2)$  and let the prior distribution on  $\theta$  be  $n(\mu, \tau^2)$  where  $\sigma^2$ ,  $\mu$ , and  $\tau^2$  are known. Consider testing  $H_0 : \theta \leq \theta_0$  versus  $H_1 : \theta > \theta_0$ . From Example 7.2.10, the posterior  $\pi(\theta|\bar{x})$  is normal with mean  $(n\tau^2\bar{x} + \sigma^2\mu)/(n\tau^2 + \sigma^2)$  and variance  $\sigma^2\tau^2/(n\tau^2 + \sigma^2)$ .

If we decide to accept  $H_0$  if and only if  $P(\theta \in \Theta_0|\mathbf{X}) \geq P(\theta \in \Theta_0^c|\mathbf{X})$ , then we will accept  $H_0$  if and only if

$$\frac{1}{2} \leq P(\theta \in \Theta_0|\mathbf{X}) = P(\theta \leq \theta_0|\mathbf{X}).$$

Since  $\pi(\theta|\mathbf{x})$  is symmetric, this is true if and only if the mean of  $\pi(\theta|\mathbf{x})$  is less than or equal to  $\theta_0$ . Therefore  $H_0$  will be accepted as true if

$$\bar{X} \leq \theta_0 + \frac{\sigma^2(\theta_0 - \mu)}{n\tau^2}$$

and  $H_1$  will be accepted as true otherwise. In particular, if  $\mu = \theta_0$  so that prior to experimentation probability  $\frac{1}{2}$  is assigned to both  $H_0$  and  $H_1$ , then  $H_0$  will be accepted as true if  $\bar{x} \leq \theta_0$  and  $H_1$  accepted otherwise.

Other methods that use the posterior distribution to make inferences in hypothesis testing problems are discussed in Chapter 10.

### 8.2.4 Union–Intersection and Intersection–Union Tests

In some situations, tests for complicated null hypotheses can be developed from tests for simpler null hypotheses. We discuss two related methods.

The *union–intersection method* of test construction might be useful when the null hypothesis is conveniently expressed as an intersection, say

$$(8.2.3) \quad H_0 : \theta \in \bigcap_{\gamma \in \Gamma} \Theta_\gamma.$$

Here  $\Gamma$  is an arbitrary index set which may be finite or infinite, depending on the problem. Suppose that tests are available for each of the problems of testing  $H_{0\gamma} : \theta \in \Theta_\gamma$  versus  $H_{1\gamma} : \theta \in \Theta_\gamma^c$ . Say the rejection region for the test of  $H_{0\gamma}$  is  $\{x : T_\gamma(x) \in R_\gamma\}$ . Then the rejection region for the union–intersection test is

$$(8.2.4) \quad \bigcup_{\gamma \in \Gamma} \{x : T_\gamma(x) \in R_\gamma\}.$$

The rationale is simple. If any one of the hypotheses  $H_{0\gamma}$  is rejected, then  $H_0$ , which, by (8.2.3), is true only if  $H_{0\gamma}$  is true for every  $\gamma$ , must also be rejected. Only if each of the hypotheses  $H_{0\gamma}$  is accepted as true will the intersection  $H_0$  be accepted as true.

In some situations a simple expression for the rejection region of a union–intersection test can be found. In particular, suppose that each of the individual tests has a rejection region of the form  $\{x : T_\gamma(x) > c\}$ , where  $c$  does not depend on  $\gamma$ . The rejection region for the union–intersection test, given in (8.2.4), can be expressed as

$$\bigcup_{\gamma \in \Gamma} \{x : T_\gamma(x) > c\} = \{x : \sup_{\gamma \in \Gamma} T_\gamma(x) > c\}.$$

Thus the test statistic for testing  $H_0$  is  $T(x) = \sup_{\gamma \in \Gamma} T_\gamma(x)$ . Some examples in which  $T(x)$  has a simple formula may be found in Chapter 11.

**Example 8.2.8:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population. Consider testing  $H_0: \mu = \mu_0$  versus  $H_1: \mu \neq \mu_0$  where  $\mu_0$  is a specified number. We can write  $H_0$  as the intersection of two sets,

$$H_0 : \{\mu: \mu \leq \mu_0\} \cap \{\mu: \mu \geq \mu_0\}.$$

The LRT of  $H_{0L}: \mu \leq \mu_0$  versus  $H_{1L}: \mu > \mu_0$  is

$$\text{reject } H_{0L}: \mu \leq \mu_0 \text{ in favor of } H_{1L}: \mu > \mu_0 \text{ if } \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq t_L$$

(see Exercise 8.47). Similarly, the LRT of  $H_{0U}: \mu \geq \mu_0$  versus  $H_{1U}: \mu < \mu_0$  is

$$\text{reject } H_{0U}: \mu \geq \mu_0 \text{ in favor of } H_{1U}: \mu < \mu_0 \text{ if } \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \leq t_U.$$

Thus the union-intersection test of  $H_0: \mu = \mu_0$  versus  $H_1: \mu \neq \mu_0$  formed from these two LRTs is

$$\text{reject } H_0 \text{ if } \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq t_L \text{ or } \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \leq t_U.$$

If  $t_L = -t_U \geq 0$ , the union-intersection test can be more simply expressed as

$$\text{reject } H_0 \text{ if } \frac{|\bar{X} - \mu_0|}{S/\sqrt{n}} \geq t_L.$$

It turns out that this union-intersection test is also the LRT for this problem (see Exercise 8.48) and is called the two-sided  $t$  test. ||

The union-intersection method of test construction is useful if the null hypothesis is conveniently expressed as an intersection. Another method, the *intersection-union method*, may be useful if the null hypothesis is conveniently expressed as a union. Suppose we wish to test the null hypothesis

$$(8.2.5) \quad H_0: \theta \in \bigcup_{\gamma \in \Gamma} \Theta_\gamma.$$

Suppose that for each  $\gamma \in \Gamma$ ,  $\{x: T_\gamma(x) \in R_\gamma\}$  is the rejection region for a test of  $H_{0\gamma}: \theta \in \Theta_\gamma$  versus  $H_{1\gamma}: \theta \in \Theta_\gamma^c$ . Then the rejection region for the intersection-union test of  $H_0$  versus  $H_1$  is

$$(8.2.6) \quad \bigcap_{\gamma \in \Gamma} \{x: T_\gamma(x) \in R_\gamma\}.$$

From (8.2.5),  $H_0$  is false if and only if *all* of the  $H_{0\gamma}$  are false, so  $H_0$  can be rejected if and only if each of the individual hypotheses  $H_{0\gamma}$  can be rejected. Again, the test can be greatly simplified if the rejection regions for the individual hypotheses are all of the form  $\{x: T_\gamma(x) \geq c\}$  ( $c$  independent of  $\gamma$ ). In such cases, the rejection region for  $H_0$  is

$$\bigcap_{\gamma \in \Gamma} \{x: T_\gamma(x) \geq c\} = \{x: \inf_{\gamma \in \Gamma} T_\gamma(x) \geq c\}.$$

The intersection-union test rejects  $H_0$  for large values of the test statistic  $\inf_{\gamma \in \Gamma} T_\gamma(\mathbf{x})$ .

**Example 8.2.9:** The topic of acceptance sampling provides an extremely useful application of an intersection-union test, as this example will illustrate. (See Berger (1982) for a more detailed treatment of this problem.)

Two parameters that are important in assessing the quality of upholstery fabric are  $\theta_1$ , the mean breaking strength, and  $\theta_2$ , the probability of passing a flammability test. Standards may dictate that  $\theta_1$  should be over 50 pounds and  $\theta_2$  should be over

.95, and the fabric is acceptable only if it meets both of these standards. This can be modeled with the hypothesis test

$$H_0 : \{\theta_1 \leq 50 \text{ or } \theta_2 \leq .95\} \quad \text{versus} \quad H_1 : \{\theta_1 > 50 \text{ and } \theta_2 > .95\},$$

where a batch of material is acceptable only if  $H_1$  is accepted.

Suppose  $X_1, \dots, X_n$  are measurements of breaking strength for  $n$  samples, and are assumed to be iid  $n(\theta_1, \sigma^2)$ . The LRT of  $H_{01} : \theta_1 \leq 50$  will reject  $H_{01}$  if  $(\bar{X} - 50)/(S/\sqrt{n}) > t$ . Suppose that we also have the results of  $m$  flammability tests, denoted by  $Y_1, \dots, Y_m$ , where  $Y_i = 1$  if the  $i$ th sample passes the test and  $Y_i = 0$  otherwise. If  $Y_1, \dots, Y_m$  are modeled as iid Bernoulli( $\theta_2$ ) random variables, the LRT will reject  $H_{02} : \theta_2 \leq .95$  if  $\sum_{i=1}^m Y_i > b$  (see Exercise 8.3). Putting all of this together, the rejection region for the intersection–union test is given by

$$\left\{ (x, y) : \frac{\bar{x} - 50}{s/\sqrt{n}} > t \text{ and } \sum_{i=1}^m y_i > b \right\}.$$

Thus the intersection–union test decides the product is acceptable, that is,  $H_1$  is true, if and only if it decides that each of the individual parameters meets its standard, that is,  $H_{1i}$  is true. If more than two parameters define a product's quality, individual tests for each parameter can be combined, by means of the intersection–union method, to yield an overall test of the product's quality. ||

## 8.3 Methods of Evaluating Tests

In deciding to accept or reject the null hypothesis  $H_0$ , an experimenter might be making a mistake. Usually, hypothesis tests are evaluated and compared through their probabilities of making mistakes. In this section we discuss how these error probabilities can be controlled. In some cases, it can even be determined which tests have the smallest possible error probabilities.

### 8.3.1 Error Probabilities and the Power Function

A hypothesis test of  $H_0 : \theta \in \Theta_0$  versus  $H_1 : \theta \in \Theta_0^c$  might make one of two types of errors. These two types of errors traditionally have been given the non-mnemonic names, Type I Error and Type II Error. If  $\theta \in \Theta_0$  but the hypothesis test incorrectly decides to reject  $H_0$ , then the test has made a *Type I Error*. If, on the other hand,  $\theta \in \Theta_0^c$  but the test decides to accept  $H_0$ , a *Type II Error* has been made. These two different situations are depicted in Figure 8.3.1.

Suppose  $R$  denotes the rejection region for a test. Then for  $\theta \in \Theta_0$ , the test will make a mistake if  $x \in R$  so the probability of a Type I Error is  $P_\theta(X \in R)$ . For  $\theta \in \Theta_0^c$ , the probability of a Type II Error is  $P_\theta(X \in R^c)$ . This switching from  $R$  to  $R^c$  is a bit confusing, but, if we realize that  $P_\theta(X \in R^c) = 1 - P_\theta(X \in R)$ , then the function of  $\theta$ ,  $P_\theta(X \in R)$ , contains all the information about the test with rejection region  $R$ . We have

		Decision	
		Accept $H_0$	Reject $H_0$
True hypothesis	$H_0$	Correct decision	Type I error
	$H_1$	Type II error	Correct decision

**FIGURE 8.3.1** Two types of errors in hypothesis testing

$$P_\theta(X \in R) = \begin{cases} \text{probability of a Type I Error} & \text{if } \theta \in \Theta_0 \\ \text{one minus the probability of a Type II Error} & \text{if } \theta \in \Theta_0^c \end{cases}$$

This consideration leads to the following definition.

**DEFINITION 8.3.1:** The *power function* of a hypothesis test with rejection region  $R$  is the function of  $\theta$  defined by  $\beta(\theta) = P_\theta(X \in R)$ .

The ideal power function is zero for all  $\theta \in \Theta_0$  and one for all  $\theta \in \Theta_0^c$ . Except in trivial situations, this ideal cannot be attained. Qualitatively, a good test has power function near one for most  $\theta \in \Theta_0^c$  and near zero for most  $\theta \in \Theta_0$ .

**Example 8.3.1:** Let  $X \sim \text{binomial}(5, \theta)$ . Consider testing  $H_0 : \theta \leq \frac{1}{2}$  versus  $H_1 : \theta > \frac{1}{2}$ . Consider first the test that rejects  $H_0$  if and only if all “successes” are observed. The power function for this test is

$$\beta_1(\theta) = P_\theta(X \in R) = P_\theta(X = 5) = \theta^5.$$

The graph of  $\beta_1(\theta)$  is in Figure 8.3.2 (page 360). In examining this power function, we might decide that although the probability of a Type I Error is acceptably low ( $\beta_1(\theta) \leq (\frac{1}{2})^5 = .0312$  for all  $\theta \leq \frac{1}{2}$ ), the probability of a Type II Error is too high ( $\beta_1(\theta)$  is too small) for most  $\theta > \frac{1}{2}$ . The probability of a Type II Error is less than  $\frac{1}{2}$  only if  $\theta > (\frac{1}{2})^{1/5} = .87$ . To achieve smaller Type II Error probabilities, we might consider using the test that rejects  $H_0$  if  $X = 3, 4$ , or  $5$ . The power function for this test is

$$\beta_2(\theta) = P_\theta(X = 3, 4, \text{ or } 5) = \binom{5}{3} \theta^3(1-\theta)^2 + \binom{5}{4} \theta^4(1-\theta)^1 + \binom{5}{5} \theta^5(1-\theta)^0.$$

The graph of  $\beta_2(\theta)$  is also in Figure 8.3.2. It can be seen in Figure 8.3.2 that the second test has achieved a smaller Type II Error probability in that  $\beta_2(\theta)$  is larger for  $\theta > \frac{1}{2}$ . But the Type I Error probability is larger for the second test;  $\beta_2(\theta)$  is larger for  $\theta \leq \frac{1}{2}$ . If a choice is to be made between these two tests, the researcher must

decide which error structure, that described by  $\beta_1(\theta)$  or that described by  $\beta_2(\theta)$ , is more acceptable.

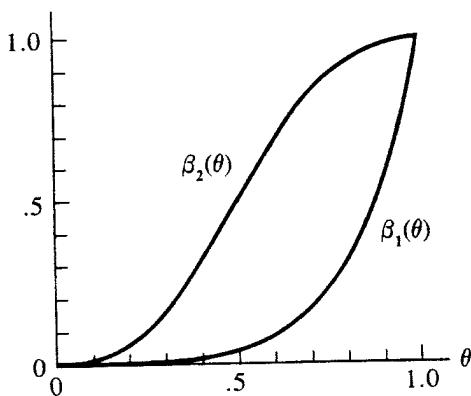


FIGURE 8.3.2 Power functions for Example 8.3.2

**Example 8.3.2:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population,  $\sigma^2$  known. An LRT of  $H_0: \theta \leq \theta_0$  versus  $H_1: \theta > \theta_0$  is a test that rejects  $H_0$  if  $(\bar{X} - \theta_0)/(\sigma/\sqrt{n}) > c$  (Exercise 8.47). The constant  $c$  can be any positive number. The power function of this test is

$$\begin{aligned}\beta(\theta) &= P_\theta \left( \frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c \right) \\ &= P_\theta \left( \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} \right) \\ &= P \left( Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} \right)\end{aligned}$$

where  $Z$  is a standard normal random variable, since  $(\bar{X} - \theta)/(\sigma/\sqrt{n}) \sim n(0, 1)$ . As  $\theta$  increases from  $-\infty$  to  $\infty$ , it is easy to see that this normal probability increases from zero to one. Therefore, it follows that  $\beta(\theta)$  is an increasing function of  $\theta$ , with

$$\lim_{\theta \rightarrow -\infty} \beta(\theta) = 0, \lim_{\theta \rightarrow \infty} \beta(\theta) = 1, \text{ and } \beta(\theta_0) = \alpha \text{ if } P(Z > c) = \alpha.$$

A graph of  $\beta(\theta)$  for  $c = 1.28$  is given in Figure 8.3.3.

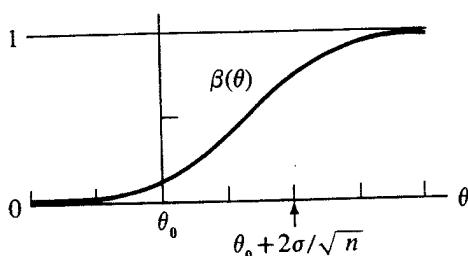


FIGURE 8.3.3 Power function for Example 8.3.2

Typically, the power function of a test will depend on the sample size  $n$ . If  $n$  can be chosen by the experimenter, consideration of the power function might help determine what sample size is appropriate in an experiment.

**Example 8.3.2 (Continued):** Suppose the experimenter wishes to have a maximum Type I Error probability of .1. Suppose, in addition, the experimenter wishes to have a maximum Type II Error probability of .2 if  $\theta \geq \theta_0 + \sigma$ . We now show how to choose  $c$  and  $n$  to achieve these goals, using a test that rejects  $H_0 : \theta \leq \theta_0$  if  $(\bar{X} - \theta_0)/(\sigma/\sqrt{n}) > c$ . As noted above, the power function of such a test is

$$\beta(\theta) = P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right).$$

Because  $\beta(\theta)$  is increasing in  $\theta$ , the requirements will be met if

$$\beta(\theta_0) = .1 \quad \text{and} \quad \beta(\theta_0 + \sigma) = .8.$$

By choosing  $c = 1.28$ , we achieve  $\beta(\theta_0) = P(Z > 1.28) = .1$  (actually .1003 from Table 1), regardless of  $n$ . Now we wish to choose  $n$  so that  $\beta(\theta_0 + \sigma) = P(Z > 1.28 - \sqrt{n}) = .8$ . But, from Table 1,  $P(Z > -.84) = .8$ . So setting  $1.28 - \sqrt{n} = -.84$  and solving for  $n$  yields  $n = 4.49$ . Of course  $n$  must be an integer. So choosing  $c = 1.28$  and  $n = 5$  yields a test with error probabilities controlled as specified by the experimenter. ||

For a fixed sample size, it is usually impossible to make both types of error probabilities arbitrarily small. In searching for a good test, it is common to restrict consideration to tests that control the Type I Error probability at a specified level. Within this class of tests we then search for tests that have Type II Error probability that is as small as possible. The following two terms are useful when discussing tests that control Type I Error probabilities.

**DEFINITION 8.3.2:** For  $0 \leq \alpha \leq 1$ , a test with power function  $\beta(\theta)$  is a *size  $\alpha$  test* if  $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$ .

**DEFINITION 8.3.3:** For  $0 \leq \alpha \leq 1$ , a test with power function  $\beta(\theta)$  is a *level  $\alpha$  test* if  $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$ .

Some authors do not make the distinction between the terms *size* and *level* that we have made, and sometimes these terms are used interchangeably. But according to our definitions, the set of level  $\alpha$  tests contains the set of size  $\alpha$  tests. Moreover, the distinction becomes important in complicated models and complicated testing situations, where it is often computationally impossible to construct a size  $\alpha$  test. In such situations, an experimenter must be satisfied with a level  $\alpha$  test, realizing that some compromises may be made. We will see some examples, especially in conjunction with union-intersection and intersection-union tests.

Experimenters commonly specify the level of the test they wish to use, with typical choices being  $\alpha = .01, .05$ , and  $.10$ . Be aware that, in fixing the level of

the test, the experimenter is controlling only the Type I Error probabilities, not the Type II Error. If this approach is taken, the experimenter should specify the null and alternative hypotheses so that it is most important to control the Type I Error probability. For example, suppose an experimenter expects an experiment to give support to a particular hypothesis, but she does not wish to make the assertion unless the data really do give convincing support. The test can be set up so that the alternative hypothesis is the one that she expects the data to support, and hopes to prove. (The alternative hypothesis is sometimes called the *research hypothesis* in this context.) By using a level  $\alpha$  test with small  $\alpha$ , the experimenter is guarding against saying the data support the research hypothesis when it is false.

The methods of Section 8.2 usually yield test statistics and general forms for rejection regions. However, they do not generally lead to one specific test. For example, an LRT (Definition 8.2.1) is one that rejects  $H_0$  if  $\lambda(\mathbf{X}) \leq c$ , but  $c$  was unspecified, so not one but an entire class of LRTs is defined, one for each value of  $c$ . The restriction to size  $\alpha$  tests may now lead to the choice of one out of the class of tests.

**Example 8.3.3:** In general, a size  $\alpha$  LRT is constructed by choosing  $c$  such that  $\sup_{\theta \in \Theta_0} P_\theta(\lambda(\mathbf{X}) \leq c) = \alpha$ . How that  $c$  is determined depends on the particular problem. For example, in Example 8.2.1  $\Theta_0$  consists of the single point  $\theta = \theta_0$  and  $\sqrt{n}(\bar{X} - \theta_0) \sim n(0, 1)$  if  $\theta = \theta_0$ . So the test

$$\text{reject } H_0 \text{ if } |\bar{X} - \theta_0| \geq z_{\alpha/2}/\sqrt{n},$$

where  $z_{\alpha/2}$  satisfies  $P(Z \geq z_{\alpha/2}) = \alpha/2$  with  $Z \sim n(0, 1)$ , is the size  $\alpha$  LRT. Specifically, this corresponds to choosing  $c = \exp(-z_{\alpha/2}^2/2)$ , but this is not an important point.

For the problem described in Example 8.2.2, finding a size  $\alpha$  LRT is complicated by the fact that the null hypothesis  $H_0: \theta \leq \theta_0$  consists of more than one point. The LRT rejects  $H_0$  if  $X_{(1)} \geq c$  where  $c$  is chosen so that this is a size  $\alpha$  test. But if  $c = (-\log \alpha)/n + \theta_0$  then

$$P_{\theta_0}(X_{(1)} \geq c) = e^{-n(c-\theta_0)} = \alpha.$$

Since  $\theta$  is a location parameter for  $X_{(1)}$ ,

$$P_\theta(X_{(1)} \geq c) \leq P_{\theta_0}(X_{(1)} \geq c) \quad \text{for any } \theta \leq \theta_0.$$

Thus

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \sup_{\theta \leq \theta_0} P_\theta(X_{(1)} \geq c) = P_{\theta_0}(X_{(1)} \geq c) = \alpha$$

and this  $c$  yields the size  $\alpha$  LRT. ||

*A note on notation:* In the above example we used the notation  $z_{\alpha/2}$  to denote the point having probability  $\alpha/2$  to the right of it for a standard normal pdf. We will

use this notation in general, not just for the normal, but for other distributions as well (defining what we need to for clarity's sake). For example, the point  $z_\alpha$  satisfies  $P(Z > z_\alpha) = \alpha$ , where  $Z \sim N(0, 1)$ ;  $t_{n-1, \alpha/2}$  satisfies  $P(T_{n-1} > t_{n-1, \alpha/2}) = \alpha/2$ , where  $T_{n-1} \sim t_{n-1}$ ; and  $\chi^2_{p, 1-\alpha}$  satisfies  $P(\chi^2_p > \chi^2_{p, 1-\alpha}) = 1 - \alpha$ , where  $\chi^2_p$  is a chi squared random variable with  $p$  degrees of freedom. Points like  $z_{\alpha/2}, z_\alpha, t_{n-1, \alpha/2}$ , and  $\chi^2_{p, 1-\alpha}$  are known as *cutoff points*.

**Example 8.3.4:** The invariant test statistic for the testing problem described in Example 8.2.6 is  $T = \bar{X}/\sqrt{S^2/n}$ . Invariance considerations told us nothing about which values of  $T$  should lead to rejection of  $H_0: \mu \leq 0$ . Since  $\bar{X}$  is a good estimate of  $\mu$ , it seems reasonable to reject  $H_0$  for large values of  $\bar{X}$ , that is, for large values of  $T$ . To find a size  $\alpha$  invariant test we must determine  $c$  such that

$$\sup_{\theta \in \Theta_0} P_\theta(T \geq c) = \alpha, \quad \Theta_0 = \{(\mu, \sigma^2): \mu \leq 0, \sigma^2 > 0\}.$$

If  $c = t_{n-1, \alpha}$  then  $P_\theta(T_{n-1} \geq c) = \alpha$  for any  $\theta = (\mu, \sigma^2)$ . For general  $(\mu, \sigma^2) = \theta \in \Theta_0, \mu \leq 0$  and so

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \geq \frac{\bar{X}}{\sqrt{S^2/n}}.$$

Now  $(\bar{X} - \mu)/\sqrt{S^2/n}$  has a Student's  $t$  distribution with  $n - 1$  degrees of freedom, so for any  $\theta \in \Theta_0$ ,

$$P_\theta(T_{n-1} \geq c) \leq P_\theta\left(\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \geq c\right) = \alpha.$$

Hence the choice of  $c = t_{n-1, \alpha}$  yields a size  $\alpha$  invariant test. ||

**Example 8.3.5:** The problem of finding a size  $\alpha$  union-intersection test in Example 8.2.8 involves finding constants  $t_L$  and  $t_U$  such that

$$\sup_{\theta \in \Theta_0} P_\theta\left(\frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \geq t_L \quad \text{or} \quad \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \leq t_U\right) = \alpha.$$

But for any  $(\mu, \sigma^2) = \theta \in \Theta_0, \mu = \mu_0$  and thus  $(\bar{X} - \mu_0)/\sqrt{S^2/n}$  has a Student's  $t$  distribution with  $n - 1$  degrees of freedom. So any choice of  $t_U = t_{n-1, 1-\alpha_1}$  and  $t_L = t_{n-1, \alpha_2}$ , with  $\alpha_1 + \alpha_2 = \alpha$ , will yield a test with Type I Error probability of exactly  $\alpha$  for all  $\theta \in \Theta_0$ . The usual choice is  $t_L = -t_U = t_{n-1, \alpha/2}$ . ||

After a hypothesis test is done, the conclusions must be reported in some statistically meaningful way. One method of reporting the results of a hypothesis test is to report the size,  $\alpha$ , of the test used and the decision to reject  $H_0$  or accept  $H_0$ . The size of the test carries important information. If  $\alpha$  is small, the decision to reject  $H_0$

is fairly convincing, but if  $\alpha$  is large, the decision to reject  $H_0$  is not very convincing because the test has a large probability of incorrectly making that decision.

Another way of reporting the results of a hypothesis test, one that is data-dependent, is to report the *p-value*. Note that in Examples 8.3.3 and 8.3.4, an entire class of tests are defined, a different test being defined for each value of  $\alpha$ . Furthermore, the rejection region corresponding to a smaller  $\alpha$  is a subset of the rejection region corresponding to a larger  $\alpha$ . Thus, if a sample point  $\mathbf{x}$  is in the rejection region of a size  $\alpha$  test, then it will also be in the rejection region of the test with size  $\alpha' > \alpha$ . The p-value for the sample point  $\mathbf{x}$  is the smallest value of  $\alpha$  for which this sample point will lead to rejection of  $H_0$ . It is important to note, however, that although a p-value is defined in terms of  $\alpha$  levels, a p-value is not an  $\alpha$  level, and should not be interpreted as such: p-values are data-dependent and do not have the error rate interpretation that  $\alpha$  levels have.

Because rejection of  $H_0$  using a test with small size is more convincing evidence that  $H_1$  is true than rejection of  $H_0$  with a test with large size, the interpretation of p-values goes in the same way. The smaller the p-value, the stronger the sample evidence that  $H_1$  is true.

In many situations, the rejection region of a size  $\alpha$  test has the form

$$\text{reject } H_0 \text{ if and only if } W(\mathbf{X}) \geq c_\alpha,$$

where  $W(\mathbf{X})$  is a test statistic appropriate for the problem, and the constant  $c_\alpha$  is chosen so that the test has size  $\alpha$ . In this case, the p-value for the sample point  $\mathbf{x}$  is

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(W(\mathbf{X}) \geq W(\mathbf{x})).$$

For example, consider the first problem in Example 8.3.3. Suppose  $\theta_0 = 10$  and  $\bar{x} = 11.75$  is observed. The p-value is

$$p(\mathbf{x}) = P(|\bar{X} - 10| \geq |11.75 - 10| \mid \theta = 10) = P(|Z| \geq 1.75) = .0802.$$

This p-value is generally interpreted as some slight, but not strong, evidence that  $H_1$  is true. The null hypothesis would be rejected if the test with size  $\alpha = .10$  were used since  $.10 \geq .0802$ . But  $H_0$  would be accepted if the test with size  $\alpha = .05 < .0802$  were used.

Other than  $\alpha$  levels and p-values, there are other features of a test that might also be of concern. For example, we would like a test to be more likely to reject  $H_0$  if  $\theta \in \Theta_0^c$  than if  $\theta \in \Theta_0$ . All of the power functions in Figures 8.3.2 and 8.3.3 have this property, yielding tests that are called unbiased.

**DEFINITION 8.3.4:** A test with power function  $\beta(\theta)$  is *unbiased* if  $\beta(\theta') \geq \beta(\theta'')$  for every  $\theta' \in \Theta_0^c$  and  $\theta'' \in \Theta_0$ .

**Example 8.3.2 (Continued):** An LRT of  $H_0: \theta \leq \theta_0$  versus  $H_1: \theta > \theta_0$  has power function

$$\beta(\theta) = P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right),$$

where  $Z \sim N(0, 1)$ . Since  $\beta(\theta)$  is an increasing function of  $\theta$  (for fixed  $\theta_0$ ), it follows that

$$\beta(\theta) > \beta(\theta_0) = \max_{t \leq \theta_0} \beta(t), \quad \text{for all } \theta > \theta_0,$$

and, hence, that the test is unbiased. ||

In most problems there are many unbiased tests. (See Exercise 8.53.) Likewise, there are many size  $\alpha$  tests, likelihood ratio tests, invariant tests, etc. In some cases we have imposed enough restrictions to narrow consideration to one test. For the two problems in Example 8.3.3, there is only one size  $\alpha$  likelihood ratio test. In other cases there remain many tests from which to choose. In Example 8.3.4 there are many size  $\alpha$ , invariant tests. We discussed only the one that rejects  $H_0$  for large values of  $T$ . In the following sections we will discuss other criteria for selecting one out of a class of tests, criteria that are all related to the power functions of the tests.

### 8.3.2 Most Powerful Tests

In previous sections we have described various classes of hypothesis tests. Some of these classes control the probability of a Type I Error, for example, level  $\alpha$  tests have Type I Error probabilities at most  $\alpha$  for all  $\theta \in \Theta_0$ . A good test in such a class would also have a small Type II Error probability, that is, a large power function for  $\theta \in \Theta_0^c$ . If one test had a smaller Type II Error probability than all other tests in the class, it would certainly be a strong contender for the best test in the class, a notion that is formalized in the next definition.

**DEFINITION 8.3.5:** Let  $\mathcal{C}$  be a class of tests for testing  $H_0 : \theta \in \Theta_0$  versus  $H_1 : \theta \in \Theta_0^c$ . A test in class  $\mathcal{C}$ , with power function  $\beta(\theta)$ , is a *uniformly most powerful (UMP) class  $\mathcal{C}$  test* if  $\beta(\theta) \geq \beta'(\theta)$  for every  $\theta \in \Theta_0^c$  and every  $\beta'(\theta)$  that is a power function of a test in class  $\mathcal{C}$ .

In this section, the class  $\mathcal{C}$  will be the class of *all* level  $\alpha$  tests. The test described in Definition 8.3.5 is then called a UMP level  $\alpha$  test. In Section 8.3.3 the class  $\mathcal{C}$  will be either the class of level  $\alpha$  tests that are also unbiased or the class of level  $\alpha$  tests that are invariant. The names used for these tests, in Section 8.3.3, are the UMP unbiased level  $\alpha$  test and UMP invariant level  $\alpha$  test. For the test described in Definition 8.3.5 to be interesting, restriction to the class  $\mathcal{C}$  must involve some restriction on the Type I Error probability. A minimization of the Type II Error probability without some control of the Type I Error probability is not very interesting. (For example, a test function that rejects  $H_0$  with probability 1 will never make a Type II Error. See Exercise 8.15.)

The requirements in Definition 8.3.5 are so strong that UMP tests do not exist in many realistic problems. But in problems that have UMP tests, a UMP test might

well be considered the best test in the class. Thus, we would like to be able to identify UMP tests if they exist. The following famous theorem clearly describes which tests are UMP level  $\alpha$  tests in the situation where the null and alternative hypotheses both consist of only one probability distribution for the sample (that is, when both  $H_0$  and  $H_1$  are *simple* hypotheses).

**THEOREM 8.3.1 (Neyman–Pearson Lemma):** Consider testing  $H_0: \theta = \theta_0$  versus  $H_1: \theta = \theta_1$ , where the pdf or pmf corresponding to  $\theta_i$  is  $f(\mathbf{x}|\theta_i)$ ,  $i = 0, 1$ , using a test with rejection region  $R$  that satisfies

$$\begin{aligned} \mathbf{x} \in R &\text{ if } f(\mathbf{x}|\theta_1) > kf(\mathbf{x}|\theta_0) \\ (8.3.1) \quad \text{and} \\ \mathbf{x} \in R^c &\text{ if } f(\mathbf{x}|\theta_1) < kf(\mathbf{x}|\theta_0), \end{aligned}$$

for some  $k \geq 0$ , and

$$(8.3.2) \quad \alpha = P_{\theta_0}(X \in R).$$

Then

- a. (*Sufficiency*) Any test that satisfies (8.3.1) and (8.3.2) is a UMP level  $\alpha$  test.
- b. (*Necessity*) If there exists a test satisfying (8.3.1) and (8.3.2) with  $k > 0$  then every UMP level  $\alpha$  test is a size  $\alpha$  test (satisfies (8.3.2)) and every UMP level  $\alpha$  test satisfies (8.3.1) except perhaps on a set  $A$  satisfying  $P_{\theta_0}(X \in A) = P_{\theta_1}(X \in A) = 0$ .

*Proof:* We will prove the theorem for the case that  $f(\mathbf{x}|\theta_0)$  and  $f(\mathbf{x}|\theta_1)$  are pdfs of continuous random variables. The proof for discrete random variables can be accomplished by replacing integrals with sums. (See Exercise 8.24.)

Note first that any test satisfying (8.3.2) is a size  $\alpha$  and, hence, a level  $\alpha$  test because  $\sup_{\theta \in \Theta_0} P_\theta(X \in R) = P_{\theta_0}(X \in R) = \alpha$ , since  $\Theta_0$  has only one point.

Let  $\phi(\mathbf{x})$  be the test function of a test satisfying (8.3.1) and (8.3.2). (Recall Definition 8.2.2.) Let  $\phi'(\mathbf{x})$  be the test function of any other level  $\alpha$  test, and let  $\beta(\theta)$  and  $\beta'(\theta)$  be the power functions corresponding to the tests  $\phi$  and  $\phi'$ , respectively. Because  $0 \leq \phi'(\mathbf{x}) \leq 1$ , (8.3.1) implies that  $(\phi(\mathbf{x}) - \phi'(\mathbf{x}))(f(\mathbf{x}|\theta_1) - kf(\mathbf{x}|\theta_0)) \geq 0$  for every  $\mathbf{x}$  (since  $\phi = 1$  if  $f(\mathbf{x}|\theta_1) > kf(\mathbf{x}|\theta_0)$  and  $\phi = 0$  if  $f(\mathbf{x}|\theta_1) < kf(\mathbf{x}|\theta_0)$ ). Thus

$$(8.3.3) \quad 0 \leq \int [\phi(\mathbf{x}) - \phi'(\mathbf{x})][f(\mathbf{x}|\theta_1) - kf(\mathbf{x}|\theta_0)]d\mathbf{x} = \beta(\theta_1) - \beta'(\theta_1) - k(\beta(\theta_0) - \beta'(\theta_0)).$$

Statement (a) is proved by noting that, since  $\phi'$  is a level  $\alpha$  test and  $\phi$  is a size  $\alpha$  test,  $\beta(\theta_0) - \beta'(\theta_0) = \alpha - \beta'(\theta_0) \geq 0$ . Thus (8.3.3) and  $k \geq 0$  imply that

$$0 \leq \beta(\theta_1) - \beta'(\theta_1) - k(\beta(\theta_0) - \beta'(\theta_0)) \leq \beta(\theta_1) - \beta'(\theta_1),$$

showing that  $\beta(\theta_1) \geq \beta'(\theta_1)$ , and hence  $\phi$  has greater power than  $\phi'$ . Since  $\phi'$  was an arbitrary level  $\alpha$  test and  $\theta_1$  is the only point in  $\Theta_0^c$ ,  $\phi$  is a UMP level  $\alpha$  test.

To prove statement (b), let  $\phi'$  now be the test function for any UMP level  $\alpha$  test. By part (a),  $\phi$ , the test satisfying (8.3.1) and (8.3.2), is also a UMP level  $\alpha$  test, thus  $\beta(\theta_1) = \beta'(\theta_1)$ . This fact, (8.3.3), and  $k > 0$  imply

$$\alpha - \beta'(\theta_0) = \beta(\theta_0) - \beta'(\theta_0) \leq 0.$$

Now, since  $\phi'$  is a level  $\alpha$  test,  $\beta'(\theta_0) \leq \alpha$ . Thus  $\beta'(\theta_0) = \alpha$ , that is,  $\phi'$  is a size  $\alpha$  test, and this also implies that (8.3.3) is an equality in this case. But the nonnegative integrand  $(\phi(\mathbf{x}) - \phi'(\mathbf{x}))(f(\mathbf{x}|\theta_1) - kf(\mathbf{x}|\theta_0))$  will have a zero integral only if  $\phi'$  satisfies (8.3.1), except perhaps on a set  $A$  with  $\int_A f(\mathbf{x}|\theta_i)d\mathbf{x} = 0$ . This implies that the last assertion in statement (b) is true.  $\square$

The following corollaries provide two useful ramifications of the Neyman-Pearson Lemma.

**COROLLARY 8.3.1:** Consider the hypothesis problem posed in Theorem 8.3.1. Suppose  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$  and  $g(t|\theta_i)$  is the pdf or pmf of  $T$  corresponding to  $\theta_i$ ,  $i = 0, 1$ . Then any test based on  $T$  with rejection region  $S$  (a subset of the sample space of  $T$ ) is a UMP level  $\alpha$  test if it satisfies

$$(8.3.4) \quad \begin{aligned} t \in S \text{ if } g(t|\theta_1) &> kg(t|\theta_0) \\ \text{and} \end{aligned}$$

$$t \in S^c \text{ if } g(t|\theta_1) < kg(t|\theta_0),$$

for some  $k \geq 0$ , where

$$(8.3.5) \quad \alpha = P_{\theta_0}(T \in S).$$

*Proof:* In terms of the original sample  $\mathbf{X}$ , the test based on  $T$  has the rejection region  $R = \{\mathbf{x}: T(\mathbf{x}) \in S\}$ . By the Factorization Theorem, the pdf or pmf of  $\mathbf{X}$  can be written as  $f(\mathbf{x}|\theta_i) = g(T(\mathbf{x})|\theta_i)h(\mathbf{x})$ ,  $i = 0, 1$ , for some nonnegative function  $h(\mathbf{x})$ . Multiplying the inequalities in (8.3.4) by this nonnegative function, we see that  $R$  satisfies

$$\mathbf{x} \in R \text{ if } f(\mathbf{x}|\theta_1) = g(T(\mathbf{x})|\theta_1)h(\mathbf{x}) > kg(T(\mathbf{x})|\theta_0)h(\mathbf{x}) = kf(\mathbf{x}|\theta_0)$$

and

$$\mathbf{x} \in R^c \text{ if } f(\mathbf{x}|\theta_1) = g(T(\mathbf{x})|\theta_1)h(\mathbf{x}) < kg(T(\mathbf{x})|\theta_0)h(\mathbf{x}) = kf(\mathbf{x}|\theta_0).$$

Also, by (8.3.5),

$$P_{\theta_0}(\mathbf{X} \in R) = P_{\theta_0}(T(\mathbf{X}) \in S) = \alpha.$$

So, by the sufficiency part of the Neyman–Pearson Lemma, the test based on  $T$  is a UMP level  $\alpha$  test.  $\square$

Hypotheses, such as  $H_0$  and  $H_1$  in the Neyman–Pearson Lemma, that specify only one possible distribution for the sample  $X$  are called simple hypotheses. In most realistic problems, the hypotheses of interest specify more than one possible distribution for the sample. Such hypotheses are called *composite hypotheses*. Since Definition 8.3.5 requires a UMP test to be most powerful against *each* individual  $\theta \in \Theta_0^c$ , the Neyman–Pearson Lemma is usually used to find UMP tests in problems involving composite hypotheses. Corollary 8.3.2 indicates how this may be done.

**COROLLARY 8.3.2:** Consider testing  $H_0: \theta \in \Theta_0$  versus  $H_1: \theta \in \Theta_0^c$ . Suppose a test based on a sufficient statistic  $T$  with rejection region  $S$  satisfies the following three conditions.

- a. The test is a level  $\alpha$  test.
- b. There exists a  $\theta_0 \in \Theta_0$  such that  $P_{\theta_0}(T \in S) = \alpha$ .
- c. Let  $g(t|\theta)$  denote the pdf or pmf of  $T$ . For the same  $\theta_0$  as in (b), and for each  $\theta' \in \Theta_0^c$ , there exists a  $k' \geq 0$  such that

$$t \in S \text{ if } g(t|\theta') > k'g(t|\theta_0) \quad \text{and} \quad t \in S^c \text{ if } g(t|\theta') < k'g(t|\theta_0).$$

Then this test is a UMP level  $\alpha$  test of  $H_0$  versus  $H_1$ .

*Proof:* Let  $\beta(\theta)$  be the power function of the test with rejection region  $S$ . Fix  $\theta' \in \Theta_0^c$ . Consider testing  $H'_0: \theta = \theta_0$  versus  $H'_1: \theta = \theta'$ . Corollary 8.3.1 and (a), (b), and (c) imply  $\beta(\theta') \geq \beta^*(\theta')$  where  $\beta^*(\theta)$  is the power function for any other level  $\alpha$  test of  $H'_0$ , that is, any test satisfying  $\beta(\theta_0) \leq \alpha$ . However, any level  $\alpha$  test of  $H_0$  satisfies  $\beta^*(\theta_0) \leq \sup_{\theta \in \Theta_0} \beta^*(\theta) \leq \alpha$ . Thus,  $\beta(\theta') \geq \beta^*(\theta')$  for any level  $\alpha$  test of  $H_0$ . Since  $\theta'$  was arbitrary, the result follows.  $\square$

Notice that conditions (a) and (b) of the corollary imply not just that the test is a size  $\alpha$  test, but something stronger. A test can be size  $\alpha$  and not have any parameter point in  $H_0$  that attains power  $\alpha$ , since the size is the *supremum* of the power. Conditions (a) and (b) ensure that this supremum is, in fact, a *maximum*, that is, it is attained for a finite parameter value.

When deriving a test that satisfies the inequalities (8.3.1) or (8.3.4), and hence is a UMP level  $\alpha$  test, it is usually easier to rewrite the inequalities as  $f(x|\theta_1)/f(x|\theta_0) > k$ . (We must be careful about dividing by zero.) This method is used in the following examples.

**Example 8.3.6:** Let  $X \sim \text{binomial}(2, \theta)$ . We want to test  $H_0: \theta = \frac{1}{2}$  versus  $H_1: \theta = \frac{3}{4}$ . Calculating the ratios of the pmfs gives

$$\frac{f(0|\theta = \frac{3}{4})}{f(0|\theta = \frac{1}{2})} = \frac{1}{4}, \quad \frac{f(1|\theta = \frac{3}{4})}{f(1|\theta = \frac{1}{2})} = \frac{3}{4}, \quad \text{and} \quad \frac{f(2|\theta = \frac{3}{4})}{f(2|\theta = \frac{1}{2})} = \frac{9}{4}.$$

SECTION 8.3 Methods of Evaluating Tests

369

By choosing  $\frac{3}{4} < k < \frac{9}{4}$  the Neyman-Pearson Lemma says that the test that rejects  $H_0$  if  $X = 2$  is the UMP level  $\alpha = P(X = 2|\theta = \frac{1}{2}) = \frac{1}{4}$  test. By choosing  $k < \frac{1}{4}$  or  $k > \frac{9}{4}$  yields point if  $H_0$  if the Neyman-Pearson Lemma says that the test that rejects  $H_0$  if  $X = 1$  or  $X = 2$  is the UMP level  $\alpha = P(X = 1 \text{ or } 2|\theta = \frac{1}{2}) = \frac{3}{4}$  test. Choosing  $k < \frac{1}{4}$  or  $k > \frac{9}{4}$  yields point if  $H_0$  if UMP level  $\alpha = 1$  or level  $\alpha = 0$  test.

Note that if  $k = \frac{3}{4}$ , then (8.3.1) says we must reject  $H_0$  for  $x = 1$  undetermined. If we accept  $H_0$  for  $x = 2$  and accept  $H_0$  for  $x = 0$  but leaves our action for  $x = 1$  we get the UMP level  $\alpha = \frac{1}{4}$  test as above. If we accept  $H_0$  for  $x = 1$  we get the UMP level  $\alpha = \frac{3}{4}$  test as above.

The above example also shows that when dealing with a discrete pmf with level at any one we are dealing. (No such problem arises in the continuous case. Any  $\alpha$  test at these are attained.) A method of dealing with this situation and producing tests, while usefulness in practice is suspect.

**Example 8.3.7:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population,  $\sigma^2$  known. The sample mean  $\bar{X}$  is a sufficient statistic for  $\theta$ . Consider testing  $H_0: \theta = \theta_0$  versus  $H_1: \theta = \theta_1$  where  $\theta_0 > \theta_1$ . The inequality (8.3.4),  $g(\bar{x}|\theta_1) > g(\bar{x}|\theta_0)$ , is equivalent to

$$\bar{x} < \frac{(2\sigma^2 \log k)/n - \theta_0^2 + \theta_1^2}{2(\theta_1 - \theta_0)}.$$

The fact that  $\theta_1 - \theta_0 < 0$  was used to obtain this inequality. The right-hand side increases from  $-\infty$  to  $\infty$  as  $k$  increases from 0 to  $\infty$ . Thus, by Corollary 8.3.1, the test with rejection region  $\bar{x} < c$  is the UMP level  $\alpha$  test where  $c = P_{\theta_0}(\bar{X} \leq c) = -\sigma z_\alpha / \sqrt{n} + \theta_0$ . If a particular  $\alpha$  is specified, then the UMP test rejects  $H_0$  if  $\bar{X} \leq c$ . This choice of  $c$  ensures that (8.3.5) is true.

Now consider testing  $H'_0: \theta \geq \theta_0$  versus  $H'_1: \theta < \theta_0$ . The test that rejects  $H'_0$  if

$$\bar{X} < -\frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0$$

is also a UMP level  $\alpha$  test in this problem. Condition (b) of Corollary 8.3.2 is obviously satisfied. Condition (c) is true because, in the above argument, only the fact that  $\theta_1 < \theta_0$ , not the exact value of  $\theta_1$ , was used in determining the UMP level  $\alpha$  test of  $H_0: \theta = \theta_0$  versus  $H_1: \theta = \theta_1$ . Condition (a) is true because the power function of this test,

$$\beta(\theta) = P_\theta \left( \bar{X} < -\frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0 \right),$$

is a decreasing function of  $\theta$  since  $\theta$  is a location parameter in the distribution of  $\bar{X}$ . Thus  $\sup_{\theta \geq \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha$  and the test is a level  $\alpha$  test.  $\square$

Hypotheses that assert that a univariate parameter is large, for example,  $H: \theta \geq \theta_0$ , or small, for example,  $H: \theta < \theta_0$ , are called *one-sided hypotheses*. Hypotheses that assert that a parameter is either large or small, for example,  $H: \theta \neq \theta_0$ , are called *two-sided hypotheses*. A large class of problems that admit UMP level  $\alpha$  tests involve one-sided hypotheses and pdfs or pmfs with the monotone likelihood ratio property.

**DEFINITION 8.3.6:** A family of pdfs or pmfs  $\{g(t|\theta) : \theta \in \Theta\}$  for a univariate random variable  $T$  with real-valued parameter  $\theta$  has a *monotone likelihood ratio* (MLR) if, for every  $\theta_2 > \theta_1$ ,  $g(t|\theta_2)/g(t|\theta_1)$  is a nondecreasing function of  $t$  on  $\{t : g(t|\theta_1) > 0 \text{ or } g(t|\theta_2) > 0\}$ . Note that  $c/0$  is defined as  $\infty$  if  $0 < c$ .

Many common families of distributions have an MLR. For example, the normal (known variance, unknown mean), Poisson, and binomial all have an MLR. Indeed, any regular exponential family with  $g(t|\theta) = h(t)c(\theta)e^{w(\theta)t}$  has an MLR if  $w(\theta)$  is a nondecreasing function (see Exercise 8.28).

**THEOREM 8.3.2 (Karlin–Rubin):** Consider testing  $H_0 : \theta \leq \theta_0$  versus  $H_1 : \theta > \theta_0$ . Suppose that  $T$  is a sufficient statistic for  $\theta$  and the family of pdfs or pmfs  $\{g(t|\theta) : \theta \in \Theta\}$  of  $T$  has an MLR. Then for any  $t_0$ , the test that rejects  $H_0$  if and only if  $T > t_0$  is a UMP level  $\alpha$  test where  $\alpha = P_{\theta_0}(T > t_0)$ .

*Proof:* Since the family of pdfs or pmfs of  $T$  has an MLR, the power function  $\beta(\theta) = P_\theta(T > t_0)$  is nondecreasing (Exercise 8.37). So  $\sup_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha$ , and this is a level  $\alpha$  test. Condition (b) of Corollary 8.3.2 is true by assumption. Condition (c) can be verified using

$$k' = \inf_{t \in T} \frac{g(t|\theta')}{g(t|\theta_0)}$$

where  $T = \{t : t > t_0 \text{ and either } g(t|\theta') > 0 \text{ or } g(t|\theta_0) > 0\}$ . Thus by Corollary 8.3.2, the test is a UMP level  $\alpha$  test.  $\square$

By an analogous argument, it can be shown that under the conditions of Theorem 8.3.2, the test that rejects  $H_0 : \theta \geq \theta_0$  in favor of  $H_1 : \theta < \theta_0$  if and only if  $T < t_0$  is a UMP level  $\alpha = P_{\theta_0}(T < t_0)$  test. The test in Example 8.3.7 is of this form.

### 8.3.3 Unbiased and Invariant Tests

In Section 8.3.2, we discussed uniformly most powerful (UMP) level  $\alpha$  tests. A UMP level  $\alpha$  test is a good test in that, when compared with other level  $\alpha$  tests, it has the highest possible power, that is, the lowest possible Type II Error probability, at *every* parameter point specified in the alternative hypothesis. Most experimenters would choose to use a UMP level  $\alpha$  test if they knew of one.

Unfortunately, for many problems there is no UMP level  $\alpha$  test. That is, no UMP test exists because the class of level  $\alpha$  tests is so large that no one test dominates all the others in terms of power. In such cases, a common method of continuing the search for a good test is to consider some subset of the class of level  $\alpha$  tests and attempt to find a UMP test in this subset. This tactic should be reminiscent of what we did in Chapter 7, when we restricted attention to unbiased point estimators in order to investigate optimality. In this section we first restrict attention to the subset consisting of unbiased tests, and later to the subset consisting of invariant tests.

First we consider an example that illustrates a typical situation in which a UMP level  $\alpha$  test does not exist.

**Example 8.3.8:** Let  $X_1, \dots, X_n$  be iid  $n(\theta, \sigma^2)$ ,  $\sigma^2$  known. Consider testing  $H_0: \theta = \theta_0$  versus  $H_1: \theta \neq \theta_0$ . For a specified value of  $\alpha$ , a level  $\alpha$  test in this problem is any test that satisfies

$$(8.3.6) \quad P_{\theta_0}(\text{reject } H_0) \leq \alpha.$$

Consider an alternative parameter point  $\theta_1 < \theta_0$ . The analysis in Example 8.3.7 shows that, among all tests that satisfy (8.3.6), the test that rejects  $H_0$  if  $\bar{X} < -\sigma z_\alpha / \sqrt{n} + \theta_0$  has the highest possible power at  $\theta_1$ . Call this Test 1. Furthermore, by part (b) (necessity) of the Neyman–Pearson Lemma, any other level  $\alpha$  test that has as high a power as Test 1 at  $\theta_1$  must have the same rejection region as Test 1 except possibly for a set  $A$  satisfying  $\int_A f(x|\theta_i)dx = 0$ . Thus, if a UMP level  $\alpha$  test exists for this problem, it must be Test 1 because no other test has as high a power as Test 1 at  $\theta_1$ .

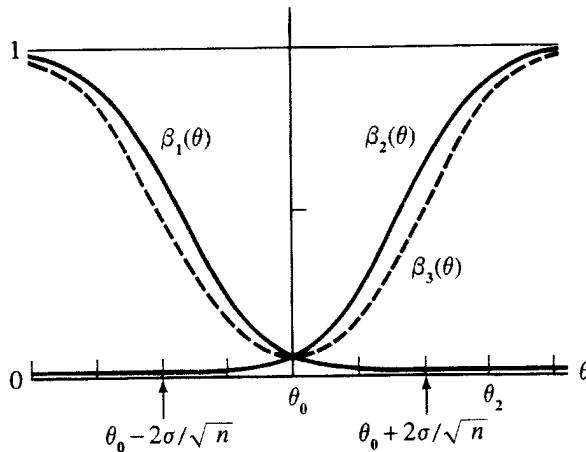
Now consider Test 2, which rejects  $H_0$  if  $\bar{X} > \sigma z_\alpha / \sqrt{n} + \theta_0$ . Test 2 is also a level  $\alpha$  test. Let  $\beta_i(\theta)$  denote the power function of Test  $i$ . For any  $\theta_2 > \theta_0$ ,

$$\begin{aligned} \beta_2(\theta_2) &= P_{\theta_2}\left(\bar{X} > \frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0\right) \\ &= P_{\theta_2}\left(\frac{\bar{X} - \theta_2}{\sigma/\sqrt{n}} > z_\alpha + \frac{\theta_0 - \theta_2}{\sigma/\sqrt{n}}\right) \\ &> P(Z > z_\alpha) && \left( \begin{array}{l} Z \sim n(0, 1) \\ > \text{since } \theta_0 - \theta_2 < 0. \end{array} \right) \\ &= P(Z < -z_\alpha) \\ &> P_{\theta_2}\left(\frac{\bar{X} - \theta_2}{\sigma/\sqrt{n}} < -z_\alpha + \frac{\theta_0 - \theta_2}{\sigma/\sqrt{n}}\right) && \left( \begin{array}{l} \text{again, } > \text{since} \\ \theta_0 - \theta_2 < 0. \end{array} \right) \\ &= P_{\theta_2}\left(\bar{X} < -\frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0\right) \\ &= \beta_1(\theta_2). \end{aligned}$$

Thus Test 1 is not a UMP level  $\alpha$  test because Test 2 has a higher power than Test 1 at  $\theta_2$ . Earlier we showed that if there were a UMP level  $\alpha$  test, it would have to be Test 1. Therefore, no UMP level  $\alpha$  test exists in this problem. ||

Example 8.3.8 illustrates again the usefulness of the Neyman–Pearson Lemma. In Section 8.3.2, the *sufficiency* part of the Lemma was used to construct UMP level  $\alpha$  tests, but to show the nonexistence of a UMP level  $\alpha$  test, the *necessity* part of the Lemma is used.

When no UMP level  $\alpha$  test exists within the class of all tests, we might try to find a UMP level  $\alpha$  test within the class of unbiased tests. The power function of an unbiased test,  $\beta_3(\theta)$ , as well as  $\beta_1(\theta)$  and  $\beta_2(\theta)$  from Example 8.3.8, is shown in Figure 8.3.4. Note that although Test 1 and Test 2 have slightly higher powers than the unbiased test for some parameter points, the unbiased test has much higher power than Test 1 and Test 2 at other parameter points. For example,  $\beta_3(\theta_2)$  is near one whereas  $\beta_1(\theta_2)$  is near zero. If the interest is in rejecting  $H_0$  for both large and small values of  $\theta$ , Figure 8.3.4 shows that the unbiased test is better overall than either Test 1 or Test 2.



**FIGURE 8.3.4** Power functions for three tests in Example 8.3.8.  $\beta_3(\theta)$  is the power function of the UMP unbiased level  $\alpha = .05$  test.

The Neyman–Pearson Lemma describes how to maximize a sum or integral (the power) subject to one constraint (level  $\alpha$ ). To find a UMP unbiased, level  $\alpha$  test, we need to maximize a sum or integral (the power) subject to two constraints (unbiasedness and level  $\alpha$ ). We use the following partial generalization of the Neyman–Pearson Lemma that considers multiple constraints. A more complete generalization of the Neyman–Pearson Lemma, that includes both necessary and sufficient conditions, may be found in Lehmann (1986).

**THEOREM 8.3.3:** Let  $c_1, \dots, c_m$  be constants and  $f_1(x), \dots, f_{m+1}(x)$  be real-valued functions. Let  $\mathcal{C}$  be the class of functions  $\phi(x)$  satisfying  $0 \leq \phi(x) \leq 1$  for every  $x$  and

$$(8.3.7) \quad \int \phi(x) f_i(x) dx = c_i, \quad i = 1, \dots, m.$$

If  $\phi^*(x)$  is a function in  $\mathcal{C}$  that satisfies

$$\phi^*(\mathbf{x}) = 1 \quad \text{if } f_{m+1}(\mathbf{x}) > \sum_{i=1}^m k_i f_i(\mathbf{x})$$

(8.3.8) and

$$\phi^*(\mathbf{x}) = 0 \quad \text{if } f_{m+1}(\mathbf{x}) < \sum_{i=1}^m k_i f_i(\mathbf{x})$$

for some constants  $k_1, \dots, k_m$ , then  $\phi^*(\mathbf{x})$  maximizes  $\int \phi(\mathbf{x}) f_{m+1}(\mathbf{x}) d\mathbf{x}$  among all functions  $\phi$  in  $\mathcal{C}$ .

*Proof:* Let  $\phi^*(\mathbf{x})$  be a function in  $\mathcal{C}$  satisfying (8.3.8) and let  $\phi(\mathbf{x})$  be any other function in  $\mathcal{C}$ . Because  $0 \leq \phi(\mathbf{x}) \leq 1$  for every  $\mathbf{x}$ , (8.3.8) implies (similar to the proof of the Neyman–Pearson Lemma) that

$$[\phi^*(\mathbf{x}) - \phi(\mathbf{x})][f_{m+1}(\mathbf{x}) - \sum_{i=1}^m k_i f_i(\mathbf{x})] \geq 0 \quad \text{for every } \mathbf{x}.$$

Thus

$$\begin{aligned} 0 &\leq \int [\phi^*(\mathbf{x}) - \phi(\mathbf{x})][f_{m+1}(\mathbf{x}) - \sum_{i=1}^m k_i f_i(\mathbf{x})] d\mathbf{x} \\ &= \int \phi^*(\mathbf{x}) f_{m+1}(\mathbf{x}) d\mathbf{x} - \int \phi(\mathbf{x}) f_{m+1}(\mathbf{x}) d\mathbf{x} \\ &\quad - \sum_{i=1}^m k_i \left[ \int \phi^*(\mathbf{x}) f_i(\mathbf{x}) d\mathbf{x} - \int \phi(\mathbf{x}) f_i(\mathbf{x}) d\mathbf{x} \right]. \end{aligned}$$

But every term,  $[\int \phi^*(\mathbf{x}) f_i(\mathbf{x}) d\mathbf{x} - \int \phi(\mathbf{x}) f_i(\mathbf{x}) d\mathbf{x}]$ , is zero since  $\phi^*(\mathbf{x})$  and  $\phi(\mathbf{x})$  both satisfy (8.3.7). Hence

$$\int \phi^*(\mathbf{x}) f_{m+1}(\mathbf{x}) d\mathbf{x} - \int \phi(\mathbf{x}) f_{m+1}(\mathbf{x}) d\mathbf{x} \geq 0,$$

as was to be shown.  $\square$

Note that if  $m = 1$  and we let  $f_1(\mathbf{x}) = f(\mathbf{x}|\theta_0)$  and  $f_2(\mathbf{x}) = f(\mathbf{x}|\theta_1)$  we then have the Neyman–Pearson Lemma. Furthermore, the Neyman–Pearson Lemma can be made to apply to general functions, not just to pdfs or pmfs.

The class of functions  $\phi(\mathbf{x})$  in Theorem 8.3.3 is actually larger than the class of test functions in Definition 8.2.2. The test functions took on only the values zero and one, while the functions of Theorem 8.3.3 were not so constrained. The above functions  $\phi(\mathbf{x})$ , satisfying  $0 \leq \phi(\mathbf{x}) \leq 1$ , are *randomized test functions*, as mentioned after Example 8.3.6. These can have the interpretation that  $\phi(\mathbf{x})$  is the probability that

$H_0$  is rejected if the sample  $\bar{x}$  is observed, and  $1 - \phi(\bar{x})$  is the probability that  $H_0$  is accepted if the sample  $\bar{x}$  is observed. If  $\bar{x}$  is observed and  $\phi(\bar{x}) = .8$ , for example, an independent Bernoulli experiment with  $p = .8$  is conducted and  $H_0$  is rejected if the Bernoulli experiment results in a “success.” Randomized tests are mainly of theoretical interest, because most researchers do not want the result of an experiment that is independent of the sample to determine their decision regarding  $H_0$ . However, randomized tests can be useful theoretical tools, allowing simple statements of some important results like Theorem 8.3.3, for example.

Now we use Theorem 8.3.3 to find a UMP unbiased level  $\alpha$  test for the hypothesis testing problem posed in Example 8.3.8.

**Example 8.3.9:** For the problem in Example 8.3.8, we now show that Test 3, which rejects  $H_0: \theta = \theta_0$  in favor of  $H_1: \theta \neq \theta_0$  if and only if  $\bar{X} > \sigma z_{\alpha/2}/\sqrt{n} + \theta_0$  or  $\bar{X} < -\sigma z_{\alpha/2}/\sqrt{n} + \theta_0$ , is a UMP unbiased level  $\alpha$  test, that is, it is UMP in the class of unbiased tests.

By an argument similar to that in Corollary 8.3.1, to find a test satisfying (8.3.8), it suffices to consider the sufficient statistic  $\bar{X}$  and its pdf in constructing a UMP unbiased test. Let  $\tau^2 = \sigma^2/n$  be the known variance of  $\bar{X}$  and  $f(\bar{x}|\theta)$  the pdf of  $\bar{X}$ . We first use Theorem 8.3.3 to show that Test 3 is the UMP test among all tests that satisfy

$$(8.3.9) \quad \int \phi(\bar{x}) f(\bar{x}|\theta_0) d\bar{x} = \alpha$$

and

$$(8.3.10) \quad \int \phi(\bar{x}) \left( \frac{\partial}{\partial \theta} f(\bar{x}|\theta) \Big|_{\theta=\theta_0} \right) d\bar{x} = 0.$$

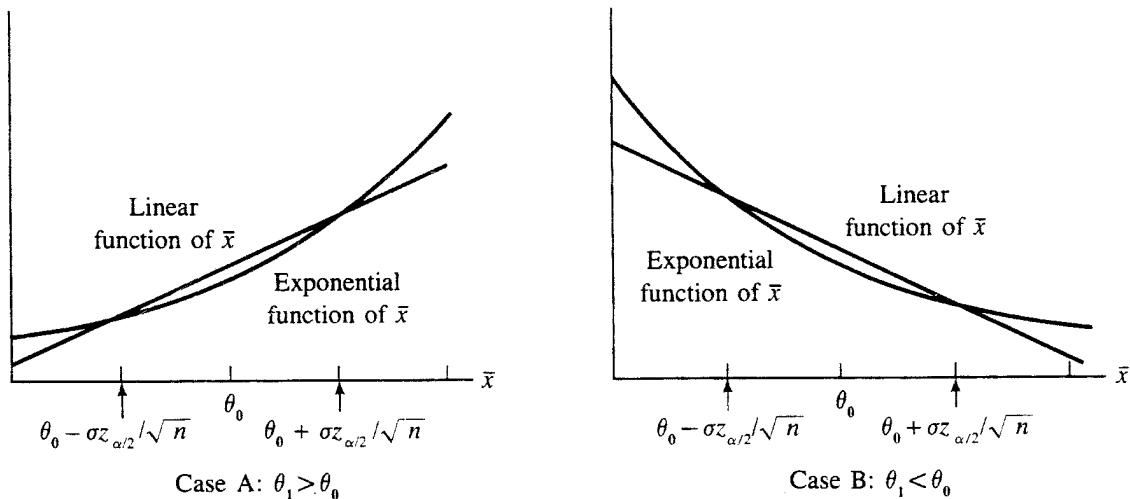
Thus  $f(\bar{x}|\theta_0)$  and  $\frac{\partial}{\partial \theta} f(\bar{x}|\theta) \Big|_{\theta=\theta_0}$  will play the roles of  $f_1$  and  $f_2$  of Theorem 8.3.3.

Before we proceed, note that (8.3.9) and (8.3.10) do not define the class of level  $\alpha$ , unbiased tests; a little more is needed. Consider what these restrictions imply. First, (8.3.9) restricts attention to size  $\alpha$  tests so we will have to argue further to make the extension to level  $\alpha$  tests. Equation (8.3.10) is related to the unbiasedness constraint. For any unbiased test the power function satisfies  $\beta(\theta_1) \geq \beta(\theta_0)$  for all  $\theta_1 \neq \theta_0$ , that is,  $\theta_0$  is a minimum of the power function for any unbiased test. Since  $f(\bar{x}|\theta)$  is an exponential family, the power function of any test is differentiable and we can evaluate the derivative by differentiating under the integral sign (see Exercise 8.43). The power function for any unbiased test satisfies  $\frac{d}{d\theta} \beta(\theta) \Big|_{\theta=\theta_0} = 0$  since  $\theta_0$  is a minimum. Equation (8.3.10) defines the class of tests whose power function has a zero derivative at  $\theta_0$ . Thus the class of tests that satisfy (8.3.10) contains the class of unbiased tests. If we show that Test 3 is a UMP level  $\alpha$  test among all tests satisfying (8.3.10) and that Test 3 is unbiased, then we can conclude that Test 3 is a UMP unbiased level  $\alpha$  test.

Fix  $\theta_1 \neq \theta_0$ . We now show that Test 3 satisfies (8.3.8) for  $f_3(\bar{x}) = f(\bar{x}|\theta_1)$ ,  $f_2(\bar{x}) = \frac{\partial}{\partial \theta} f(\bar{x}|\theta) \Big|_{\theta=\theta_0}$ , and  $f_1(\bar{x}) = f(\bar{x}|\theta_0)$ . For this choice of  $f_1$ ,  $f_2$ , and  $f_3$  the first inequality in (8.3.8) is equivalent to

$$(8.3.11) \quad \exp\left(\frac{\bar{x}(\theta_1 - \theta_0)}{\tau^2} - \frac{(\theta_1^2 - \theta_0^2)}{2\tau^2}\right) > k_1 - k_2 \frac{\theta_0}{2\tau^2} + k_2 \frac{\bar{x}}{2\tau^2}.$$

The exponential function of  $\bar{x}$  on the left-hand side of (8.3.11) is increasing or decreasing depending on whether  $\theta_1 > \theta_0$  or  $\theta_1 < \theta_0$ , but in either case the constants  $k_1$  and  $k_2$  can be chosen so that the linear function on the right-hand side of (8.3.11) crosses the exponential function on the left-hand side at the two points  $\bar{x} = \sigma z_{\alpha/2}/\sqrt{n} + \theta_0$  and  $\bar{x} = -\sigma z_{\alpha/2}/\sqrt{n} + \theta_0$  (Figure 8.3.5). For this choice of  $k_1$  and  $k_2$ , since the exponential function is a convex function, the exponential function will be greater than the linear function for  $\bar{x} > \sigma z_{\alpha/2}/\sqrt{n} + \theta_0$  and  $\bar{x} < -\sigma z_{\alpha/2}/\sqrt{n} + \theta_0$ . That is, (8.3.11) and the first inequality in (8.3.8) are true for exactly the  $\bar{x}$  in the rejection region of Test 3. It is easily verified that Test 3 satisfies (8.3.9) and (8.3.10) (Exercise 8.55). Thus, by Theorem 8.3.3, Test 3 maximizes  $\int \phi(\bar{x})f(\bar{x}|\theta_1)d\bar{x}$ , the power at  $\theta_1$ , among all tests satisfying (8.3.9) and (8.3.10). Furthermore, the test function  $\phi^*(\bar{x}) = \alpha$  also satisfies (8.3.9) and (8.3.10). So the power function of Test 3 must satisfy  $\beta_3(\theta_1) = \int \phi_3(\bar{x})f(\bar{x}|\theta_1)d\bar{x} \geq \int \phi^*(\bar{x})f(\bar{x}|\theta_1)d\bar{x} = \alpha$ . Since  $\theta_1$  was arbitrary, this says that Test 3 is an unbiased test. Thus we have shown that Test 3 is the UMP unbiased size  $\alpha$  test.



**FIGURE 8.3.5**  $k_1$  and  $k_2$  can be chosen so the linear function crosses the exponential function at  $\theta_0 + \sigma z_{\alpha/2}/\sqrt{n}$  and  $\theta_0 - \sigma z_{\alpha/2}/\sqrt{n}$  in Example 8.3.9.

To extend this to level  $\alpha$  tests, we note that by the same argument, for any  $\alpha' < \alpha$ , the UMP unbiased size  $\alpha'$  test is the test that rejects  $H_0$  if and only if  $\bar{X} > \sigma z_{\alpha'/2}/\sqrt{n} + \theta_0$  or  $\bar{X} < -\sigma z_{\alpha'/2}/\sqrt{n} + \theta_0$ . But  $-z_{\alpha'/2} < -z_{\alpha/2}$  and  $z_{\alpha'/2} > z_{\alpha/2}$  so this rejection region is contained in the rejection region for Test 3. Therefore, the power function for Test 3 is greater than the power function for the UMP unbiased size  $\alpha'$  test for every value of  $\theta$ . The power function of the UMP unbiased size  $\alpha'$  test, in turn, is greater than the power function for any other unbiased size  $\alpha'$  test. Therefore Test 3 is the UMP unbiased level  $\alpha$  test. ||

Perhaps a more realistic model is the normal model, like Example 8.3.8, in which the variance  $\sigma^2$  is unknown. It can be shown that the  $t$  test that rejects  $H_0: \theta = \theta_0$

in favor of  $H_1: \theta \neq \theta_0$  if and only if  $|\bar{X} - \theta_0|/(S/\sqrt{n}) > t_{\alpha/2}$  is the UMP unbiased level  $\alpha$  test. Theorem 8.3.3 can be used in the proof of this as it was in Example 8.3.9 (Exercise 8.48). See Lehmann (1986), Section 5.2.

Thus far in this section we have restricted attention to level  $\alpha$  tests that were also unbiased and looked for a UMP test in this class. We now consider a different restriction, to level  $\alpha$  tests that are also invariant, and look for a UMP test in this class. Recall that invariant tests were introduced in Section 8.2.2.

Searching for UMP invariant tests is a rather straightforward proposition, given all of the techniques we have for finding UMP tests. Using the invariance restriction, we can reduce the class of tests under consideration. Then, within this reduced class, we look for the UMP test.

In Example 8.2.6, the class of invariant tests was characterized as those that depended on the sample only through the statistic  $T = \bar{X}/(S/\sqrt{n})$ . In fact, such a characterization in terms of a statistic can always be made (see Lehmann (1986)). Having made such a characterization, the results of the previous sections can be used to search for a UMP invariant level  $\alpha$  test. The statistic  $T$  plays the role of the sample  $x$  in the Neyman–Pearson Lemma and a sufficient statistic based on the family of distributions of  $T$ ,  $\{f(t|\theta) : \theta \in \Theta\}$ , frequently  $T$  itself, is used in Corollaries 8.3.1 and 8.3.2 and Theorem 8.3.2.

**Example 8.3.10:** Consider the problem discussed in Example 8.2.6. The statistic  $T = \bar{X}/(S/\sqrt{n})$  has a *noncentral t distribution* with  $n - 1$  degrees of freedom and noncentrality parameter  $\delta = \sqrt{n}\mu/\sigma$ . The noncentral  $t$  family of distributions has an MLR in the parameter  $\delta$ . (See Exercise 8.38 for more about the noncentral  $t$  distribution.) Furthermore, the hypotheses  $H_0: \mu \leq 0$  and  $H_1: \mu > 0$  are equivalent to  $H'_0: \delta \leq 0$  and  $H'_1: \delta > 0$ . By Theorem 8.3.2, the test that rejects  $H'_0$  if and only if  $T > t_0$  is a UMP level  $\alpha = P(T > t_0 | \mu = \delta = 0)$  test based on  $T$ . That is, this test is a UMP invariant level  $\alpha$  test. To achieve a specified level  $\alpha$ ,  $t_0 = t_{n-1,\alpha}$  is used, as explained in Example 8.3.4. ||

### 8.3.4 Locally Most Powerful Tests

In the previous section we used the following rationale: If a UMP test did not exist, then we could restrict the class of tests and look for a test in this class that had high power at all alternative parameter values. Another way to handle the situation in which no UMP test exists is to restrict the alternative parameter values we consider. That is, we could look for tests that have high power at some particular alternatives.

In most problems, alternative parameter values that are close to the null hypothesis are hard to detect. Thus, it is important to use a test that has as high a power as possible for alternative parameter values that are close to the null hypothesis. A locally most powerful test is one that maximizes the power for alternative values that are close to  $H_0$ .

**DEFINITION 8.3.7:** A test with power function  $\beta(\theta)$  is a *locally most powerful (LMP) test* for testing  $H_0: \theta \leq \theta_0$  versus  $H_1: \theta > \theta_0$  if, for any other test with power

function  $\beta'(\theta)$  satisfying  $\beta(\theta_0) = \beta'(\theta_0)$ , there exists  $\Delta > 0$  such that  $\beta(\theta) \geq \beta'(\theta)$  for every  $\theta$  satisfying  $\theta_0 < \theta \leq \theta_0 + \Delta$ .

Usually, the tests we consider have differentiable power functions. In such a case, an LMP test will maximize  $\frac{d}{d\theta}\beta(\theta)|_{\theta=\theta_0}$ . If the derivative may be taken inside the integral sign, then an LMP level  $\alpha$  test will maximize

$$(8.3.12) \quad \frac{d}{d\theta}\beta(\theta)\Big|_{\theta=\theta_0} = \int \phi(x) \frac{\partial}{\partial\theta} f(x|\theta)\Big|_{\theta=\theta_0} dx$$

subject to the size constraint

$$(8.3.13) \quad \int \phi(x) f(x|\theta_0) dx = \alpha.$$

Theorem 8.3.3 (the Generalized Neyman–Pearson Lemma) implies that if a test satisfies (8.3.13) and is defined by

$$(8.3.14) \quad \phi(x) = \begin{cases} 1 & \text{if } \frac{\partial}{\partial\theta} f(x|\theta)|_{\theta=\theta_0} > kf(x|\theta_0) \\ 0 & \text{if } \frac{\partial}{\partial\theta} f(x|\theta)|_{\theta=\theta_0} < kf(x|\theta_0) \end{cases},$$

then it will maximize (8.3.12). Furthermore, the test that maximizes (8.3.12) is unique. (It is conceivable that two size  $\alpha$  tests have the same derivative at  $\theta_0$  but different power functions.) Although a test might maximize (8.3.12) and not be the LMP test, if the test that maximizes (8.3.12) is unique then it must be the LMP test.

Before considering an example, note that the first inequality in (8.3.14),  $\frac{\partial}{\partial\theta} f(x|\theta)|_{\theta=\theta_0} > kf(x|\theta_0)$ , can be written as

$$\frac{\partial}{\partial\theta} \log f(x|\theta)\Big|_{\theta=\theta_0} > \log k$$

provided, of course, that  $f(x|\theta) > 0$  so the log function is well defined. Furthermore, if  $X_1, \dots, X_n$  are a random sample from a population with pdf or pmf  $f(x|\theta)$ , then this inequality can be written as

$$\sum_{i=1}^n \frac{\partial}{\partial\theta} \log f(x_i|\theta)\Big|_{\theta=\theta_0} > \log k.$$

These facts may simplify the derivation of an LMP test in some cases.

**Example 8.3.11:** Suppose that  $X_1, \dots, X_n$  are iid with pdf

$$f(x|\theta) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)} \left( \frac{(x-\theta)^2 + \nu}{\nu} \right)^{-(\nu+1)/2},$$

a  $t$  distribution with  $\nu$  degrees of freedom and location parameter  $\theta$ . It is easily calculated that

$$\frac{\partial}{\partial \theta} \log f(x|\theta) = \frac{(\nu + 1)(x - \theta)}{(x - \theta)^2 + \nu}.$$

Thus the test with rejection region consisting of the  $\mathbf{x}$ s that satisfy

$$\sum_{i=1}^n \frac{(\nu + 1)(x_i - \theta_0)}{(x_i - \theta_0)^2 + \nu} > k'$$

is an LMP test.

Although this test is LMP it does have its faults. For example, the power function of this test,  $\beta(\theta)$ , satisfies  $\lim_{\theta \rightarrow \infty} \beta(\theta) = 0$  if  $\alpha$  is small enough to make  $k' > 0$  (Exercise 8.56). The test does a poor job of detecting values of  $\theta$  that are much larger than  $\theta_0$ . Of course, this family of distributions is not an exponential family. No UMP test of  $H_0: \theta \leq \theta_0$  versus  $H_1: \theta > \theta_0$  exists (Exercise 8.32 is a special case of this). ||

### 8.3.5 Sizes of Union–Intersection and Intersection–Union Tests

Because of the simple way in which they are constructed, the sizes of union-intersection (UIT) and intersection-union (IUT) tests can often be bounded above by the sizes of some other tests. Such bounds are useful if a level  $\alpha$  test is wanted, but the size of the UIT or IUT is too difficult to evaluate. In this section we discuss these bounds and give examples in which the bounds are sharp, that is, the size of the test is equal to the bound.

First consider UITs. Recall that, in this situation, we are testing a null hypothesis of the form  $H_0: \theta \in \Theta_0$  where  $\Theta_0 = \bigcap_{\gamma \in \Gamma} \Theta_\gamma$ . To be specific, let  $\lambda_\gamma(\mathbf{x})$  be the LRT statistic for testing  $H_{0\gamma}: \theta \in \Theta_\gamma$  versus  $H_{1\gamma}: \theta \in \Theta_\gamma^c$  and let  $\lambda(\mathbf{x})$  be the LRT statistic for testing  $H_0: \theta \in \Theta_0$  versus  $H_1: \theta \in \Theta_0^c$ . Then we have the following relationships between the overall LRT and the UIT based on  $\lambda_\gamma(\mathbf{x})$ .

**THEOREM 8.3.4:** Consider testing  $H_0: \theta \in \Theta_0$  versus  $H_1: \theta \in \Theta_0^c$ , where  $\Theta_0 = \bigcap_{\gamma \in \Gamma} \Theta_\gamma$  and  $\lambda_\gamma(\mathbf{x})$  is defined in the previous paragraph. Define  $T(\mathbf{x}) = \inf_{\gamma \in \Gamma} \lambda_\gamma(\mathbf{x})$ , and form the UIT with rejection region

$$\{\mathbf{x}: \lambda_\gamma(\mathbf{x}) < c \text{ for some } \gamma \in \Gamma\} = \{\mathbf{x}: T(\mathbf{x}) < c\}.$$

Also consider the usual LRT with rejection region  $\{\mathbf{x}: \lambda(\mathbf{x}) < c\}$ . Then

- a.  $T(\mathbf{x}) \geq \lambda(\mathbf{x})$  for every  $\mathbf{x}$ ;
- b. If  $\beta_T(\theta)$  and  $\beta_\lambda(\theta)$  are the power functions for the tests based on  $T$  and  $\lambda$ , respectively, then  $\beta_T(\theta) \leq \beta_\lambda(\theta)$  for every  $\theta \in \Theta$ ; and
- c. If the LRT is a level  $\alpha$  test then the UIT is a level  $\alpha$  test.

*Proof:* Since  $\Theta_0 = \bigcap_{\gamma \in \Gamma} \Theta_\gamma \subset \Theta_\gamma$  for any  $\gamma$ , from Definition 8.2.1 we see that for any  $x$ ,

$$\lambda_\gamma(x) \geq \lambda(x) \quad \text{for each } \gamma \in \Gamma,$$

because the region of maximization is bigger for the individual  $\lambda_\gamma$ . Thus  $T(x) = \inf_{\gamma \in \Gamma} \lambda_\gamma(x) \geq \lambda(x)$ , proving (a). By (a),  $\{x : T(x) < c\} \subset \{x : \lambda(x) < c\}$ , so

$$\beta_T(\theta) = P_\theta(T(X) < c) \leq P_\theta(\lambda(X) < c) = \beta_\lambda(\theta),$$

proving (b). Since (b) holds for every  $\theta$ ,  $\sup_{\theta \in \Theta_0} \beta_T(\theta) \leq \sup_{\theta \in \Theta_0} \beta_\lambda(\theta) \leq \alpha$ , proving (c).  $\square$

**Example 8.3.12:** In some situations,  $T(x) = \lambda(x)$  in Theorem 8.3.4. The UIT built up from individual LRTs is the same as the overall LRT. This was the case in Example 8.2.8. There the UIT formed from two one-sided  $t$  tests was equivalent to the two-sided LRT.  $\parallel$

Since the LRT is uniformly more powerful than the UIT in Theorem 8.3.4, we might ask why should we use the UIT. One reason is that the UIT has a smaller Type I Error probability for every  $\theta \in \Theta_0$ . Furthermore, if  $H_0$  is rejected, we may wish to look at the individual tests of  $H_{0\gamma}$  to see why. As yet, we have not discussed inferences for the individual  $H_{0\gamma}$ . The error probabilities for such inferences would have to be examined before such an inference procedure were adopted. But the possibility of gaining additional information by looking at the  $H_{0\gamma}$  individually, rather than looking only at the overall LRT, is evident.

Now we investigate the sizes of IUTs. A simple bound for the size of an IUT is related to the sizes of the individual tests that are used to define the IUT. Recall that in this situation the null hypothesis is expressible as a *union*, that is, we are testing

$$H_0: \theta \in \Theta_0 \quad \text{versus} \quad H_1: \theta \in \Theta_0^c, \quad \text{where } \Theta_0 = \bigcup_{\gamma \in \Gamma} \Theta_\gamma.$$

An IUT has a rejection region of the form  $R = \bigcap_{\gamma \in \Gamma} R_\gamma$  where  $R_\gamma$  is the rejection region for a test of  $H_{0\gamma}: \theta \in \Theta_\gamma$ .

**THEOREM 8.3.5:** Let  $\alpha_\gamma$  be the size of the test of  $H_{0\gamma}$  with rejection region  $R_\gamma$ . Then the IUT with rejection region  $R = \bigcap_{\gamma \in \Gamma} R_\gamma$  is a level  $\alpha = \sup_{\gamma \in \Gamma} \alpha_\gamma$  test.

*Proof:* Let  $\theta \in \Theta_0$ . Then  $\theta \in \Theta_\gamma$  for some  $\gamma$  and

$$P_\theta(X \in R) \leq P_\theta(X \in R_\gamma) \leq \alpha_\gamma \leq \alpha.$$

Since  $\theta \in \Theta_0$  was arbitrary, the IUT is a level  $\alpha$  test.  $\square$

Typically, the individual rejection regions  $R_\gamma$  are chosen so that  $\alpha_\gamma = \alpha$  for all  $\gamma$ . In such a case, Theorem 8.3.5 states that the resulting IUT is a level  $\alpha$  test.

Theorem 8.3.5, which provides an upper bound for the size of an IUT, is somewhat more useful than Theorem 8.3.4, which provides an upper bound for the size of a UIT. Theorem 8.3.4 applies only to UITs constructed from likelihood ratio tests. In contrast, Theorem 8.3.5 applies to any IUT.

The bound in Theorem 8.3.4 is the size of the LRT which, in a complicated problem, may be difficult to compute. In Theorem 8.3.5, however, the LRT need not be used to obtain the upper bound. Any test of  $H_{0\gamma}$  with known size  $\alpha_\gamma$  can be used and then the upper bound on the size of the IUT is given in terms of the known sizes  $\alpha_\gamma, \gamma \in \Gamma$ . If we attempt to bound the size of the UIT in terms of the sizes of the individual tests, the inequality naturally goes in the *opposite* direction and a lower bound is obtained (Exercise 8.57). A lower bound on the size of a test is not very useful.

**Example 8.3.13:** In Example 8.2.9, let  $n = m = 58$ ,  $t = 1.672$ , and  $b = 57$ . Then each of the individual tests has size  $\alpha = .05$  (approximately). Therefore, by Theorem 8.3.5, the IUT is a level  $\alpha = .05$  test, that is, the probability of deciding the product is good, when in fact it is not, is no more than .05. In fact, this test is a size  $\alpha = .05$  test. To see this consider a sequence of parameter points

$$\theta_n = (\theta_{1n}, \theta_2), \quad \text{with } \theta_{1n} \rightarrow \infty \text{ as } n \rightarrow \infty, \quad \text{and } \theta_2 = .95.$$

All such parameter points are in  $\Theta_0$  since  $\theta_2 \leq .95$ . For these parameter points, the power function of the IUT satisfies

$$\begin{aligned} \lim_{n \rightarrow \infty} P_{\theta_n}(R_1 \cap R_2) &= \lim_{n \rightarrow \infty} (1 - P_{\theta_n}(R_1^c \cup R_2^c)) && \left( \begin{array}{l} R_1 \cap R_2 \text{ is the} \\ \text{rejection region} \end{array} \right) \\ &\geq 1 - \lim_{n \rightarrow \infty} (P_{\theta_n}(R_1^c) + P_{\theta_n}(R_2^c)). && \left( \begin{array}{l} \text{Bonferroni} \\ \text{Inequality} \end{array} \right) \end{aligned}$$

But  $P_{\theta_n}(R_1^c) \rightarrow 0$  as  $\theta_{1n} \rightarrow \infty$  while  $P_{\theta_n}(R_2^c) = .95$  for all  $n$  since  $\theta_2 = .95$ . Thus

$$.05 \geq \sup_{\theta \in \Theta_0} P_\theta(R_1 \cap R_2) \geq \lim_{n \rightarrow \infty} P_{\theta_n}(R_1 \cap R_2) \geq .05,$$

the first inequality being the result of Theorem 8.3.5. Thus the IUT is a size  $\alpha$  test. ||

Note that, in Example 8.3.13, only the marginal distributions of the  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_m$  were used to find the size of the test. This point is extremely important, and directly relates to the usefulness of IUTs, because the joint distribution is often difficult to know and, if known, often difficult to work with. For example,  $X_i$  and  $Y_i$  may be related if they are measurements on the same piece of fabric, but this relationship would have to be modeled and used to calculate the exact power of the IUT at any particular parameter value.

## 8.4 Other Considerations

As in Section 7.4, this section describes a few methods for deriving *some* tests in complicated problems. We are thinking of problems in which no optimal test, as defined in earlier sections, exists (for example, no UMP unbiased test exists) or is known. In such situations, the derivation of any reasonable test might be of use. In two subsections, we will discuss large-sample properties of likelihood ratio tests and other approximate large-sample tests.

### 8.4.1 Asymptotic Distribution of LRTs

One of the most useful methods for complicated models is the likelihood ratio method of test construction because it gives an explicit definition of the test statistic,

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})},$$

and an explicit form for the rejection region,  $\{\mathbf{x}: \lambda(\mathbf{x}) \leq c\}$ . After the data  $\mathbf{X} = \mathbf{x}$  are observed the likelihood function,  $L(\theta|\mathbf{x})$ , is a completely defined function of the variable  $\theta$ . Even if the two suprema of  $L(\theta|\mathbf{x})$ , over the sets  $\Theta_0$  and  $\Theta$ , cannot be analytically obtained, they can usually be computed numerically. Thus, the test statistic  $\lambda(\mathbf{x})$  can be obtained for the observed data point even if no convenient formula defining  $\lambda(\mathbf{x})$  is available.

To define a level  $\alpha$  test, the constant  $c$  must be chosen so that

$$(8.4.1) \quad \sup_{\theta \in \Theta_0} P_{\theta} (\lambda(\mathbf{X}) \leq c) \leq \alpha.$$

If we cannot derive a simple formula for  $\lambda(\mathbf{x})$ , it might seem that it is hopeless to derive the sampling distribution of  $\lambda(\mathbf{X})$  and thus know how to pick  $c$  to ensure (8.4.1). The following general theorem allows us to ensure (8.4.1) is true, at least for large samples. A complete discussion of this topic may be found in Kendall and Stuart (1979).

**THEOREM 8.4.1:** Let  $X_1, \dots, X_n$  be a random sample from a pdf or pmf  $f(x|\theta)$ . Under some regularity conditions on the model  $f(x|\theta)$ , if  $\theta \in \Theta_0$  then the distribution of the statistic  $-2 \log \lambda(\mathbf{X})$  converges to a chi squared distribution as the sample size  $n \rightarrow \infty$ . The degrees of freedom of the limiting distribution is the difference between the number of free parameters specified by  $\theta \in \Theta_0$  and the number of free parameters specified by  $\theta \in \Theta$ .  $\square$

The “regularity conditions” needed for the model, the same as those mentioned in Chapter 7, are general conditions that are satisfied for many reasonable distributions (but not all). These conditions are mainly concerned with the existence and behavior of the derivatives (with respect to the parameter) of the likelihood function, and the support of the distribution (it cannot depend on the parameter).

Rejection of  $H_0: \theta \in \Theta_0$  for small values of  $\lambda(\mathbf{X})$  is equivalent to rejection for large values of  $-2 \log \lambda(\mathbf{X})$ . Thus,

$$H_0 \text{ is rejected if and only if } -2 \log \lambda(\mathbf{X}) \geq \chi^2_{\nu, \alpha},$$

where  $\nu$  is the degrees of freedom specified in Theorem 8.4.1. The Type I Error probability will be approximately  $\alpha$  if  $\theta \in \Theta_0$  and the sample size is large. In this way, (8.4.1) will be approximately satisfied for large sample sizes and an *asymptotic size  $\alpha$  test* has been defined. Note that the theorem will actually imply only that

$$\lim_{n \rightarrow \infty} P_\theta(\text{reject } H_0) = \alpha \quad \text{for each } \theta \in \Theta_0,$$

not that the  $\sup_{\theta \in \Theta_0} P_\theta(\text{reject } H_0)$  converges to  $\alpha$ . This is usually the case for asymptotic size  $\alpha$  tests.

The computation of the degrees of freedom for the test statistic is usually straightforward. Most often,  $\Theta$  can be represented as a subset of  $q$ -dimensional Euclidian space that contains an open subset in  $\mathbb{R}^q$  and  $\Theta_0$  can be represented as a subset of  $p$ -dimensional Euclidian space that contains an open subset in  $\mathbb{R}^p$ , where  $p < q$ . Then  $q - p = \nu$  is the degrees of freedom for the test statistic.

**Example 8.4.1:** Let  $\theta = (p_1, p_2, p_3, p_4, p_5)$  where  $\sum_{j=1}^5 p_j = 1$  and  $p_j \geq 0, j = 1, \dots, 5$ . Suppose  $X_1, \dots, X_n$  are iid discrete random variables and  $P_\theta(X_i = j) = p_j, j = 1, \dots, 5$ . Thus the pmf of  $X_i$  is  $f(j|\theta) = p_j$  and the likelihood function is

$$L(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta) = p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4} p_5^{y_5},$$

where  $y_j = \text{number of } x_1, \dots, x_n \text{ equal to } j$ . Consider testing

$$H_0: p_1 = p_2 = p_3 \text{ and } p_4 = p_5 \quad \text{versus} \quad H_1: H_0 \text{ is not true.}$$

The full parameter space,  $\Theta$ , is really a four-dimensional set. Since  $p_5 = 1 - p_1 - p_2 - p_3 - p_4$ , there are only four free parameters. The parameter set is defined by

$$\sum_{j=1}^4 p_j \leq 1 \quad \text{and} \quad p_j \geq 0, \quad j = 1, \dots, 4,$$

a subset of  $\mathbb{R}^4$  containing an open subset of  $\mathbb{R}^4$ . Thus  $q = 4$ . There is only one free parameter in the set specified by  $H_0$  because, once  $p_1, 0 \leq p_1 \leq \frac{1}{3}$ , is fixed, then  $p_2 = p_3$  must equal  $p_1$  and  $p_4 = p_5$  must equal  $\frac{1-3p_1}{2}$ . Thus  $p = 1$ , and the degrees of freedom is  $\nu = 4 - 1 = 3$ .

To calculate  $\lambda(\mathbf{x})$ , the MLE of  $\theta$  under both  $\Theta_0$  and  $\Theta$  must be determined. By setting

$$\frac{\partial}{\partial p_j} \log L(\theta|x) = 0 \quad \text{for each of } j = 1, \dots, 4,$$

and using the facts that  $p_5 = 1 - p_1 - p_2 - p_3 - p_4$  and  $y_5 = n - y_1 - y_2 - y_3 - y_4$ , it can be verified that the MLE of  $p_j$  under  $\Theta$  is  $\hat{p}_j = y_j/n$ . Under  $H_0$ , the likelihood function reduces to

$$L(\theta|x) = p_1^{y_1+y_2+y_3} \left( \frac{1-3p_1}{2} \right)^{y_4+y_5}.$$

Again, the usual method of setting the derivative equal to zero shows that the MLE of  $p_1$  under  $H_0$  is  $\hat{p}_{10} = (y_1 + y_2 + y_3)/(3n)$ . Then  $\hat{p}_{10} = \hat{p}_{20} = \hat{p}_{30}$  and  $\hat{p}_{40} = \hat{p}_{50} = (1-3\hat{p}_{10})/2$ . Substituting these values and the  $\hat{p}_j$  values into  $L(\theta|x)$  and combining terms with the same exponent yields

$$\lambda(x) = \left( \frac{y_1 + y_2 + y_3}{3y_1} \right)^{y_1} \left( \frac{y_1 + y_2 + y_3}{3y_2} \right)^{y_2} \left( \frac{y_1 + y_2 + y_3}{3y_3} \right)^{y_3} \left( \frac{y_4 + y_5}{2y_4} \right)^{y_4} \left( \frac{y_4 + y_5}{2y_5} \right)^{y_5}.$$

Thus the test statistic is

$$(8.4.2) \quad -2 \log \lambda(x) = 2 \sum_{i=1}^5 y_i \log \left( \frac{y_i}{m_i} \right)$$

where  $m_1 = m_2 = m_3 = (y_1 + y_2 + y_3)/3$  and  $m_4 = m_5 = (y_4 + y_5)/2$ . The asymptotic size  $\alpha$  test rejects  $H_0$  if  $-2 \log \lambda(x) \geq \chi^2_{3,\alpha}$ . This example is one of a large class of testing problems for which the asymptotic theory of the likelihood ratio test is extensively used. This class of problems is called *log linear models*. The expression (8.4.2) is a typical expression for the LRT statistic in these problems. A thorough discussion of log linear models may be found in Bishop, Fienberg, and Holland (1975). ||

## 8.4.2 Other Large-Sample Tests

Another common method of constructing a large-sample test statistic is based on an estimator that has an asymptotic normal distribution. Suppose we wish to test a hypothesis about a real-valued parameter  $\theta$ , and  $W_n = W(X_1, \dots, X_n)$  is a point estimator of  $\theta$ , based on a sample of size  $n$ , that has been derived by some method. For example,  $W_n$  might be the MLE of  $\theta$ . An approximate test, based on a normal approximation, can be justified in the following way. If  $\sigma_n^2$  denotes the variance of  $W_n$  and if we can use some form of the Central Limit Theorem to show that, as  $n \rightarrow \infty$ ,  $(W_n - \theta)/\sigma_n$  converges in distribution to a standard normal random variable, then  $(W_n - \theta)/\sigma_n$  can be compared to a  $n(0, 1)$  distribution. We therefore have the basis for an approximate test.

There are, of course, many details to be verified in the argument of the previous paragraph, but this idea does have application in many situations. For example, if  $W_n$

is an MLE the above arguments are usually valid. Note that the distribution of  $W_n$  and, perhaps, the value of  $\sigma_n$  depend on the value of  $\theta$ . The convergence, therefore, more formally says that for each fixed value of  $\theta \in \Theta$ , if we use the corresponding distribution for  $W_n$  and the corresponding value for  $\sigma_n$ ,  $(W_n - \theta)/\sigma_n$  converges to a standard normal. If for each  $n$ ,  $\sigma_n$  is a calculable constant (which may depend on  $\theta$  but not any other unknown parameters), then a test based on  $(W_n - \theta)/\sigma_n$  might be derived.

In some instances,  $\sigma_n$  also depends on unknown parameters. In such a case, we look for an estimate  $S_n$  of  $\sigma_n$  with the property that  $\sigma_n/S_n$  converges in probability to one. Then, using Slutsky's Theorem (as in Example 5.3.5) we can deduce that  $(W_n - \theta)/S_n$  also converges in distribution to a standard normal distribution. A large-sample test may be based on this fact.

Suppose we wish to test the two-sided hypothesis  $H_0: \theta = \theta_0$  versus  $H_1: \theta \neq \theta_0$ . An approximate test can be based on the statistic  $Z_n = (W_n - \theta_0)/S_n$ , and would reject  $H_0$  if and only if  $Z_n < -z_{\alpha/2}$  or  $Z_n > z_{\alpha/2}$ . If  $H_0$  is true, then  $\theta = \theta_0$  and  $Z_n$  converges in distribution to  $Z \sim N(0, 1)$ . Thus, the Type I Error probability,

$$P_{\theta_0}(Z_n < -z_{\alpha/2} \text{ or } Z_n > z_{\alpha/2}) \rightarrow P(Z < -z_{\alpha/2} \text{ or } Z > z_{\alpha/2}) = \alpha$$

and this is an asymptotically size  $\alpha$  test.

Now consider an alternative parameter value  $\theta \neq \theta_0$ . We can write

$$(8.4.3) \quad Z_n = \frac{W_n - \theta_0}{S_n} = \frac{W_n - \theta}{S_n} + \frac{\theta - \theta_0}{S_n}.$$

No matter what the value of  $\theta$ , the term  $(W_n - \theta)/S_n \rightarrow N(0, 1)$ . Typically, it is also the case that  $\sigma_n \rightarrow 0$  as  $n \rightarrow \infty$ . (Recall,  $\sigma_n = \sqrt{\text{Var } W_n}$ , and estimators typically become more precise as  $n \rightarrow \infty$ .) Thus,  $S_n$  will converge in probability to zero and the term  $(\theta - \theta_0)/S_n$  will converge to  $+\infty$  or  $-\infty$  in probability, depending on whether  $(\theta - \theta_0)$  is positive or negative. Thus,  $Z_n$  will converge to  $+\infty$  or  $-\infty$  in probability and

$$\begin{aligned} P_\theta(\text{reject } H_0) &= P_\theta(Z_n < -z_{\alpha/2} \text{ or } Z_n > z_{\alpha/2}) \\ &\rightarrow 1, \quad \text{as } n \rightarrow \infty. \end{aligned}$$

In this way, a test with asymptotic size  $\alpha$  and asymptotic power one can be constructed.

If we wish to test the one-sided hypothesis,  $H_0: \theta \leq \theta_0$  versus  $H_1: \theta > \theta_0$ , a similar test might be constructed. Again, the test statistic  $Z_n = (W_n - \theta_0)/S_n$  would be used and the test would reject  $H_0$  if and only if  $Z_n > z_\alpha$ . Using reasoning similar to the above, we could conclude that the power function of this test converges to zero,  $\alpha$ , or one according as  $\theta < \theta_0$ ,  $\theta = \theta_0$ , or  $\theta > \theta_0$ . Thus this test too has reasonable asymptotic power properties.

**Example 8.4.2:** Let  $X_1, \dots, X_n$  be a random sample from a Bernoulli( $p$ ) population. Consider testing  $H_0: p \leq p_0$  versus  $H_1: p > p_0$  where  $0 < p_0 < 1$  is a specified value. The MLE of  $p$ , based on a sample of size  $n$ , is  $\hat{p}_n = \sum_{i=1}^n X_i/n$ . Since

$\hat{p}_n$  is just a sample mean, the Central Limit Theorem applies and states that for any  $p$ ,  $0 < p < 1$ ,  $(\hat{p}_n - p)/\sigma_n$  converges to a standard normal random variable. Here  $\sigma_n = \sqrt{p(1-p)/n}$ , a value that depends on the unknown parameter  $p$ . A reasonable estimate of  $\sigma_n$  is  $S_n = \sqrt{\hat{p}_n(1-\hat{p}_n)/n}$  and it can be shown (Exercise 5.15) that  $\sigma_n/S_n$  converges in probability to one. Thus, for any  $p$ ,  $0 < p < 1$ ,

$$\frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}} \rightarrow n(0, 1).$$

The test statistic  $Z_n$  is defined by replacing  $p$  by  $p_0$  and the large-sample test rejects  $H_0$  if  $Z_n > z_\alpha$ .

If there was interest in testing the two-sided hypothesis  $H_0 : p = p_0$  versus  $H_1 : p \neq p_0$  where  $0 < p_0 < 1$  is a specified value, the above strategy is again applicable. However, in this case, there is an alternative approximate test. By the Central Limit Theorem, for any  $p$ ,  $0 < p < 1$ ,

$$\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \rightarrow n(0, 1).$$

Therefore, it follows that, if the null hypothesis is true, the statistic

$$Z'_n = \frac{\hat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}} \sim n(0, 1). \quad (\text{approximately})$$

The approximate level  $\alpha$  test rejects  $H_0$  if  $|Z'_n| > z_{\alpha/2}$ .

In cases where both tests are applicable, for example when testing  $H_0 : p = p_0$ , it is not clear which test is to be preferred. The power functions (actual, not approximate) cross one another, so each test is more powerful in a certain portion of the parameter space. (Ghosh (1979)) gives some insights into this problem. A related binomial controversy, that of the two-sample problem, is discussed by Robbins (1977) and Eberhardt and Fligner (1977). Two different test statistics for this problem are given in Exercise 8.59.)

Of course, any comparison of power functions is confounded by the fact that these are *approximate* tests, and do not necessarily maintain level  $\alpha$ . The use of a continuity correction (Example 3.2.2) can help in this problem. In many cases, approximate procedures that use the continuity correction turn out to be *conservative*, that is, they maintain their nominal  $\alpha$  level. This will be further discussed in Chapter 9. ||

## EXERCISES

---

- 8.1 In 1,000 tosses of a coin, 560 heads and 440 tails appear. Is it reasonable to assume that the coin is fair? Justify your answer.

- 8.2 In a given city it is assumed that the number of automobile accidents in a given year follows a Poisson distribution. In past years the average number of accidents per year was 15, and this year it was 10. Is it justified to claim that the accident rate has dropped?
- 8.3 Here, the LRT alluded to in Example 8.2.9 will be derived. Suppose that we observe  $m$  iid  $\text{Bernoulli}(\theta)$  random variables, denoted by  $Y_1, \dots, Y_m$ . Show that the LRT of  $H_0: \theta \leq \theta_0$  versus  $H_1: \theta > \theta_0$  will reject  $H_0$  if  $\sum_{i=1}^m Y_i > b$ .
- 8.4 Prove the assertion made in the text after Definition 8.2.1. If  $f(x|\theta)$  is the pmf of a discrete random variable then the numerator of  $\lambda(\mathbf{x})$ , the LRT statistic, is the maximum probability of the observed sample when the maximum is computed over parameters in the null hypothesis. Furthermore, the denominator of  $\lambda(\mathbf{x})$  is the maximum probability of the observed sample over all possible parameters.
- 8.5 A random sample,  $X_1, \dots, X_n$ , is drawn from a Pareto population with pdf

$$f(x|\theta, \nu) = \frac{\theta\nu^\theta}{x^{\theta+1}} I_{[\nu, \infty)}(x), \quad \theta > 0, \quad \nu > 0.$$

- a. Find the MLEs of  $\theta$  and  $\nu$ .  
 b. Show that the LRT of

$$H_0: \theta = 1, \nu \text{ unknown} \quad \text{versus} \quad H_1: \theta \neq 1, \nu \text{ unknown},$$

has critical region of the form  $\{\mathbf{x}: T(\mathbf{x}) \leq c_1 \text{ or } T(\mathbf{x}) \geq c_2\}$ , where  $0 < c_1 < c_2$  and

$$T = \log \left[ \frac{\prod_{i=1}^n X_i}{(\min_i X_i)^n} \right].$$

- c. Show that, under  $H_0$ ,  $2T$  has a chi squared distribution, and find the number of degrees of freedom. (*Hint:* Obtain the joint distribution of the  $n - 1$  nontrivial terms  $X_i/(\min_i X_i)$  conditional on  $\min_i X_i$ . Put these  $n - 1$  terms together, and notice that the distribution of  $T$  given  $\min_i X_i$  does not depend on  $\min_i X_i$ , so it is the unconditional distribution of  $T$ .)
- 8.6 Suppose that we have two independent random samples:  $X_1, \dots, X_n$  are exponential( $\theta$ ), and  $Y_1, \dots, Y_m$  are exponential( $\mu$ ).  
 a. Find the LRT of  $H_0: \theta = \mu$  versus  $H_1: \theta \neq \mu$ .  
 b. Show that the test in part (a) can be based on the statistic

$$T = \frac{\sum X_i}{\sum X_i + \sum Y_i}.$$

- c. Find the distribution of  $T$  when  $H_0$  is true.  
 8.7 We have already seen the usefulness of the LRT in dealing with problems with nuisance parameters. We now look at some other nuisance parameter problems.  
 a. Find the LRT of

$$H_0: \theta \leq 0 \quad \text{versus} \quad H_1: \theta > 0,$$

based on a sample  $X_1, \dots, X_n$  from a population with probability density function  $f(x|\theta, \lambda) = \frac{1}{\lambda} e^{-(x-\theta)/\lambda} I_{(\theta, \infty)}(x)$ , where both  $\theta$  and  $\lambda$  are unknown.

b. We have previously seen that the exponential pdf is a special case of a gamma pdf. Generalizing in another way, the exponential pdf can be considered as a special case of the Weibull( $\gamma, \beta$ ). The Weibull pdf, which reduces to the exponential if  $\gamma = 1$ , is very important in modeling reliability of systems. Suppose that  $X_1, \dots, X_n$  is a random sample from a Weibull population with both  $\gamma$  and  $\beta$  unknown. Find the LRT of

$$H_0 : \gamma = 1 \quad \text{versus} \quad H_1 : \gamma \neq 1.$$

- 8.8** A special case of a normal family is one in which the mean and the variance are related, the  $n(\theta, a\theta)$  family. If we are interested in testing this relationship, regardless of the value of  $\theta$ , we are again faced with a nuisance parameter problem.
- a. Find the LRT of  $H_0: a = 1$  versus  $H_1: a \neq 1$  based on a sample  $X_1, \dots, X_n$  from a  $n(\theta, a\theta)$  family, where  $\theta$  is unknown.
- b. A similar question can be asked about a related family, the  $n(\theta, a\theta^2)$  family. Thus, if  $X_1, \dots, X_n$  are iid  $n(\theta, a\theta^2)$ , where  $\theta$  is unknown, find the LRT of  $H_0: a = 1$  versus  $H_1: a \neq 1$ .
- 8.9** Let  $X_1, \dots, X_n$  be iid Poisson( $\lambda$ ), and let  $\lambda$  have a gamma( $\alpha, \beta$ ) distribution, the conjugate family for the Poisson. In Exercise 7.25 the posterior distribution of  $\lambda$  was found, including the posterior mean and variance. Now consider a Bayesian test of  $H_0: \lambda \leq \lambda_0$  versus  $H_1: \lambda > \lambda_0$ .
- a. Calculate expressions for the posterior probabilities of  $H_0$  and  $H_1$ .
- b. If  $\alpha = \frac{5}{2}$  and  $\beta = 2$ , the prior distribution is a chi squared distribution with 5 degrees of freedom. Explain how a chi squared table could be used to perform a Bayesian test.
- 8.10** In Exercise 7.24 the posterior distribution of  $\sigma^2$ , the variance of a normal population, given  $S^2$ , the sample variance based on a sample of size  $n$ , was found using a conjugate prior for  $\sigma^2$  (the inverted gamma pdf with parameters  $\alpha$  and  $\beta$ ). Based on observing  $S^2$ , a decision about the hypotheses  $H_0: \sigma \leq 1$  versus  $H_1: \sigma > 1$  is to be made.
- a. Find the region of the sample space for which  $P(\sigma \leq 1|S^2) > P(\sigma > 1|S^2)$ , the region for which a Bayes test will decide that  $\sigma \leq 1$ .
- b. Compare the region in part (a) with the acceptance region of an LRT. Is there any choice of prior parameters for which the regions agree?
- 8.11** For samples of size  $n = 1, 4, 16, 64, 100$ , from a normal population with mean  $\mu$  and known variance  $\sigma^2$ , plot the power function of the following tests. Take  $\alpha = .05$ .
- a.  $H_0: \mu \leq 0$  versus  $H_1: \mu > 0$
- b.  $H_0: \mu = 0$  versus  $H_1: \mu \neq 0$
- 8.12** Let  $X_1, X_2$  be iid uniform( $\theta, \theta + 1$ ). For testing  $H_0: \theta = 0$  versus  $H_1: \theta > 0$ , we have two competing tests:

$$\phi_1(X_1) : \text{Reject } H_0 \text{ if } X_1 > .95,$$

$$\phi_2(X_1, X_2) : \text{Reject } H_0 \text{ if } X_1 + X_2 > C.$$

- a. Find the value of  $C$  so that  $\phi_2$  has the same size as  $\phi_1$ .
- b. Calculate the power function of each test. Draw a well-labeled graph of each power function.
- c. Prove or disprove:  $\phi_2$  is a more powerful test than  $\phi_1$ .
- d. Show how to get a test that has the same size, but is more powerful than  $\phi_2$ .
- 8.13** For a random sample  $X_1, \dots, X_n$  of Bernoulli( $p$ ) variables, it is desired to test

$$H_0 : p = .49 \quad \text{versus} \quad H_1 : p = .51.$$

Use the Central Limit Theorem to determine, approximately, the sample size needed so that the two probabilities of error are both about .01. Use a test function that rejects  $H_0$  if  $\sum_{i=1}^n X_i$  is large.

- 8.14** Show that for a random sample  $X_1, \dots, X_n$  from a  $n(0, \sigma^2)$  population, the most powerful test of  $H_0: \sigma = \sigma_0$  versus  $H_1: \sigma = \sigma_1$ , where  $\sigma_0 < \sigma_1$ , is given by

$$\phi(\Sigma X_i^2) = \begin{cases} 1 & \text{if } \Sigma X_i^2 > c \\ 0 & \text{if } \Sigma X_i^2 \leq c \end{cases}.$$

For a given value of  $\alpha$ , the size of the Type I Error, show how the value of  $c$  is explicitly determined.

- 8.15** One very striking abuse of  $\alpha$  levels is to choose them *after* seeing the data, and to choose them in such a way as to force rejection (or acceptance) of a null hypothesis. To see what the *true* Type I and Type II Error probabilities of such a procedure are, calculate size and power of the following two trivial tests:

- Always reject  $H_0$ , no matter what data are obtained (equivalent to the practice of choosing the  $\alpha$  level to force rejection of  $H_0$ ).
- Always accept  $H_0$ , no matter what data are obtained (equivalent to the practice of choosing the  $\alpha$  level to force acceptance of  $H_0$ ).

- 8.16** Suppose that  $X_1, \dots, X_n$  are iid with a  $\text{beta}(\mu, 1)$  pdf and  $Y_1, \dots, Y_m$  are iid with a  $\text{beta}(\theta, 1)$  pdf. Also assume that the  $X$ s are independent of the  $Y$ s.

- Find an LRT of  $H_0: \theta = \mu$  versus  $H_1: \theta \neq \mu$ .
- Show that the test in part (a) can be based on the statistic

$$T = \frac{\sum \log X_i}{\sum \log X_i + \sum \log Y_i}.$$

- Find the distribution of  $T$  when  $H_0$  is true, and then show how to get a test of size  $\alpha = .10$ .

- 8.17** Let  $X_1, \dots, X_n$  be iid  $n(\theta, \sigma^2)$ ,  $\sigma^2$  known, and let  $\theta$  have a double exponential distribution, that is,  $\pi(\theta) = e^{-|\theta|/a}/(2a)$ ,  $a$  known. A Bayesian test of the hypotheses  $H_0: \theta \leq 0$  versus  $H_1: \theta > 0$  will decide in favor of  $H_1$  if its posterior probability is large.

- For a given constant  $K$ , calculate the posterior probability that  $\theta > K$ , that is,  $P(\theta > K | x_1, \dots, x_n, a)$ .
- Find an expression for  $\lim_{a \rightarrow \infty} P(\theta > K | x_1, \dots, x_n, a)$ .
- Compare your answer in part (b) to the p-value associated with the classical hypothesis test.

- 8.18** In each of the following situations, calculate the p-value of the observed data.

- For testing  $H_0: \theta \leq \frac{1}{2}$  versus  $H_1: \theta > \frac{1}{2}$ , 7 successes are observed out of 10 Bernoulli trials.
- For testing  $H_0: \lambda \leq 1$  versus  $H_1: \lambda > 1$ ,  $X = 3$  is observed, where  $X \sim \text{Poisson}(\lambda)$ .
- For testing  $H_0: \lambda \leq 1$  versus  $H_1: \lambda > 1$ ,  $X_1 = 3$ ,  $X_2 = 5$ , and  $X_3 = 1$  is observed, where  $X_i \sim \text{Poisson}(\lambda)$ , independent.

- 8.19** In Example 8.2.7 we saw an example of a one-sided Bayesian hypothesis test. Now we will consider a similar situation, but with a two-sided test. We want to test

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \neq 0,$$

and we observe  $X_1, \dots, X_n$ , a random sample from a  $n(\theta, \sigma^2)$  population,  $\sigma^2$  known. A type of prior distribution that is often used in this situation is a mixture of a point mass on  $\theta = 0$  and a pdf spread out over  $H_1$ . A typical choice is to take  $P(\theta = 0) = \frac{1}{2}$ , and if  $\theta \neq 0$ , take the prior distribution to be  $\frac{1}{2}n(0, \tau^2)$ , where  $\tau^2$  is known.

- a. Show that the prior defined above is proper, that is,  $P(-\infty < \theta < \infty) = 1$ .
- b. Calculate the posterior probability that  $H_0$  is true,  $P(\theta = 0|x_1, \dots, x_n)$ .
- c. Find an expression for the p-value corresponding to a value of  $\bar{x}$ .
- d. For the special case  $\sigma^2 = \tau^2 = 1$ , compare  $P(\theta = 0|x_1, \dots, x_n)$  and the p-value for a range of values of  $\bar{x}$ . Show that the Bayes probability is bigger than the p-value for moderately large values of  $\bar{x}$ .

Note that small values of  $P(\theta = 0|x_1, \dots, x_n)$  are evidence *against*  $H_0$ , and thus this quantity is similar in spirit to a p-value. The fact that these two quantities can have very different values was noted by Lindley (1957), and is also examined by Berger and Sellke (1987). (See the *Miscellanea* section.)

- 8.20** The discrepancies between p-values and Bayes posterior probabilities are not as dramatic in the one-sided problem, as is discussed by Casella and Berger (1987), also mentioned in the *Miscellanea* section. Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population, and suppose that the hypotheses to be tested are

$$H_0 : \theta \leq 0 \quad \text{versus} \quad H_1 : \theta > 0.$$

The prior distribution on  $\theta$  is  $n(0, \tau^2)$ ,  $\tau^2$  known, which is symmetric about the hypotheses in the sense that  $P(\theta \leq 0) = P(\theta > 0) = \frac{1}{2}$ .

- a. Calculate the posterior probability that  $H_0$  is true,  $P(\theta \leq 0|x_1, \dots, x_n)$ .
- b. Find an expression for the p-value corresponding to a value of  $\bar{x}$ , using tests that reject for large values of  $\bar{X}$ .
- c. For the special case  $\sigma^2 = \tau^2 = 1$ , compare  $P(\theta \leq 0|x_1, \dots, x_n)$  and the p-value for values of  $\bar{x} > 0$ . Show that the Bayes probability is always bigger than the p-value.
- d. Using the expression derived in parts (a) and (b), show that

$$\lim_{\tau^2 \rightarrow \infty} P(\theta \leq 0|x_1, \dots, x_n) = \text{p-value},$$

an equality that does not occur in the two-sided problem.

- 8.21** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population,  $\sigma^2$  known. An LRT of  $H_0: \theta = \theta_0$  versus  $H_1: \theta \neq \theta_0$  is a test that rejects  $H_0$  if  $|\bar{X} - \theta_0|/(\sigma/\sqrt{n}) > c$ .
- a. Find an expression, in terms of standard normal probabilities, for the power function of this test.
  - b. The experimenter desires a Type I Error probability of .05, and a maximum Type II Error probability of .25 at  $\theta = \theta_0 + \sigma$ . Find values of  $n$  and  $c$  that will achieve this.
- 8.22** The random variable  $X$  has pdf  $f(x) = e^{-x}$ ,  $x > 0$ . One observation is obtained on the random variable  $Y = X^\theta$ , and a test of  $H_0: \theta = 1$  versus  $H_1: \theta = 2$  needs to be constructed. Find the UMP level  $\alpha = .10$  test and compute the Type II Error probability.
- 8.23** Let  $X$  be a random variable whose pmf under  $H_0$  and  $H_1$  is given by

$x$	1	2	3	4	5	6	7
$f(x H_0)$	.01	.01	.01	.01	.01	.94	
$f(x H_1)$	.06	.05	.04	.03	.02	.01	.79

Use the Neyman–Pearson Lemma to find the most powerful test for  $H_0$  versus  $H_1$  with size  $\alpha = .04$ . Compute the probability of Type II Error for this test.

- 8.24** In the proof of Theorem 8.3.1 (Neyman–Pearson Lemma), it was stated that the proof, which was given for continuous random variables, can easily be adapted to cover discrete random variables.

a. Provide the details, that is, prove the Neyman–Pearson Lemma for discrete random variables. Assume that the  $\alpha$  level is such that it is attainable *without* randomization. (See the *Miscellanea* section.)

b. Comment on what is to be done if the  $\alpha$  level *cannot* be attained without randomization.

- 8.25** Let  $X_1, \dots, X_{10}$  be iid Bernoulli( $p$ ).

a. Find the most powerful test of size  $\alpha = .0547$  of the hypotheses  $H_0: p = \frac{1}{2}$  versus  $H_1: p = \frac{1}{4}$ . Find the power of this test.

b. For testing  $H_0: p \leq \frac{1}{2}$  versus  $H_1: p > \frac{1}{2}$ , find the size and sketch the power function of the test that rejects  $H_0$  if  $\sum_{i=1}^{10} X_i \geq 6$ .

c. For what  $\alpha$  levels does there exist a (nonrandomized) UMP test of the hypotheses in part (a)?

- 8.26** Suppose  $X$  is one observation from a population with beta( $\theta, 1$ ) pdf.

a. For testing  $H_0: \theta \leq 1$  versus  $H_1: \theta > 1$ , find the size and sketch the power function of the test that rejects  $H_0$  if  $X > \frac{1}{2}$ .

b. Find the most powerful level  $\alpha$  test of  $H_0: \theta = 1$  versus  $H_1: \theta = 2$ .

c. Is there a UMP test of  $H_0: \theta \leq 1$  versus  $H_1: \theta > 1$ ? If so, find it. If not, prove so.

- 8.27** Find the LRT of a *simple*  $H_0$  versus a *simple*  $H_1$ . Is this test equivalent to the one obtained from the Neyman–Pearson Lemma? (This relationship is treated in some detail by Solomon (1975).)

- 8.28** Show that each of the following families has an MLR.

a.  $n(\theta, \sigma^2)$  family with  $\sigma^2$  known

b. Poisson( $\theta$ ) family

c. binomial( $n, \theta$ ) family with  $n$  known

- 8.29** a. Show that if a family of pdfs  $\{f(x|\theta): \theta \in \Theta\}$  has an MLR, then the corresponding family of cdfs is *stochastically increasing* in  $\theta$ . (See the *Miscellanea* section.)  
 b. Show that the converse of part (a) is false, that is, give an example of a family of cdfs that is stochastically increasing in  $\theta$  for which the corresponding family of pdfs does not have an MLR.

- 8.30** Suppose  $g(t|\theta) = h(t)c(\theta)e^{w(\theta)t}$  is a one-parameter exponential family for the random variable  $T$ . Show that this family has an MLR if  $w(\theta)$  is an increasing function of  $\theta$ . Give three examples of such a family.

- 8.31** Let  $f(x|\theta)$  be the logistic location pdf

$$f(x|\theta) = \frac{e^{(x-\theta)}}{(1 + e^{(x-\theta)})^2}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

- a. Show that this family has MLR.

- b. Based on one observation,  $X$ , find the most powerful size  $\alpha$  test of  $H_0: \theta = 0$  versus  $H_1: \theta = 1$ . For  $\alpha = .2$ , find the size of the Type II Error.

- c. Show that the test in part (b) is UMP size  $\alpha$  for testing  $H_0: \theta \leq 0$  versus  $H_1: \theta > 0$ . What can be said about UMP tests in general for the logistic location family?

- 8.32** Let  $X$  be one observation from a Cauchy( $\theta$ ) distribution.

- a. Show that this family does not have an MLR.

b. Show that the test

$$\phi(x) = \begin{cases} 1 & \text{if } 1 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

is most powerful of its size for testing  $H_0: \theta = 0$  versus  $H_1: \theta = 1$ . Calculate the Type I and Type II Error probabilities.

c. Prove or disprove: The test in part (b) is UMP for testing  $H_0: \theta \leq 0$  versus  $H_1: \theta > 0$ . What can be said about UMP tests in general for the Cauchy location family?

8.33 Let  $f(x|\theta)$  be the Cauchy scale pdf

$$f(x|\theta) = \frac{\theta}{\pi(\theta^2 + x^2)}, \quad -\infty < x < \infty, \quad \theta > 0.$$

a. Show that this family does not have an MLR.

b. If  $X$  is one observation from  $f(x|\theta)$ , show that  $|X|$  is sufficient for  $\theta$ , and that the distribution of  $|X|$  does have an MLR.

8.34 Let  $X_1, \dots, X_n$  be iid Poisson( $\lambda$ ).

a. Find a UMP test of  $H_0: \lambda \leq \lambda_0$  versus  $H_1: \lambda > \lambda_0$ .

b. Consider the specific case  $H_0: \lambda \leq 1$  versus  $H_1: \lambda > 1$ . Use the Central Limit Theorem to determine the sample size  $n$  so a UMP test satisfies  $P(\text{reject } H_0|\lambda = 1) = .05$  and  $P(\text{reject } H_0|\lambda = 2) = .9$ .

8.35 Let  $X_1, \dots, X_n$  be iid  $n(\theta, 1)$ , and let  $\theta_0$  be a specified value of  $\theta$ .

a. Find the UMP, size  $\alpha$ , test of  $H_0: \theta \geq \theta_0$  versus  $H_1: \theta < \theta_0$ .

b. Show that there does not exist a UMP, size  $\alpha$ , test of  $H_0: \theta = \theta_0$  versus  $H_1: \theta \neq \theta_0$ .

8.36 Let  $X_1, \dots, X_n$  be a random sample from the uniform( $\theta, \theta + 1$ ) distribution. To test  $H_0: \theta = 0$  versus  $H_1: \theta > 0$ , use the test

$$\text{reject } H_0 \text{ if } Y_n \geq 1 \text{ or } Y_1 \geq k,$$

where  $k$  is a constant,  $Y_1 = \min\{X_1, \dots, X_n\}$ ,  $Y_n = \max\{X_1, \dots, X_n\}$ .

a. Determine  $k$  so that the test will have size  $\alpha$ .

b. Find an expression for the power function of the test in part (a).

c. Prove that the test is UMP size  $\alpha$ .

d. Find values of  $n$  and  $k$  so that the UMP .10 level test will have power at least .8 if  $\theta > 1$ .

8.37 In each of the following two situations, show that for any number  $c$ , if  $\theta_1 \leq \theta_2$  then

$$P_{\theta_1}(T > c) \leq P_{\theta_2}(T > c).$$

a.  $\theta$  is a location parameter in the distribution of the random variable  $T$ .

b. The family of pdfs of  $T$ ,  $\{g(t|\theta): \theta \in \Theta\}$ , has an MLR.

8.38 The usual  $t$  distribution, as derived in Section 5.4.2, is also known as a *central t distribution*. It can be thought of as the pdf of a random variable of the form  $T = n(0, 1)/\sqrt{\chi_\nu^2/\nu}$ , where the normal and the chi squared random variables are independent. A generalization of the  $t$  distribution, the *noncentral t*, discussed in Example 8.3.10, is of the form  $T' = n(\mu, 1)/\sqrt{\chi_\nu^2/\nu}$ , where the normal and the chi squared random variables are independent and we can have  $\mu \neq 0$ . (We have already seen a noncentral pdf, the noncentral chi squared, in (4.4.3).) Formally, if  $X \sim n(\mu, 1)$ , and

$Y \sim \chi^2_\nu$ , independent of  $X$ , then  $T' = X/\sqrt{Y/\nu}$  has a noncentral  $t$  distribution with  $\nu$  degrees of freedom and noncentrality parameter  $\delta = \sqrt{\mu^2}$ .

- Calculate the mean and variance of  $T'$ .
- The pdf of  $T'$  is given by

$$f_{T'}(t|\delta) = \frac{e^{-\delta^2/2}}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})\sqrt{\nu}} \sum_{k=0}^{\infty} \frac{(2/\nu)^{k/2}(\delta t)^k}{k!} \frac{\Gamma([\nu+k+1]/2)}{(1+(t^2/\nu))^{(\nu+k+1)/2}}.$$

Show that this pdf reduces to that of a central  $t$  if  $\delta = 0$ .

- Starting from the definition of  $T'$ , derive its pdf.
  - Establish the claim made in Example 8.3.10, that the pdf of  $T'$  has MLR in its noncentrality parameter.
- 8.39** These next two exercises will look more closely at a few randomized tests. First, we will look at the example treated in the Miscellanea section. Let  $X_1$  and  $X_2$  be iid  $\text{Poisson}(\lambda)$  random variables, and consider the test of  $H_0: \lambda \leq 1$  versus  $H_1: \lambda > 1$  given by

$$\phi(X_1) = \begin{cases} 1 & \text{if } X_1 \geq 2 \\ 0 & \text{otherwise} \end{cases}.$$

- Show that  $\phi(X_1)$  has size .26.
- Derive the test function  $E[\phi(X_1)|X_1 + X_2]$ , and show that it also has size .26.
- Plot the power functions of  $\phi(X_1)$  and  $E[\phi(X_1)|X_1 + X_2]$ .
- We can construct another randomized test, one having test function

$$\phi'(X_1 + X_2) = \begin{cases} 1 & \text{if } X_1 + X_2 > 3 \\ .65 & \text{if } X_1 + X_2 = 3 \\ 0 & \text{if } X_1 + X_2 < 3 \end{cases}.$$

Show that  $\phi'$  has size .26, and draw a graph that shows that  $\phi'$  is more powerful than both  $\phi(X_1)$  and  $E[\phi(X_1)|X_1 + X_2]$ .

- Prove that  $\phi'$  is more powerful than the other two tests. Try to formulate a general theorem. (See Corollary 8.3.2.)

There are two types of randomized rules. The ones in this exercise, which randomize after observing the data, are called *behavioral decision rules*. Another type randomly chooses a test procedure, prior to observing the data. This type is called a *randomized decision rule*. Relationships between these two types of randomization are discussed by Ferguson (1967).

- 8.40** As another illustration of a randomized test, consider the binomial distribution. Let  $X_1, \dots, X_n$  be iid  $\text{Bernoulli}(p)$ .

- Show that the test

$$\phi(\Sigma X_i) = \begin{cases} 1 & \text{if } \Sigma X_i > k \\ \gamma & \text{if } \Sigma X_i = k \\ 0 & \text{if } \Sigma X_i < k \end{cases}$$

is UMP of size  $\alpha$  for testing  $H_0: p \leq p_0$  versus  $H_1: p > p_0$ , where  $\gamma$  satisfies

$$P(\Sigma X_i > k | p = p_0) + \gamma P(\Sigma X_i = k | p = p_0) = \alpha.$$

b. If  $n = 4$ , show that the test

$$\phi(\sum X_i) = \begin{cases} 1 & \text{if } \sum X_i > 3 \\ .144 & \text{if } \sum X_i = 3 \\ 0 & \text{if } \sum X_i < 3 \end{cases}$$

is UMP of size .05 for testing  $H_0: p \leq .3$  versus  $H_1: p > .3$ . Calculate its power at  $p = .6$ .

c. Two obvious nonrandomized competitors to the test in part (b) are

$$\phi_1(\sum X_i) = \begin{cases} 1 & \text{if } \sum X_i \geq 3 \\ 0 & \text{if } \sum X_i < 3 \end{cases}$$

and

$$\phi_2(\sum X_i) = \begin{cases} 1 & \text{if } \sum X_i > 3 \\ 0 & \text{if } \sum X_i \leq 3 \end{cases}$$

Calculate the sizes of  $\phi_1$  and  $\phi_2$ , and their powers at  $p = .6$ . Is either of them a reasonable alternative to the test in part (b)?

- 8.41** We have one observation from a beta( $1, \theta$ ) population, and we want to test  $H_0: \theta_1 \leq \theta \leq \theta_2$  versus  $H_1: \theta < \theta_1$  or  $\theta > \theta_2$ , where  $\theta_1 = 1$  and  $\theta_2 = 2$ . A test satisfies  $E_{\theta_1}\phi = .5$  and  $E_{\theta_2}\phi = .3$ . Find a test that is as good, and explain why it is as good.
- 8.42** For the same set-up as in the previous exercise, consider testing  $H_0: \theta = \theta_1$  versus  $H_1: \theta \neq \theta_1$ , with  $\theta_1 = 1$ . Find a two-sided test that satisfies  $E_{\theta_1}\phi = .1$  and

$$\frac{d}{d\theta} E_{\theta}(\phi) \Big|_{\theta=\theta_1} = 0.$$

- 8.43** Prove that if  $f(x|\theta)$  is an exponential family, then

- a. The power function of any test is differentiable.
- b. The derivative of the power function can be evaluated by differentiating under the integral sign.

- 8.44** Let  $X_1, \dots, X_n$  be a random sample from a population with pdf  $f$ , where  $f$  can be one of two specified location densities  $f_i(x - \theta)$ ,  $i = 0$  or  $1$ . It is desired to test the hypotheses

$$H_0: f(x - \theta) = f_0(x - \theta) \quad \text{versus} \quad H_1: f(x - \theta) = f_1(x - \theta),$$

using a test that is invariant under the group of transformations

$$g_a(x_1, \dots, x_n) = (x_1 + a, \dots, x_n + a),$$

where  $a$  is a constant,  $-\infty < a < \infty$ .

- a. Show that the hypothesis testing problem is invariant, as stated in Definition 8.2.4.
- b. Show that an invariant test is a function of  $(X_1, \dots, X_n)$  only through the  $n - 1$  differences  $Y = (X_1 - X_n, \dots, X_{n-1} - X_n)$ .
- c. If  $H_i$  is true,  $i = 0$  or  $1$ , show that  $Y$  has pdf

$$g_i(y) = \int_{-\infty}^{\infty} \prod_{j=1}^n f_i(x_j + t) dt = \int_{-\infty}^{\infty} \prod_{j=1}^{n-1} f_i(y_j + t) f_i(t) dt,$$

independent of  $\theta$ .

- d. Use the Neyman-Pearson Lemma to show that the UMP invariant test is of the form:  
Reject  $H_0$  if  $g_1(\mathbf{Y})/g_0(\mathbf{Y}) > C$ , for some constant  $C$ .
- e. If  $X_1, \dots, X_n$  is a random sample from a uniform( $\theta, \theta + \lambda$ ) find the UMP invariant test of  $H_0: \lambda \leq \lambda_0$  versus  $H_1: \lambda > \lambda_0$ .
- f. If  $X_1, \dots, X_n$  is a random sample from a  $n(\theta, \sigma^2)$  population find the UMP invariant test of  $H_0: \sigma \leq \sigma_0$  versus  $H_1: \sigma > \sigma_0$ . Compare this test to the LRT.
- 8.45** In each of the following situations, find an invariant test. If possible, find the UMP invariant test. Compare the test to that obtained by using the LRT.
- a.  $X_1, \dots, X_n$  is a random sample from a  $n(\mu, \sigma^2)$  population, both unknown. It is desired to test the hypotheses  $H_0: \mu \leq \mu_0$  versus  $H_1: \mu > \mu_0$  using a test that is invariant under the group of transformations  $g_c(x_1, \dots, x_n) = (cx_1, \dots, cx_n)$ , where  $c$  is a constant,  $0 < c < \infty$ . (Note the similarity to Example 8.2.6, but here we are starting with the entire sample.)
- b.  $X_1, \dots, X_n$  is a random sample from a double exponential population( $\theta, \lambda$ ) where  $\theta$  and  $\lambda$  are both unknown. Find a test of the hypotheses  $H_0: \theta \leq \theta_0$  versus  $H_1: \theta > \theta_0$  that is invariant under the group of transformations  $g_c(x_1, \dots, x_n) = (cx_1, \dots, cx_n)$ , where  $c$  is a constant,  $0 < c < \infty$ .
- c. Based on what was done in Exercise 8.44, and the first two parts of this exercise, formulate a general invariant test with respect to the scale group. Arguing analogously to the previous exercise, a best test that is invariant under the group of transformations  $g_c(x_1, \dots, x_n) = (cx_1, \dots, cx_n)$ , where  $c$  is a constant,  $0 < c < \infty$ , should be of the form

$$\text{reject } H_0 \text{ if } \frac{\int_0^\infty \prod_{j=1}^n t^{n-1} f_1(tX_j) dt}{\int_0^\infty \prod_{j=1}^n t^{n-1} f_0(tX_j) dt} > C.$$

- 8.46** Let  $X \sim \text{binomial}(n, p)$ , and consider invariant tests using the group of Example 8.2.5.
- a. Find the UMP invariant test of  $H_0: p = \frac{1}{2}$  versus  $H_1: p \neq \frac{1}{2}$ .
- b. What other hypotheses are invariant with respect to this group? Explain.
- 8.47** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population. Consider testing

$$H_0: \theta \leq \theta_0 \quad \text{versus} \quad H_1: \theta > \theta_0.$$

- a. If  $\sigma^2$  is known, show that the test that rejects  $H_0$  when

$$\bar{X} > \theta_0 + z_\alpha \sqrt{\sigma^2/n}$$

is a test of size  $\alpha$ . Show that the test can be derived as an LRT.

- b. Show that the test in part (a) is a UMP test.
- c. If  $\sigma^2$  is unknown, show that the test that rejects  $H_0$  when

$$\bar{X} > \theta_0 + t_{n-1, \alpha} \sqrt{S^2/n}$$

is a test of size  $\alpha$ . Show that the test can be derived as an LRT.

- d. Show that the test in part (c) is a UMP unbiased test.

- 8.48** Let  $X_1, \dots, X_n$  be iid  $n(\theta, \sigma^2)$ , where  $\theta_0$  is a specified value of  $\theta$ , and  $\sigma^2$  is unknown. We are interested in testing

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

- a. Show that the test that rejects  $H_0$  when

$$|\bar{X} - \theta_0| > t_{n-1, \alpha/2} \sqrt{S^2/n}$$

is a test of size  $\alpha$ .

- b. Show that the test in part (a) can be derived as an LRT.  
c. Show that the test in part (a) is a UMP unbiased test.

- 8.49** Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be a random sample from a bivariate normal distribution with parameters  $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$ . We are interested in testing

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y.$$

- a. Show that the random variables  $W_i = X_i - Y_i$  are iid  $n(\mu_W, \sigma_W^2)$ .  
b. Show that the above hypothesis can be tested with the statistic

$$T_W = \frac{\bar{W}}{\sqrt{\frac{1}{n} S_W^2}}$$

where  $\bar{W} = \frac{1}{n} \sum_{i=1}^n W_i$  and  $S_W^2 = \frac{1}{(n-1)} \sum_{i=1}^n (W_i - \bar{W})^2$ . Furthermore, show that, under  $H_0$ ,  $T_W \sim \text{Student's } t$  with  $n-1$  degrees of freedom. (This test is known as the *paired-sample t test*.)

- c. Does the paired-sample  $t$  test have any optimality properties? Prove any claims.  
**8.50** Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be a random sample from a bivariate normal distribution with parameters  $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$ .  
a. Derive the LRT of

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y,$$

where  $\sigma_X^2, \sigma_Y^2$ , and  $\rho$  are unspecified and unknown.

- b. Show that the test derived in part (a) is equivalent to the paired  $t$  test of Exercise 8.49. (*Hint:* Straightforward maximization of the bivariate likelihood is possible, but somewhat nasty. Filling in the gaps of the following argument gives a more elegant proof.)

Make the transformation  $u = x - y, v = x + y$ . Let  $f(x, y)$  denote the bivariate normal pdf, and write

$$f(x, y) = g(v|u)h(u),$$

where  $g(v|u)$  is the conditional pdf of  $V$  given  $U$ , and  $h(u)$  is the marginal pdf of  $U$ . Argue that (1) the likelihood can be equivalently factored and (2) the piece involving  $g(v|u)$  has the same maximum whether or not the means are restricted. Thus, it can be ignored (since it will cancel) and the LRT is just based only on  $h(u)$ . However,  $h(u)$  is a normal pdf with mean  $\mu_X - \mu_Y$ , and the LRT is the usual one-sample  $t$  test, as derived in Exercise 8.48.

- 8.51** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu_X, \sigma_X^2)$ , and let  $Y_1, \dots, Y_m$  be an independent random sample from a  $n(\mu_Y, \sigma_Y^2)$ . We are interested in testing

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y,$$

with the assumption that  $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ .

a. Derive the LRT for these hypotheses. Show that the LRT can be based on the statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2 \left( \frac{1}{n} + \frac{1}{m} \right)}},$$

where

$$S_p^2 = \frac{1}{(n+m-2)} \left( \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2 \right).$$

(The quantity  $S_p^2$  is sometimes referred to as a *pooled variance estimate*. This type of estimate will be used extensively in Chapter 11.)

- b. Show that, under  $H_0$ ,  $T \sim t_{n+m-2}$ . (This test is known as the *two-sample t test*.)  
 c. Does the two-sample *t* test have any optimality properties? Prove any claims.  
 d. Samples of wood were obtained from the core and periphery of a certain Byzantine church. The date of the wood was determined, giving the following data.

<i>Core</i>		<i>Periphery</i>	
1294	1251	1284	1274
1279	1248	1272	1264
1274	1240	1256	1256
1264	1232	1254	1250
1263	1220	1242	
1254	1218		
1251	1210		

Use the two-sample *t* test to determine if the mean age of the core is the same as the mean age of the periphery.

- 8.52** The assumption of equal variances, which was made in Exercise 8.51, is not always tenable. In such a case, the distribution of the statistic is no longer a *t*. Indeed, there is doubt as to the wisdom of calculating a pooled variance estimate. (This problem, of making inference on means when variances are unequal, is, in general, quite a difficult one. It is known as the *Behrens–Fisher Problem*, discussed briefly in Section 11.2.1.) A natural test to try is the following modification of the two-sample *t* test: Test

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y,$$

where we do not assume that  $\sigma_X^2 = \sigma_Y^2$ , using the statistic

$$T' = \frac{\bar{X} - \bar{Y}}{\sqrt{\left( \frac{S_X^2}{n} + \frac{S_Y^2}{m} \right)}},$$

where

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{and} \quad S_Y^2 = \frac{1}{m-1} \sum_{i=1}^m (Y_i - \bar{Y})^2.$$

The exact distribution of  $T'$  is not pleasant, but we can approximate the distribution using Satterthwaite's approximation (Example 7.2.3).

a. Show that

$$\frac{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \sim \frac{\chi_{\nu}^2}{\nu}, \quad (\text{approximately})$$

where  $\nu$  can be estimated with

$$\hat{\nu} = \frac{\left( \frac{S_X^2}{n} + \frac{S_Y^2}{m} \right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}}.$$

b. Argue that the distribution of  $T'$  can be approximated by a  $t$  distribution with  $\hat{\nu}$  degrees of freedom.

c. Re-examine the data from Exercise 8.51 using the approximate  $t$  test of this exercise, that is, test if the mean age of the core is the same as the mean age of the periphery using the  $T'$  statistic.

d. Is there any statistical evidence that the variance of the data from the core may be different from the variance of the data from the periphery? (Recall Example 5.4.1.)

**8.53** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population. Consider testing

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0.$$

Let  $\bar{X}_m$  denote the sample mean of the first  $m$  observations,  $X_1, \dots, X_m$ , for  $m = 1, \dots, n$ . If  $\sigma^2$  is known, show that for each  $m = 1, \dots, n$ , the test that rejects  $H_0$  when

$$\bar{X}_m > \theta_0 + z_\alpha \sqrt{\sigma^2/m}$$

is an unbiased size  $\alpha$  test. Graph the power function for each of these tests if  $n = 4$ .

**8.54** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population. Consider testing

$$H_0 : \theta_1 \leq \theta \leq \theta_2 \quad \text{versus} \quad H_1 : \theta < \theta_1 \text{ or } \theta > \theta_2.$$

a. Show that the test,

$$\text{reject } H_0 \text{ if } \bar{X} > \theta_2 + t_{n-1, \alpha/2} \sqrt{S^2/n} \quad \text{or} \quad \bar{X} < \theta_1 - t_{n-1, \alpha/2} \sqrt{S^2/n},$$

is not a size  $\alpha$  test.

b. Show that, for an appropriately chosen constant  $k$ , a size  $\alpha$  test is given by

$$\text{reject } H_0 \text{ if } |\bar{X} - \bar{\theta}| > k \sqrt{S^2/n},$$

where  $\bar{\theta} = (\theta_1 + \theta_2)/2$ .

c. Show that the tests in parts (a) and (b) are unbiased of their size. (Assume that the noncentral  $t$  distribution has an MLR.)

**8.55** Verify that Test 3, the UMP unbiased test in Example 8.3.9, satisfies (8.3.9) and (8.3.10).

**8.56** Let  $\beta(\theta)$  be the power function for the LMP test in Example 8.3.11. Show that if  $\alpha$  is sufficiently small then  $\lim_{\theta \rightarrow \infty} \beta(\theta) = 0$ .

**8.57** Let  $\alpha_\gamma$  be the size of the test of  $H_0$  with rejection region  $R_\gamma$ . Show that the UIT with rejection region  $R = \bigcup_{\gamma \in \Gamma} R_\gamma$  has size at least equal to  $\alpha = \sup_{\gamma \in \Gamma} \alpha_\gamma$ .

**8.58** Let  $X_n$  be a sequence of random variables that converges in distribution to a random variable  $X$ . Let  $Y_n$  be a sequence of random variables with the property that for any finite number  $c$ ,

$$\lim_{n \rightarrow \infty} P(Y_n > c) = 1.$$

Show that for any finite number  $c$ ,

$$\lim_{n \rightarrow \infty} P(X_n + Y_n > c) = 1.$$

(This is the type of result used in the discussion of the power properties of the tests described in Section 8.4.2.)

**8.59** Binomial data gathered from more than one population are often presented in a *contingency table*. For the case of two populations, the table might look like

		Population		
		1	2	Total
Successes		$S_1$	$S_2$	$S = S_1 + S_2$
		$F_1$	$F_2$	$F = F_1 + F_2$
Total		$n_1$	$n_2$	$n = n_1 + n_2$

where Population 1 is binomial( $n_1, p_1$ ), with  $S_1$  successes and  $F_1$  failures, and Population 2 is binomial( $n_2, p_2$ ), with  $S_2$  successes and  $F_2$  failures. A hypothesis that is usually of interest is

$$H_0 : p_1 = p_2 \quad \text{versus} \quad H_1 : p_1 \neq p_2.$$

a. Show that a test can be based on the statistic

$$T = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)(\hat{p}(1 - \hat{p}))},$$

where  $\hat{p}_1 = S_1/n_1$ ,  $\hat{p}_2 = S_2/n_2$ , and  $\hat{p} = (S_1 + S_2)/(n_1 + n_2)$ . Also, show that as  $n_1, n_2 \rightarrow \infty$ , the distribution of  $T$  approaches  $\chi^2_1$ . (This is a special case of a test known as a *chi squared test of independence*.)

b. Another way of measuring departure from  $H_0$  is by calculating an *expected frequency table*. This table is constructed by conditioning on the marginal totals, and filling in the table according to  $H_0: p_1 = p_2$ , that is,

		Expected frequencies		
		1	2	Total
Successes		$\frac{n_1 S}{n_1 + n_2}$	$\frac{n_2 S}{n_1 + n_2}$	$S = S_1 + S_2$
	Failures	$\frac{n_1 F}{n_1 + n_2}$	$\frac{n_2 F}{n_1 + n_2}$	$F = F_1 + F_2$
Total		$n_1$	$n_2$	$n = n_1 + n_2$

Using the expected frequency table, a statistic  $T^*$  is computed by going through the cells of the tables and computing

$$\begin{aligned} T^* &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{\left(S_1 - \frac{n_1 S}{n_1 + n_2}\right)^2}{\frac{n_1 S}{n_1 + n_2}} + \dots + \frac{\left(F_2 - \frac{n_2 F}{n_1 + n_2}\right)^2}{\frac{n_2 F}{n_1 + n_2}}. \end{aligned}$$

Show, algebraically, that  $T^* = T$  and hence that  $T^*$  is asymptotically chi squared.

c. Another statistic that could be used to test equality of  $p_1$  and  $p_2$  is

$$T^{**} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

Show that, under  $H_0$ ,  $T^{**}$  is asymptotically  $n(0, 1)$ , and hence its square is asymptotically  $\chi_1^2$ . Furthermore, show that  $(T^{**})^2 \neq T^*$ .

d. Under what circumstances is one statistic preferable to the other?

e. A famous medical experiment was conducted by Joseph Lister in the late 1800s. Mortality associated with surgery was quite high and Lister conjectured that the use of a disinfectant, carbolic acid, would help. Over a period of several years Lister performed 75 amputations with and without using carbolic acid. The data are

		Carbolic acid used?	
		Yes	No
Patient lived?	Yes	34	19
	No	6	16

Use these data to test whether the use of carbolic acid is associated with patient mortality.

- 8.60 a. Let  $(X_1, \dots, X_n) \sim \text{multinomial}(m, p_1, \dots, p_n)$ . Consider testing  $H_0: p_1 = p_2$  versus  $H_1: p_1 \neq p_2$ . A test that is often used, called *McNemar's Test*, rejects  $H_0$  if

$$\frac{(X_1 - X_2)^2}{X_1 + X_2} > \chi_{1,\alpha}^2.$$

Show that this test statistic has the form (as in Exercise 8.59)

$$\sum_1^n \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

where the  $X_i$ s are the observed cell frequencies and the expected cell frequencies are the MLEs of  $mp_i$ , under the assumption that  $p_1 = p_2$ .

b. McNemar's Test is often used in the following type of problem. Subjects are asked if they agree or disagree with a statement. Then they read some information about the statement and are asked again if they agree or disagree. The numbers of responses in each category are summarized in a  $2 \times 2$  table like this.

		<i>Before</i>	
		Agree	Disagree
<i>After</i>	Agree	$X_3$	$X_2$
	Disagree	$X_1$	$X_4$

The hypothesis  $H_0: p_1 = p_2$  states that the proportion of people who change from agree to disagree is the same as the proportion of people who change from disagree to agree. Another hypothesis that might be tested is that the proportion of those who initially agree, then change, is the same as the proportion of those who initially disagree, then change. Express this hypothesis in terms of conditional probabilities and show that it is different from the above  $H_0$ . (This hypothesis can be tested with a  $\chi^2$  test like those in Exercise 8.59.)

## Miscellanea

---

### Sufficiency and Randomized Tests

In Chapter 7 we proved the Rao–Blackwell Theorem, which said, in looking for good point estimators of a parameter, we need only to look at estimators that are functions of a sufficient statistic. We have not proved a similar theorem for hypothesis tests, and it would be nice if we had one. It does seem reasonable to expect such a theorem to exist.

Such a theorem does exist, but to properly implement it, we need to use randomized tests. This is an instance of the theoretical usefulness of randomized tests. Although we could prove a theorem similar to the one below without using randomized tests, its statement would be somewhat cumbersome.

**Theorem:** Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a random sample from a population with pdf or pmf  $f(x|\theta)$ ,  $\theta \in \Theta$ , and let  $T$  be a sufficient statistic for  $\theta$ . For any test function  $\phi(\mathbf{X})$ , there exists another (possibly randomized) test function  $\phi'(T)$ , that is a function of only the sufficient statistic, and satisfies

$$E_\theta \phi'(T) = E_\theta \phi(\mathbf{X}), \quad \text{for all } \theta \in \Theta.$$

Hence,  $\phi$  and  $\phi'$  have the same error probabilities.

*Proof:* As in the proof of the Rao–Blackwell Theorem, the conditional expectation  $E[\phi(\mathbf{X})|T]$  is a statistic that is function of  $T$  alone, and satisfies  $E_\theta(E[\phi(\mathbf{X})|T]) = E_\theta\phi(\mathbf{X})$ , for all  $\theta \in \Theta$ . Taking  $\phi'(T) = E[\phi(\mathbf{X})|T]$  completes the proof.  $\square$

When  $f(x|\theta)$  is discrete,  $E[\phi(\mathbf{X})|T]$  will almost certainly not be a test function as defined in Definition 8.2.2, but will probably be a randomized test, as the following example illustrates. Let  $X_1$  and  $X_2$  be iid Poisson( $\lambda$ ) random variables, and consider the test of  $H_0: \lambda \leq 1$  versus  $H_1: \lambda > 1$  given by

$$\phi(X_1) = \begin{cases} 1 & \text{if } X_1 \geq 2 \\ 0 & \text{otherwise} \end{cases},$$

with size given by

$$\sup_{\lambda \leq 1} E_\lambda \phi(X_1) = P(X_1 \geq 2 | \lambda = 1) = .26.$$

We know that  $X_1 + X_2$  is sufficient for  $\lambda$ , and taking conditional expectations we get

$$\begin{aligned} E[\phi(X_1)|X_1 + X_2 = t] &= P(X_1 \geq 2 | X_1 + X_2 = t) \\ &= 1 - \frac{1}{2^t} \left[ \binom{t}{0} + \binom{t}{1} \right] \end{aligned}$$

where we have used the fact that the conditional distribution of  $X_1$  given  $X_1 + X_2 = t$  is binomial (Exercise 4.15).

The randomized test  $\phi'(T)$  that rejects  $H_0$  with probability  $1 - 2^{-T} [\binom{T}{0} + \binom{T}{1}]$ , where  $T = X_1 + X_2$ , is a size .26 test, with the same power function as  $\phi(X_1)$ . (See Exercise 8.39.)

### Monotonic Power Function

In this chapter we used the property of MLR quite extensively, particularly in relation to properties of power functions of tests. The concept of *stochastic ordering* can also be used to obtain properties of power functions. (Recall that *stochastic ordering* has already been encountered in previous chapters, for example, in Exercises 1.56, 3.36–3.38, and 5.32. A cdf  $F$  is *stochastically greater* than a cdf  $G$  if  $F(x) < G(x)$  for all  $x$ , which implies that if  $X \sim F, Y \sim G$ , then  $P(X > x) > P(Y > x)$  for all  $x$ . In other words,  $F$  gives more probability to greater values.)

In terms of hypothesis testing, it is often the case that the distribution under the alternative is stochastically greater than under the null distribution. For example, if we have a random sample from a  $n(\theta, \sigma^2)$  population and are interested in testing  $H_0: \theta \leq \theta_0$  versus  $H_1: \theta > \theta_0$ , it is true that all the distributions in the alternative are stochastically greater than all those in the null. Gilat (1977) uses the property of stochastic ordering, rather than MLR, to prove monotonicity of power functions under general conditions.

### p-Values and Posterior Probabilities

In Section 8.2.3, where Bayes tests were discussed, we saw that the posterior probability that  $H_0$  is true is a measure of the evidence the data provide against (or for) the null hypothesis. We also saw, in Section 8.3.1, that p-values provide a measure of data-based evidence against  $H_0$ . A natural question to ask is whether these two different measures ever agree, that is, can they

be reconciled? Berger (James, not Roger) and Sellke (1987) contended that, in the two-sided testing problem, these measures could not be reconciled, and the Bayes measure was superior. Casella and Berger (Roger, 1987) argued that the two-sided Bayes problem is artificial, and that in the more natural one-sided problem, the measures of evidence can be reconciled. These articles are published together with discussion, and make for reasonably lively reading.

### ***Unbiasedness and Invariance***

Lehmann (1986) has an interesting discussion about the relationship between unbiasedness and invariance, which we might interpret as saying that, except in rare circumstances, these concepts are complementary. That is, we can expect only one concept to be useful in any given problem.

An interesting theorem is also given. It says that if a unique UMP unbiased test exists and if a unique UMP invariant test exists, then the tests coincide. This situation, as we might expect, is rather rare and it seems that, if we are outside of this case, we can benefit from either unbiasedness or invariance but not both. Furthermore, Lehmann shows that, in situations where *neither* unbiasedness nor invariance leads to a good solution, a combination of these ideas can be helpful in obtaining a solution.

# 9 Interval Estimation

*"As a rule," said Holmes, "the more bizarre a thing is the less mysterious it proves to be."*

**Sherlock Holmes**  
*The Red-Headed League*

## 9.1 Introduction

In Chapter 7 we discussed point estimation of a parameter  $\theta$ , where the inference is a guess of a single value as the value of  $\theta$ . In this chapter we discuss interval estimation and, more generally, set estimation. The inference in a set estimation problem is the statement that " $\theta \in C$ " where  $C \subset \Theta$  and  $C = C(\mathbf{x})$  is a set determined by the value of the data  $\mathbf{X} = \mathbf{x}$  observed. If  $\theta$  is real-valued, then we usually prefer the set estimate  $C$  to be an interval. Interval estimators will be the main topic of this chapter.

As in the previous two chapters, this chapter is divided into two parts, the first concerned with finding interval estimators and the second part concerned with evaluating the worth of the estimators. We begin with a formal definition of interval estimator, a definition as vague as the definition of point estimator.

**DEFINITION 9.1.1:** An *interval estimate* of a real-valued parameter  $\theta$  is any pair of functions,  $L(x_1, \dots, x_n)$  and  $U(x_1, \dots, x_n)$ , of a sample that satisfy  $L(\mathbf{x}) \leq U(\mathbf{x})$  for all  $\mathbf{x} \in \mathcal{X}$ . If  $\mathbf{X} = \mathbf{x}$  is observed, the inference  $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$  is made. The random interval  $[L(\mathbf{X}), U(\mathbf{X})]$  is called an *interval estimator*.

We will use our previously defined conventions and write  $[L(\mathbf{X}), U(\mathbf{X})]$  for an interval estimator of  $\theta$  based on the random sample  $\mathbf{X} = (X_1, \dots, X_n)$  and  $[L(\mathbf{x}), U(\mathbf{x})]$  for the realized value of the interval. Although in the majority of cases we will work with finite values for  $L$  and  $U$ , there is sometimes interest in *one-sided* interval estimates. For instance, if  $L(\mathbf{x}) = -\infty$ , then we have the one-sided interval  $(-\infty, U(\mathbf{x}))$  and the assertion is that " $\theta \leq U(\mathbf{x})$ ," with no mention of a lower bound. We could similarly take  $U(\mathbf{x}) = \infty$  and have a one-sided interval  $[L(\mathbf{x}), \infty)$ .

Although the definition mentions a closed interval  $[L(\mathbf{x}), U(\mathbf{x})]$ , it will sometimes be more natural to use an open interval  $(L(\mathbf{x}), U(\mathbf{x}))$  or even a half-open and half-closed interval, as in the previous paragraph. We will use whichever seems most appropriate for the particular problem at hand although the preference will be for a closed interval.

**Example 9.1.1:** For a sample  $X_1, X_2, X_3, X_4$  from a  $n(\mu, 1)$ , an interval estimator of  $\mu$  is  $[\bar{X} - 1, \bar{X} + 1]$ . This means that we will assert that  $\mu$  is in this interval. ||

At this point, it is natural to inquire as to what is gained by using an interval estimator. Previously, we estimated  $\mu$  with  $\bar{X}$  and now we have the less precise estimator  $[\bar{X} - 1, \bar{X} + 1]$ . We surely must gain something! By giving up some precision in our estimate (or assertion about  $\mu$ ), we have gained some confidence, or assurance, that our assertion is correct.

**Example 9.1.1 (Continued):** When we estimate  $\mu$  by  $\bar{X}$ , the probability that we are exactly correct, that is,  $P(\bar{X} = \mu)$ , is zero. However, with an interval estimator, we have a positive probability of being correct. The probability that  $\mu$  is covered by the interval  $[\bar{X} - 1, \bar{X} + 1]$  can be calculated as

$$\begin{aligned} P(\mu \in [\bar{X} - 1, \bar{X} + 1]) &= P(\bar{X} - 1 \leq \mu \leq \bar{X} + 1) \\ &= P(-1 \leq \bar{X} - \mu \leq 1) \\ &= P\left(-2 \leq \frac{\bar{X} - \mu}{\sqrt{1/4}} \leq 2\right) \\ &= P(-2 \leq Z \leq 2) \quad \left( \frac{\bar{X} - \mu}{\sqrt{1/4}} \text{ is standard normal} \right) \\ &= .9544. \end{aligned} \tag{Table 1}$$

Thus we have over a 95% chance of covering the unknown parameter with our interval estimator. Sacrificing some precision in our estimate, in moving from a point to an interval, has resulted in increased confidence that our assertion is correct. ||

The purpose of using an interval estimator, rather than a point estimator, is to have some guarantee of capturing the parameter of interest. The certainty of this guarantee is quantified in the following definitions.

**DEFINITION 9.1.2:** For an interval estimator  $[L(\mathbf{X}), U(\mathbf{X})]$  of a parameter  $\theta$ , the *coverage probability* of  $[L(\mathbf{X}), U(\mathbf{X})]$  is the probability that the random interval  $[L(\mathbf{X}), U(\mathbf{X})]$  covers the true parameter,  $\theta$ . In symbols, it is denoted by either  $P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$  or  $P(\theta \in [L(\mathbf{X}), U(\mathbf{X})]|\theta)$ .

**DEFINITION 9.1.3:** For an interval estimator  $[L(\mathbf{X}), U(\mathbf{X})]$  of a parameter  $\theta$ , the *confidence coefficient* of  $[L(\mathbf{X}), U(\mathbf{X})]$  is the infimum of the coverage probabilities,  $\inf_\theta P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$ .

There are a number of things to be aware of in these definitions. One, it is important to keep in mind that the *interval* is the random quantity, not the parameter. Therefore, when we write probability statements such as  $P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$ , these probability statements refer to  $\mathbf{X}$ , not  $\theta$ . In other words, think of  $P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$ , which might look like a statement about a random  $\theta$ , as the algebraically equivalent  $P_\theta(L(\mathbf{X}) \leq \theta, U(\mathbf{X}) \geq \theta)$ , a statement about a random  $\mathbf{X}$ .

Interval estimators, together with a measure of confidence (usually a confidence coefficient) are sometimes known as *confidence intervals*. We will often use this

term interchangeably with *interval estimator*. Although we are mainly concerned with confidence *intervals*, we occasionally will work with more general sets. When working in general, and not being quite sure of the exact form of our sets, we will speak of confidence *sets*. A confidence set with confidence coefficient equal to some value, say  $1 - \alpha$ , is simply called a  $1 - \alpha$  *confidence set*.

Another important point is concerned with coverage probabilities and confidence coefficients. Since we do not know the true value of  $\theta$ , we can only guarantee a coverage probability equal to the infimum, the confidence coefficient. In some cases this does not matter because the coverage probability will be a constant function of  $\theta$ . In other cases, however, the coverage probability can be a fairly variable function of  $\theta$ .

**Example 9.1.2:** Let  $X_1, \dots, X_n$  be a random sample from a uniform( $0, \theta$ ) population and let  $Y = \max\{X_1, \dots, X_n\}$ . We are interested in an interval estimator of  $\theta$ . We consider two candidate estimators:  $[aY, bY]$ ,  $1 \leq a < b$ , and  $[Y + c, Y + d]$ ,  $0 \leq c < d$ , where  $a, b, c$ , and  $d$  are specified constants. (Note that  $\theta$  is necessarily larger than  $y$ .) For the first interval we have

$$\begin{aligned} P_\theta(\theta \in [aY, bY]) &= P_\theta(aY \leq \theta \leq bY) \\ &= P_\theta\left(\frac{1}{b} \leq \frac{Y}{\theta} \leq \frac{1}{a}\right) \\ &= P_\theta\left(\frac{1}{b} \leq T \leq \frac{1}{a}\right). \quad (T = Y/\theta) \end{aligned}$$

We previously saw (Example 7.3.5) that  $f_Y(y) = ny^{n-1}/\theta^n$ ,  $0 \leq y \leq \theta$ , so the pdf of  $T$  is  $f_T(t) = nt^{n-1}$ ,  $0 \leq t \leq 1$ . We therefore have

$$P_\theta\left(\frac{1}{b} \leq T \leq \frac{1}{a}\right) = \int_{1/b}^{1/a} nt^{n-1} dt = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n.$$

The coverage probability of the first interval is independent of the value of  $\theta$  and thus  $(\frac{1}{a})^n - (\frac{1}{b})^n$  is the confidence coefficient of the interval.

For the other interval, for  $\theta \geq d$  a similar calculation yields

$$\begin{aligned} P_\theta(\theta \in [Y + c, Y + d]) &= P_\theta(Y + c \leq \theta \leq Y + d) \\ &= P_\theta\left(1 - \frac{d}{\theta} \leq T \leq 1 - \frac{c}{\theta}\right) \quad (T = Y/\theta) \\ &= \int_{1-d/\theta}^{1-c/\theta} nt^{n-1} dt = \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n. \end{aligned}$$

In this case, the coverage probability depends on  $\theta$ . Furthermore, it is straightforward to calculate

$$\lim_{\theta \rightarrow \infty} \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n = 0,$$

showing that the confidence coefficient of this interval estimator is zero. ||

## 9.2 Methods of Finding Interval Estimators

We will examine five different methods of finding interval estimators, although there is some overlap in the methods. We start with the method of test inversion, a fairly general technique.

### 9.2.1 Inverting a Test Statistic

There is a very strong correspondence between hypothesis testing and interval estimation. In fact, we can say in general that every confidence set corresponds to a test and vice versa. Consider the following example.

**Example 9.2.1:** Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$  and consider testing  $H_0: \mu = \mu_0$  versus  $H_1: \mu \neq \mu_0$ . For a fixed  $\alpha$  level, a reasonable test (in fact, UMP unbiased) has rejection region  $\{\mathbf{x} : |\bar{x} - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}\}$ . Note that  $H_0$  is accepted for sample points with  $|\bar{x} - \mu_0| \leq z_{\alpha/2}\sigma/\sqrt{n}$ , or equivalently,

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Since the test has size  $\alpha$ , this means that  $P(H_0 \text{ is rejected} | \mu = \mu_0) = \alpha$ , or, stated in another way,  $P(H_0 \text{ is accepted} | \mu = \mu_0) = 1 - \alpha$ . Combining this with the above characterization of the acceptance region, we can write

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_0\right) = 1 - \alpha.$$

But this probability statement is true for every  $\mu_0$ . Hence, the statement

$$P_\mu\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

is true. The interval  $[\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}]$ , obtained by *inverting* the acceptance region of the level  $\alpha$  test, is a  $1 - \alpha$  confidence interval. ||

We have illustrated the correspondence between confidence sets and tests. The acceptance region of the hypothesis test, the set in the *sample space* for which  $H_0: \mu = \mu_0$  is accepted, is given by

$$A(\mu_0) = \left\{(\mathbf{x}_1, \dots, \mathbf{x}_n) : \mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right\},$$

and the *confidence interval*, the set in the *parameter space* with plausible values of  $\mu$ , is given by

$$C(x_1, \dots, x_n) = \left\{ \mu : \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\}.$$

These sets are connected to each other by the tautology

$$(x_1, \dots, x_n) \in A(\mu_0) \Leftrightarrow \mu_0 \in C(x_1, \dots, x_n).$$

The correspondence between testing and interval estimation for the two-sided normal problem is illustrated in Figure 9.2.1. There it is, perhaps, more easily seen that both tests and intervals ask the same question, but from a slightly different perspective. Both procedures look for consistency between sample statistics and population parameters. The hypothesis test fixes the parameter and asks what sample values (the acceptance region) are consistent with that fixed value. The confidence set fixes the sample value and asks what parameter values (the confidence interval) make this sample value most plausible.

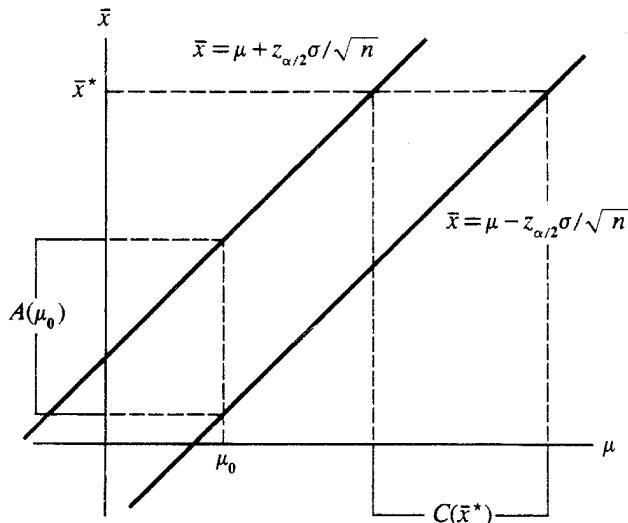


FIGURE 9.2.1 Relationship between confidence intervals and acceptance regions for tests

The correspondence between acceptance regions of tests and confidence sets holds in general. The next theorem gives a formal version of this correspondence.

**THEOREM 9.2.1:** For each  $\theta_0 \in \Theta$ , let  $A(\theta_0)$  be the acceptance region of a level  $\alpha$  test of  $H_0: \theta = \theta_0$ . For each  $x \in \mathcal{X}$ , define a set  $C(x)$  in the parameter space by

$$(9.2.1) \quad C(x) = \{\theta_0: x \in A(\theta_0)\}.$$

Then the random set  $C(X)$  is a  $1 - \alpha$  confidence set. Conversely, let  $C(X)$  be a  $1 - \alpha$  confidence set. For any  $\theta_0 \in \Theta$ , define

$$A(\theta_0) = \{x: \theta_0 \in C(x)\}.$$

Then  $A(\theta_0)$  is the acceptance region of a level  $\alpha$  test of  $H_0: \theta = \theta_0$ .

*Proof:* For the first part, since  $A(\theta_0)$  is the acceptance region of a level  $\alpha$  test,

$$P_{\theta_0}(\mathbf{X} \notin A(\theta_0)) \leq \alpha \quad \text{and, hence,} \quad P_{\theta_0}(\mathbf{X} \in A(\theta_0)) \geq 1 - \alpha.$$

Since  $\theta_0$  is arbitrary, write  $\theta$  instead of  $\theta_0$ . The above inequality, together with (9.2.1), shows that the coverage probability of the set  $C(\mathbf{X})$  is given by

$$P_\theta(\theta \in C(\mathbf{X})) = P_\theta(\mathbf{X} \in A(\theta)) \geq 1 - \alpha,$$

showing that  $C(\mathbf{X})$  is a  $1 - \alpha$  confidence set.

For the second part, the Type I Error probability for the test of  $H_0: \theta = \theta_0$  with acceptance region  $A(\theta_0)$  is

$$P_{\theta_0}(\mathbf{X} \notin A(\theta_0)) = P_{\theta_0}(\theta_0 \notin C(\mathbf{X})) \leq \alpha.$$

So this is a level  $\alpha$  test. □

Although it is common to talk about inverting a test to obtain a confidence set, Theorem 9.2.1 makes it clear that we really have a family of tests, one for each value of  $\theta_0 \in \Theta$ , that we invert to obtain one confidence set.

The fact that tests can be inverted to obtain a confidence set and vice versa is theoretically interesting, but the really useful part of Theorem 9.2.1 is the first part. It is a relatively easy task to construct a level  $\alpha$  acceptance region. The difficult task is constructing a confidence set. So the method of obtaining a confidence set by inverting an acceptance region is quite useful. All of the techniques we have for obtaining tests can immediately be applied to constructing confidence sets.

In Theorem 9.2.1, we stated only the null hypothesis  $H_0: \theta = \theta_0$ . All that is required of the acceptance region is

$$P_{\theta_0}(\mathbf{X} \in A(\theta_0)) \geq 1 - \alpha.$$

In practice, when constructing a confidence set by test inversion, we will also have in mind an alternative hypothesis such as  $H_1: \theta \neq \theta_0$  or  $H_1: \theta > \theta_0$ . The alternative will dictate the form of  $A(\theta_0)$  that is reasonable, and the form of  $A(\theta_0)$  will determine the shape of  $C(\mathbf{x})$ . Note, however, that we carefully used the word set, rather than interval. This is because there is no guarantee that the confidence set obtained by test inversion will be an interval. In most cases, however, one-sided tests give one-sided intervals, two-sided tests give two-sided intervals, strange-shaped acceptance regions give strange-shaped confidence sets. Later examples will exhibit this.

The properties of the inverted test also carry over (sometimes suitably modified) to the confidence set. For example, invariant tests, when inverted, will produce invariant confidence sets. Also, and more importantly, since we know that we can confine attention to sufficient statistics when looking for a good test, it follows that we can confine attention to sufficient statistics when looking for good confidence sets (see the *Miscellanea* section).

The method of test inversion really is most helpful in situations where our intuition deserts us and we have no good idea as to what would constitute a reasonable set. We merely fall back on our all-purpose method for constructing a reasonable test.

**Example 9.2.2:** Suppose that we want a confidence interval for the mean,  $\lambda$ , of an exponential( $\lambda$ ) population. We can obtain such an interval by inverting a level  $\alpha$  test of  $H_0: \lambda = \lambda_0$  versus  $H_1: \lambda \neq \lambda_0$ .

If we take a random sample  $X_1, \dots, X_n$ , the LRT statistic is given by

$$\frac{\frac{1}{\lambda_0^n} e^{-\sum x_i / \lambda_0}}{\sup_{\lambda} \frac{1}{\lambda^n} e^{-\sum x_i / \lambda}} = \frac{\frac{1}{\lambda_0^n} e^{-\sum x_i / \lambda_0}}{\left(\frac{1}{\sum x_i / n}\right)^n e^{-n}} = \left(\frac{\sum x_i}{n \lambda_0}\right)^n e^n e^{-\sum x_i / \lambda_0}.$$

For fixed  $\lambda_0$ , the acceptance region is given by

$$(9.2.2) \quad A(\lambda_0) = \left\{ \mathbf{x}: \left( \frac{\sum x_i}{\lambda_0} \right)^n e^{-\sum x_i / \lambda_0} \geq k^* \right\}$$

where  $k^*$  is a constant chosen to satisfy  $P_{\lambda_0}(\mathbf{X} \in A(\lambda_0)) = 1 - \alpha$ . (The constant  $e^n/n^n$  has been absorbed into  $k^*$ .) This is a set in the sample space as shown in Figure 9.2.2. Inverting this acceptance region gives the  $1 - \alpha$  confidence set

$$C(\mathbf{x}) = \left\{ \lambda: \left( \frac{\sum x_i}{\lambda} \right)^n e^{-\sum x_i / \lambda} \geq k^* \right\}.$$

This is an interval in the parameter space as shown in Figure 9.2.2. The expression

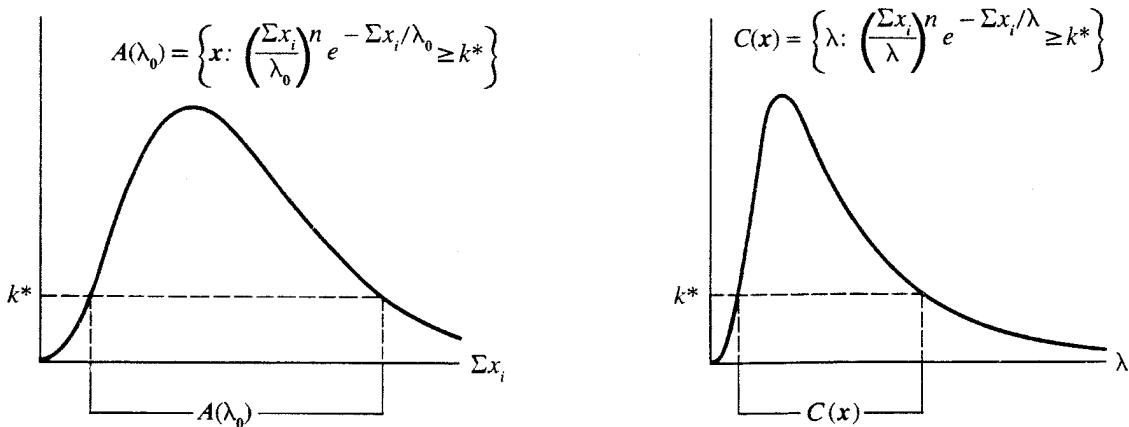


FIGURE 9.2.2 Acceptance region and confidence interval for Example 9.2.2

defining  $C(\mathbf{x})$  depends on  $\mathbf{x}$  only through  $\sum x_i$ . So the confidence interval can be expressed in the form

$$(9.2.3) \quad C(\sum x_i) = \{ \lambda: L(\sum x_i) \leq \lambda \leq U(\sum x_i) \}$$

where  $L$  and  $U$  are functions determined by the constraints that the set (9.2.2) has probability  $1 - \alpha$  and

$$(9.2.4) \quad \left( \frac{\sum x_i}{L(\sum x_i)} \right)^n e^{-\sum x_i / L(\sum x_i)} = \left( \frac{\sum x_i}{U(\sum x_i)} \right)^n e^{-\sum x_i / U(\sum x_i)}.$$

Note that if we take

$$(9.2.5) \quad L(\sum x_i) = a \sum x_i \quad \text{and} \quad U(\sum x_i) = b \sum x_i,$$

where  $a < b$  are constants, then (9.2.4) becomes

$$(9.2.6) \quad \left( \frac{1}{a} \right)^n e^{-1/a} = \left( \frac{1}{b} \right)^n e^{-1/b},$$

which yields easily to numerical solution. To work out some details, let  $n = 2$  and note that  $\sum X_i \sim \text{gamma}(2, \lambda)$  and  $\sum X_i / \lambda \sim \text{gamma}(2, 1)$ . Hence, using (9.2.5), the confidence interval becomes  $\{\lambda : a \sum x_i \leq \lambda \leq b \sum x_i\}$ , where  $a$  and  $b$  satisfy

$$P_\lambda \left( a \sum X_i \leq \lambda \leq b \sum X_i \right) = P \left( \frac{1}{b} \leq \frac{\sum X_i}{\lambda} \leq \frac{1}{a} \right) = 1 - \alpha$$

and, from (9.2.6),

$$\left( \frac{1}{a} \right)^2 e^{-1/a} = \left( \frac{1}{b} \right)^2 e^{-1/b}.$$

We have

$$\begin{aligned} P \left( \frac{1}{b} \leq \frac{\sum X_i}{\lambda} \leq \frac{1}{a} \right) &= \int_{1/b}^{1/a} t e^{-t} dt \\ &= e^{-1/b} \left( \frac{1}{b} + 1 \right) - e^{-1/a} \left( \frac{1}{a} + 1 \right). \quad (\text{integration by parts}) \end{aligned}$$

To get, for example, a 90% confidence interval, we must simultaneously satisfy the probability condition and the constraint. To three decimal places, we get  $a = .182$ ,  $b = 2.280$ , with a confidence coefficient of .901. Thus,

$$P_\lambda \left( (.182) \sum X_i \leq \lambda \leq (2.280) \sum X_i \right) = .901. \quad \|$$

The test inversion method is completely general in that we can invert any test and obtain a confidence set. Here we inverted LRTs, but we could have used a test constructed by any method. Also, note that the inversion of a two-sided test gave a two-sided interval. In the next examples, we invert one-sided tests to get one-sided intervals.

**Example 9.2.3:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population. Consider constructing a  $1 - \alpha$  upper confidence bound for  $\mu$ . That is, we want a confidence interval of the form  $C(\mathbf{x}) = (-\infty, U(\mathbf{x})]$ . To obtain such an interval using Theorem 9.2.1, we will invert one-sided tests of  $H_0: \mu = \mu_0$  versus  $H_1: \mu < \mu_0$ . (Note that we use the specification of  $H_1$  to determine the form of the confidence interval here.  $H_1$  specifies “large” values of  $\mu_0$ , so the confidence set will contain “small” values, values less than a bound. Thus, we will get an upper confidence bound.) The size  $\alpha$  LRT of  $H_0$  versus  $H_1$  rejects  $H_0$  if

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} < -t_{n-1, \alpha}$$

(similar to Example 8.2.4). Thus the acceptance region for this test is

$$A(\mu_0) = \left\{ \mathbf{x} : \bar{x} \geq \mu_0 - t_{n-1, \alpha} \frac{s}{\sqrt{n}} \right\}$$

and  $\mathbf{x} \in A(\mu_0) \Leftrightarrow \bar{x} + t_{n-1, \alpha} s / \sqrt{n} \geq \mu_0$ . According to (9.2.1), we define

$$C(\mathbf{x}) = \left\{ \mu_0 : \mathbf{x} \in A(\mu_0) \right\} = \left\{ \mu_0 : \bar{x} + t_{n-1, \alpha} \frac{s}{\sqrt{n}} \geq \mu_0 \right\}.$$

By Theorem 9.2.1, the random set  $C(\mathbf{X}) = (-\infty, \bar{X} + t_{n-1, \alpha} S / \sqrt{n}]$  is a  $1 - \alpha$  confidence set for  $\mu$ . We see that, indeed, it is the right form for an upper confidence bound. Inverting the one-sided test gave a one-sided confidence interval. ||

**Example 9.2.4:** As a more difficult example of a one-sided confidence interval, consider putting a  $1 - \alpha$  lower confidence bound on  $p$ , the success probability from a sequence of Bernoulli trials. That is, we observe  $X_1, \dots, X_n$ , where  $X_i \sim \text{Bernoulli}(p)$ , and we want the interval to be of the form  $(L(x_1, \dots, x_n), 1]$ , where  $P_p(p \in (L(X_1, \dots, X_n), 1]) \geq 1 - \alpha$ . (The interval we obtain turns out to be open on the left, as will be seen.)

Since we want a one-sided interval that gives a lower confidence bound, we consider inverting the acceptance regions from tests of

$$H_0: p = p_0 \quad \text{versus} \quad H_1: p > p_0.$$

To simplify things, we know that we can base our test on  $T = \sum_{i=1}^n X_i \sim \text{binomial}(n, p)$ , since  $T$  is sufficient for  $p$ . (See the *Miscellanea* section.) Since the binomial distribution has monotone likelihood ratio (Exercise 8.28), by the Karlin-Rubin Theorem (Theorem 8.3.2) the test that rejects  $H_0$  if  $T > k(p_0)$  is the UMP test of its size. For each  $p_0$ , we want to choose the constant  $k(p_0)$  (it can be an integer) so that we have a level  $\alpha$  test. We cannot get the size of the test to be exactly  $\alpha$ , except for certain values of  $p_0$ , because of the discreteness of  $T$ . But we choose  $k(p_0)$  so that the size of the test is as close to  $\alpha$  as possible, without being larger. Thus,  $k(p_0)$  is defined to be the integer between 0 and  $n$  that simultaneously satisfies the inequalities

$$(9.2.7) \quad \sum_{y=0}^{k(p_0)} \binom{n}{y} p_0^y (1-p_0)^{n-y} \geq 1 - \alpha$$

and

$$\sum_{y=0}^{k(p_0)-1} \binom{n}{y} p_0^y (1-p_0)^{n-y} < 1 - \alpha.$$

Because of the MLR property of the binomial, for any  $k = 0, \dots, n$ , the quantity

$$f(p_0|k) = \sum_{y=0}^k \binom{n}{y} p_0^y (1-p_0)^{n-y}$$

is a decreasing function of  $p_0$  (Exercise 8.29). Of course,  $f(0|0) = 1$ , so  $k(0) = 0$  and  $f(p_0|0)$  remains above  $1 - \alpha$  for an interval of values. Then, at some point  $f(p_0|0) = 1 - \alpha$  and for values of  $p_0$  greater than this value  $f(p_0|0) < 1 - \alpha$ . So, at this point,  $k(p_0)$  increases to one. This pattern continues. Thus,  $k(p_0)$  is an integer-valued *step-function*. It is constant for a range of  $p_0$ , then it jumps to the next bigger integer. Since  $k(p_0)$  is a nondecreasing function of  $p_0$ , this gives the lower confidence bound. (See Exercise 9.5 for an upper confidence bound.) Solving the inequalities in (9.2.7) for  $k(p_0)$  gives both the acceptance region of the test and the confidence set.

For each  $p_0$ , the acceptance region is given by  $A(p_0) = \{t : t \leq k(p_0)\}$ , where  $k(p_0)$  satisfies (9.2.7). For each value of  $t$  the confidence set is  $C(t) = \{p_0 : t \leq k(p_0)\}$ . This set, in its present form, however, does not do us much practical good. Although it is formally correct and a  $1 - \alpha$  confidence set, it is defined implicitly in terms of  $p_0$  and we want it to be defined explicitly in terms of  $p_0$ .

Since  $k(p_0)$  is nondecreasing, for a given observation  $T = t$ ,  $k(p_0) < t$  for all  $p_0$  less than or equal to some value, call it  $k^{-1}(t)$ . At  $k^{-1}(t)$ ,  $k(p_0)$  jumps up to equal  $t$  and  $k(p_0) \geq t$  for all  $p_0 > k^{-1}(t)$ . (Note that at  $p_0 = k^{-1}(t)$ ,  $f(p_0|t-1) = 1 - \alpha$ . So (9.2.7) is still satisfied by  $k(p_0) = t - 1$ . Only for  $p_0 > k^{-1}(t)$  is  $k(p_0) \geq t$ .) Thus, the confidence set is

$$(9.2.8) \quad C(t) = \{p_0 : t \leq k(p_0)\} = \{p_0 : p_0 > k^{-1}(t)\}$$

and we have constructed a  $1 - \alpha$  lower confidence bound of the form  $C(T) = (k^{-1}(T), 1]$ .

The number  $k^{-1}(t)$  can be defined as

$$(9.2.9) \quad k^{-1}(t) = \sup \left\{ p : \sum_{y=0}^{t-1} \binom{n}{y} p^y (1-p)^{n-y} \geq 1 - \alpha \right\}.$$

Realize that  $k^{-1}(t)$  is not really an inverse of  $k(p_0)$  because  $k(p_0)$  is not a one-to-one function. However, the expressions in (9.2.7) and (9.2.9) give us well-defined quantities for  $k$  and  $k^{-1}$ .

This problem of binomial confidence bounds was first treated by Clopper and Pearson (1934), who obtained answers similar to these.

### 9.2.2 Pivotal Quantities

Perhaps one of the most elegant methods of constructing set estimators and calculating coverage probabilities is the use of pivotal quantities. The use of pivotal quantities for confidence set construction, resulting in what has been called *pivotal inference*, is mainly due to G. A. Barnard (1949, 1980). Closely related is D. A. S. Fraser's theory of *structural inference* (Fraser, 1968, 1979). An interesting discussion of the strengths and weaknesses of these methods is given in Berger and Wolpert (1984).

**DEFINITION 9.2.1:** A random variable  $Q(\mathbf{X}, \theta) = Q(X_1, \dots, X_n, \theta)$ , is a *pivotal quantity* (or *pivot*) if the distribution of  $Q(\mathbf{X}, \theta)$  is independent of all parameters. That is, if  $\mathbf{X} \sim F(\mathbf{x}|\theta)$ , then  $Q(\mathbf{X}, \theta)$  has the same distribution for all values of  $\theta$ .

The function  $Q(\mathbf{x}, \theta)$  will usually explicitly contain both parameters and statistics, but for any set  $\mathcal{A}$ ,  $P_\theta(Q(\mathbf{X}, \theta) \in \mathcal{A})$  cannot depend on  $\theta$ . The technique of constructing confidence sets from pivots relies on being able to find a pivot and a set  $\mathcal{A}$  so that the set  $\{\theta : Q(\mathbf{x}, \theta) \in \mathcal{A}\}$  is a set estimate of  $\theta$ .

**Example 9.2.5:** In location and scale cases there are lots of pivotal quantities. We will show a few here; more will be found in Exercise 9.8. Let  $X_1, \dots, X_n$  be a random sample from the indicated pdfs and let  $\bar{X}$  and  $S$  be the sample mean and standard deviation.

Form of pdf	Type of pdf	Pivotal quantity
$f(x - \mu)$	location	$\bar{X} - \mu$
$\frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$	scale	$\frac{\bar{X}}{\sigma}$
$\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$	location-scale	$\frac{\bar{X}-\mu}{S}$

To prove that the above quantities are pivots, we just have to show that their pdfs are independent of parameters (details in Exercise 9.9). Notice that, in particular, if  $X_1, \dots, X_n$  is a random sample from a  $n(\mu, \sigma^2)$  population, then the  $t$  statistic  $(\bar{X} - \mu)/(S/\sqrt{n})$  is a pivot because the  $t$  distribution does not depend on the parameters  $\mu$  and  $\sigma^2$ .

Of the intervals constructed in Section 9.2.1 using the test inversion method, some turned out to be based on pivots (Examples 9.2.2 and 9.2.3) and some did not (Example 9.2.4). There is no all-purpose strategy for finding pivots. However, we can be a little clever and not rely totally on guesswork. For example, it is a relatively easy task to find pivots for location or scale parameters. In general, *differences* are pivotal for location problems, while *ratios* (or products) are pivotal for scale problems.

**Example 9.2.6:** Suppose that  $X_1, \dots, X_n$  are iid exponential( $\lambda$ ). Then  $T = \sum X_i$  is a sufficient statistic for  $\lambda$  and  $T \sim \text{gamma}(n, \lambda)$ . In the gamma pdf  $t$  and  $\lambda$  appear together as  $t/\lambda$  and, in fact the gamma( $n; \lambda$ ) pdf  $(\Gamma(n)\lambda^n)^{-1}t^{n-1}e^{-t/\lambda}$  is a scale family. Thus, if  $Q(T, \lambda) = 2T/\lambda$ , then

$$Q(T, \lambda) \sim \text{gamma}(n, \lambda(2/\lambda)) = \text{gamma}(n, 2),$$

which does not depend on  $\lambda$ . The quantity  $Q(T, \lambda) = 2T/\lambda$  is a pivot with a  $\text{gamma}(n, 2)$ , or  $\chi^2_{2n}$ , distribution.  $\parallel$

We can sometimes look to the form of the pdf to see if a pivot exists. In the above example, the quantity  $t/\lambda$  appeared in the pdf and this turned out to be a pivot. In the normal pdf, the quantity  $(\bar{x} - \mu)/\sigma$  appears and this quantity is also a pivot. In general, suppose the pdf of a statistic  $T$ ,  $f(t|\theta)$ , can be expressed in the form

$$f(t|\theta) = g(Q(t, \theta)) \left| \frac{\partial}{\partial t} Q(t, \theta) \right|,$$

for some function  $g$  and some monotone function  $Q$  (monotone in  $t$  for each  $\theta$ ). Then Theorem 2.1.2 can be used to show (Exercise 9.10) that  $Q(T, \theta)$  is a pivot.

Once we have a pivot, how do we use it to construct a confidence set? That part is really quite simple. If  $Q(\mathbf{X}, \theta)$  is a pivot, then for a specified value of  $\alpha$  we can find numbers  $a$  and  $b$ , which do not depend on  $\theta$ , to satisfy

$$(9.2.10) \quad P_\theta(a \leq Q(\mathbf{X}, \theta) \leq b) \geq 1 - \alpha.$$

Then, for each  $\theta_0 \in \Theta$ ,

$$A(\theta_0) = \{\mathbf{x}: a \leq Q(\mathbf{x}, \theta_0) \leq b\}$$

is the acceptance region for a level  $\alpha$  test of  $H_0: \theta = \theta_0$ . We will use the test inversion method to construct the confidence set but we are using the pivot to specify the specific form of our acceptance regions. Using Theorem 9.2.1, we invert these tests to obtain

$$C(\mathbf{x}) = \{\theta_0: a \leq Q(\mathbf{x}, \theta_0) \leq b\}$$

and  $C(\mathbf{x})$  is a  $1 - \alpha$  confidence set for  $\theta$ . If  $\theta$  is a real-valued parameter and if, for each  $\mathbf{x} \in \mathcal{X}$ ,  $Q(\mathbf{x}, \theta)$  is a monotone function of  $\theta$ , then  $C(\mathbf{x})$  will be an interval. In fact, if  $Q(\mathbf{x}, \theta)$  is an increasing function of  $\theta$  then  $C(\mathbf{x})$  has the form  $L(\mathbf{x}, a) \leq \theta \leq U(\mathbf{x}, b)$ . If  $Q(\mathbf{x}, \theta)$  is a decreasing function of  $\theta$  (which is typical), then  $C(\mathbf{x})$  has the form  $L(\mathbf{x}, b) \leq \theta \leq U(\mathbf{x}, a)$ .

**Example 9.2.6 (Continued):** In Example 9.2.2 we obtained a confidence interval for the mean,  $\lambda$ , of the exponential( $\lambda$ ) pdf by inverting a level  $\alpha$  LRT of  $H_0: \lambda = \lambda_0$

versus  $H_1: \lambda \neq \lambda_0$ . Now we also see that if we have a sample  $X_1, \dots, X_n$ , we can define  $T = \sum X_i$  and  $Q(T, \lambda) = 2T/\lambda \sim \chi^2_{2n}$ .

If we choose constants  $a$  and  $b$  to satisfy  $P(a \leq \chi^2_{2n} \leq b) = 1 - \alpha$ , then

$$P_\lambda \left( a \leq \frac{2T}{\lambda} \leq b \right) = P_\lambda(a \leq Q(T, \lambda) \leq b) = P(a \leq \chi^2_{2n} \leq b) = 1 - \alpha.$$

Inverting the set  $A(\lambda) = \{t: a \leq \frac{2t}{\lambda} \leq b\}$  gives  $C(t) = \{\lambda: \frac{2t}{b} \leq \lambda \leq \frac{2t}{a}\}$ , which is a  $1 - \alpha$  confidence interval. (Notice that the lower endpoint depends on  $b$  and the upper endpoint depends on  $a$ , as mentioned above.  $Q(t, \lambda) = 2t/\lambda$  is decreasing in  $\lambda$ .) For example, if  $n = 10$ , then consulting Table 3 shows that a 95% confidence interval is given by  $\{\lambda: \frac{2T}{34.17} \leq \lambda \leq \frac{2T}{9.59}\}$ . ||

For the location problem, even if the variance is unknown, construction and calculation of pivotal intervals is quite easy. In fact, we have used these ideas already but have not called them by any formal name.

**Example 9.2.7:** It follows from Theorem 5.4.1 that if  $X_1, \dots, X_n$  are iid  $n(\mu, \sigma^2)$ , then  $(\bar{X} - \mu)/(\sigma/\sqrt{n})$  is a pivot. If  $\sigma^2$  is known, we can use this pivot to calculate a confidence interval for  $\mu$ . For any constant  $a$ ,

$$P \left( -a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq a \right) = P(-a \leq Z \leq a) \quad (Z \text{ is standard normal})$$

and (by now) familiar algebraic manipulations give us the confidence interval

$$\left\{ \mu: \bar{x} - a \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + a \frac{\sigma}{\sqrt{n}} \right\}.$$

If  $\sigma^2$  is unknown we can use the location-scale pivot  $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ . Since  $\frac{\bar{X} - \mu}{S/\sqrt{n}}$  has Student's  $t$  distribution,

$$P \left( -a \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq a \right) = P(-a \leq T_{n-1} \leq a).$$

Thus, for any given  $\alpha$ , if we take  $a = t_{n-1, \alpha/2}$ , we find that a  $1 - \alpha$  confidence interval is given by

$$(9.2.11) \quad \left\{ \mu: \bar{x} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \right\},$$

which is the classic  $1 - \alpha$  confidence interval for  $\mu$  based on Student's  $t$  distribution.

Continuing with this case, suppose that we also want an interval estimate for  $\sigma$ . Because  $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ ,  $(n-1)S^2/\sigma^2$  is also a pivot. Thus, if we choose  $a$  and  $b$  to satisfy

$$P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right) = P(a \leq \chi_{n-1}^2 \leq b) = 1 - \alpha,$$

we can invert this set to obtain the  $1 - \alpha$  confidence interval

$$\left\{ \sigma^2 : \frac{(n-1)s^2}{b} \leq \sigma^2 \leq \frac{(n-1)s^2}{a} \right\},$$

or equivalently,

$$\left\{ \sigma : \sqrt{\frac{(n-1)s^2}{b}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{a}} \right\}.$$

One choice of  $a$  and  $b$  that will produce the required interval is  $a = \chi_{n-1,1-\alpha/2}^2$  and  $b = \chi_{n-1,\alpha/2}^2$ . This choice splits the probability equally, putting  $\alpha/2$  in each tail of the distribution. The  $\chi_{n-1}^2$  distribution, however, is a skewed distribution and it is not immediately clear that an equal probability split is optimal for a skewed distribution. (It is not immediately clear that an equal probability split is optimal for a symmetric distribution, but our intuition makes this latter case more plausible.) In fact, for the chi squared distribution, the equal probability split is not optimal, as will be seen in Section 9.3. (See also Exercise 9.55.)

One final note for this problem. We now have constructed confidence intervals for  $\mu$  and  $\sigma$  separately. It is entirely plausible that we would be interested in a confidence set for  $\mu$  and  $\sigma$  *simultaneously*. The Bonferroni Inequality is an easy (and relatively good) method for accomplishing this. (See Exercise 9.13.) ||

### 9.2.3 Guaranteeing an Interval

We now illustrate a specific method of confidence set construction, one that guarantees that the resulting confidence set will be an interval. (Mood, Graybill, and Boes (1974) call this method “the statistical method.”) The method looks different from the test inversion method, mainly because it takes advantage of some additional properties of the underlying cdf.

If in doubt, or in a strange situation, we would recommend constructing a confidence set based on inverting an LRT, if possible. Such a set, although not guaranteed to be optimal, will never be very bad. However, in some cases such a tactic is too difficult, either analytically or computationally; inversion of the acceptance region can sometimes be quite a chore. If the method of this section can be applied, it is rather straightforward to implement and will usually produce a set that is reasonable.

To illustrate the type of trouble that could arise from the test inversion method, without extra conditions on the exact types of acceptance regions used, consider the following example, which illustrates one of the early methods of constructing confidence sets for a binomial success probability.

**Example 9.2.8:** Sterne (1954) proposed the following method for constructing binomial confidence sets, a method that produces a set with the shortest length. Given  $\alpha$ , for each value of  $p$  find the size  $\alpha$  acceptance region composed of the most probable  $x$  values. That is, for each  $p$ , order the  $x = 0, \dots, n$  values from the most probable to the least probable and put values into the acceptance region  $A(p)$  until it has probability  $1 - \alpha$ . Then use (9.2.1) to invert these acceptance regions to get a  $1 - \alpha$  confidence set, which Sterne claimed had length optimality properties.

To see the unexpected problems with this seemingly reasonable construction, consider a small example. Let  $X \sim \text{binomial}(3, p)$  and use confidence coefficient  $1 - \alpha = .442$ . The following table gives the acceptance regions obtained by the Sterne construction.

$p$	Acceptance region = $A(p)$
[.000, .238]	0
(.238, .305)	0, 1
[.305, .362]	1
(.362, .366)	0, 1
[.366, .634]	1, 2
(.634, .638)	2, 3
[.638, .695]	2
(.695, .762)	2, 3
[.762, 1.00]	3

Now, inverting this family of tests gives the following confidence set.

$x$	Confidence set = $C(x)$
0	[.000, .305] $\cup$ (.362, .366)
1	(.238, .634]
2	[.366, .762)
3	(.634, .638) $\cup$ (.695, 1.00]

Surprisingly, the confidence set is not a confidence *interval*. This seemingly reasonable construction has led us to an unreasonable procedure. The blame is to be put on the pmf, as it does not behave as we expect. (See Exercise 9.17.) ||

We base our confidence interval construction for a parameter  $\theta$  on a real-valued statistic  $T$  with cdf  $F_T(t|\theta)$ . (In practice we would usually take  $T$  to be a sufficient statistic for  $\theta$  but this is not necessary for the following theory to go through.) We will first assume that  $T$  is a continuous random variable. The situation where  $T$  is discrete is similar, but has a few additional technical details to consider. We, therefore, state the discrete case in a separate theorem.

Recall the definitions of *stochastically increasing* and *stochastically decreasing*. (See the *Miscellanea* section of Chapter 8 and Exercise 8.29, or Exercises 3.36–3.38.) A family of cdfs  $F(t|\theta)$  is *stochastically increasing in  $\theta$*  (*stochastically decreasing*

in  $\theta$ ) if, for each  $t \in T$ , the sample space of  $T, F(t|\theta)$  is a decreasing (increasing) function of  $\theta$ . In what follows, we need only the fact that  $F$  is monotone, either increasing or decreasing. The more statistical concepts of stochastic increasing or decreasing merely serve as interpretational tools.

**THEOREM 9.2.2:** Let  $T$  be a statistic with continuous cdf  $F_T(t|\theta)$ . Let  $0 < \alpha < 1$  be a fixed value. Suppose that for each  $t \in T$ , the functions  $\theta_L(t)$  and  $\theta_U(t)$  can be defined as follows.

- a. If  $F_T(t|\theta)$  is a decreasing function of  $\theta$  for each  $t$ , define  $\theta_L(t)$  and  $\theta_U(t)$  by

$$F_T(t|\theta_U(t)) = \frac{\alpha}{2}, \quad F_T(t|\theta_L(t)) = 1 - \frac{\alpha}{2}.$$

- b. If  $F_T(t|\theta)$  is an increasing function of  $\theta$  for each  $t$ , define  $\theta_L(t)$  and  $\theta_U(t)$  by

$$F_T(t|\theta_U(t)) = 1 - \frac{\alpha}{2}, \quad F_T(t|\theta_L(t)) = \frac{\alpha}{2}.$$

Then the random interval  $[\theta_L(T), \theta_U(T)]$  is a  $1 - \alpha$  confidence interval for  $\theta$ .

*Proof:* We will prove only part (a). The proof of part (b) is similar and is left as Exercise 9.19. Since  $F_T(t|\theta)$  is a decreasing function of  $\theta$  for each  $t$  and  $1 - \frac{\alpha}{2} > \frac{\alpha}{2}$ ,  $\theta_L(t) < \theta_U(t)$  and the values  $\theta_L(t)$  and  $\theta_U(t)$  are unique. Also,  $F_T(t|\theta)$  is decreasing in  $\theta$  and hence

$$\begin{aligned} F_T(t|\theta) < \frac{\alpha}{2} &\Leftrightarrow \theta > \theta_U(t), \\ F_T(t|\theta) > 1 - \frac{\alpha}{2} &\Leftrightarrow \theta < \theta_L(t). \end{aligned}$$

Consider the interval  $[\theta_L(T), \theta_U(T)]$ . Since

$$\theta > \theta_U(t) \Leftrightarrow F_T(t|\theta) < \frac{\alpha}{2},$$

it follows that

$$P_\theta(\theta > \theta_U(T)) = P_\theta\left(F_T(T|\theta) < \frac{\alpha}{2}\right) = \frac{\alpha}{2},$$

where the last equality follows from the Probability Integral Transform (Theorem 2.1.4) which states that the random variable  $F_T(T|\theta) \sim \text{uniform}(0, 1)$ . By a similar argument, we have

$$P_\theta(\theta < \theta_L(T)) = P_\theta\left(F_T(T|\theta) > 1 - \frac{\alpha}{2}\right) = \frac{\alpha}{2}.$$

Putting these two together, we have

$$P_\theta(\theta_L(T) \leq \theta \leq \theta_U(T)) = P_\theta(\theta \leq \theta_U(T)) - P_\theta(\theta < \theta_L(T)) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha,$$

proving the theorem. □

The equations for the stochastically increasing case,

$$(9.2.12) \quad F_T(t|\theta_U(t)) = \frac{\alpha}{2}, \quad F_T(t|\theta_L(t)) = 1 - \frac{\alpha}{2},$$

can also be expressed in terms of the pdf of the statistic  $T$ . The functions  $\theta_U(t)$  and  $\theta_L(t)$  can be defined to satisfy

$$\int_{-\infty}^t f_T(u|\theta_U(t)) du = \frac{\alpha}{2} \quad \text{and} \quad \int_t^\infty f_T(u|\theta_L(t)) du = \frac{\alpha}{2}.$$

A similar set of equations holds for the stochastically decreasing case.

**Example 9.2.9:** This method can be used to get a confidence interval for the location exponential pdf. (In Exercise 9.25 the answer here is compared to that obtained by likelihood and pivotal methods. See also Exercise 9.40.)

If  $X_1, \dots, X_n$  are iid with pdf  $f(x|\mu) = e^{-(x-\mu)} I_{[\mu, \infty)}(x)$ , then  $Y = \min\{X_1, \dots, X_n\}$  is sufficient for  $\mu$  with pdf

$$f_Y(y|\mu) = n e^{-n(y-\mu)} I_{[\mu, \infty)}(y).$$

Fix  $\alpha$  and define  $\mu_L(y)$  and  $\mu_U(y)$  to satisfy

$$\int_{\mu_U(y)}^y n e^{-n(u-\mu_U(y))} du = \frac{\alpha}{2}, \quad \int_y^\infty n e^{-n(u-\mu_L(y))} du = \frac{\alpha}{2}.$$

These integrals can be evaluated to give the equations

$$1 - e^{-n(y-\mu_U(y))} = \frac{\alpha}{2}, \quad e^{-n(y-\mu_L(y))} = \frac{\alpha}{2},$$

which give us the solutions

$$\mu_U(y) = y + \frac{1}{n} \log \left( 1 - \frac{\alpha}{2} \right), \quad \mu_L(y) = y + \frac{1}{n} \log \left( \frac{\alpha}{2} \right).$$

Hence, the random interval

$$C(Y) = \left\{ \mu : Y + \frac{1}{n} \log \left( \frac{\alpha}{2} \right) \leq \mu \leq Y + \frac{1}{n} \log \left( 1 - \frac{\alpha}{2} \right) \right\}$$

a  $1 - \alpha$  confidence interval for  $\mu$ . ||

Note two things about the use of this method. First, the actual equations (9.2.12) need to be solved only for the value of the statistics actually observed. If  $T = t_0$  is

observed, then the realized confidence interval on  $\theta$  will be  $[\theta_L(t_0), \theta_U(t_0)]$ . Thus, we need to solve only the two equations

$$\int_{-\infty}^{t_0} f_T(u|\theta_U(t_0)) du = \frac{\alpha}{2} \quad \text{and} \quad \int_{t_0}^{\infty} f_T(u|\theta_L(t_0)) du = \frac{\alpha}{2}$$

for  $\theta_L(t_0)$  and  $\theta_U(t_0)$ . Second, realize that even if these equations cannot be solved analytically, we really only need to solve them numerically since the proof that we have a  $1 - \alpha$  confidence interval did not require an analytic solution.

We now consider the discrete case.

**THEOREM 9.2.3:** Let  $T$  be a discrete statistic with cdf  $F_T(t|\theta) = P(T \leq t|\theta)$ . Let  $0 < \alpha < 1$  be a fixed value. Suppose that for each  $t \in \mathcal{T}$ ,  $\theta_L(t)$  and  $\theta_U(t)$  can be defined as follows.

- a. If  $F_T(t|\theta)$  is a decreasing function of  $\theta$  for each  $t$ , define  $\theta_L(t)$  and  $\theta_U(t)$  by

$$P(T \leq t|\theta_U(t)) = \frac{\alpha}{2}, \quad P(T \geq t|\theta_L(t)) = \frac{\alpha}{2}.$$

- b. If  $F_T(t|\theta)$  is an increasing function of  $\theta$  for each  $t$ , define  $\theta_L(t)$  and  $\theta_U(t)$  by

$$P(T \geq t|\theta_U(t)) = \frac{\alpha}{2}, \quad P(T \leq t|\theta_L(t)) = \frac{\alpha}{2}.$$

Then the random interval  $[\theta_L(T), \theta_U(T)]$  is a  $1 - \alpha$  confidence interval for  $\theta$ .

*Proof:* We will only sketch the proof of part (a). The details, as well as the proof of part (b), are left to Exercise 9.20.

Define the function  $\bar{F}_T(t|\theta)$  by

$$\bar{F}_T(t|\theta) = P(T \geq t|\theta).$$

Since  $F_T(t|\theta)$  is a decreasing function of  $\theta$  for each  $t$ , it can be shown that  $\bar{F}_T(t|\theta)$  is a nondecreasing function of  $\theta$  for each  $t$ . It therefore follows that

$$\theta > \theta_U(t) \Rightarrow F_T(t|\theta) < \frac{\alpha}{2},$$

$$\theta < \theta_L(t) \Rightarrow \bar{F}_T(t|\theta) < \frac{\alpha}{2}.$$

For the interval  $[\theta_L(T), \theta_U(T)]$ ,

$$\begin{aligned} P_\theta(\theta_L(T) \leq \theta \leq \theta_U(T)) &= P_\theta(\theta \leq \theta_U(T)) - P_\theta(\theta < \theta_L(T)) \\ &= 1 - P_\theta\left(F_T(T|\theta) < \frac{\alpha}{2}\right) - P_\theta\left(\bar{F}_T(T|\theta) < \frac{\alpha}{2}\right). \end{aligned}$$

Now, from Exercise 2.11 we know that  $F_T(T|\theta)$  is stochastically greater than a uniform random variable. Furthermore, this fact also implies that  $\bar{F}_T(T|\theta)$  is stochastically greater than a uniform random variable. Thus, we have that

$$P_\theta \left( F_T(T|\theta) < \frac{\alpha}{2} \right) \leq \frac{\alpha}{2} \quad \text{and} \quad P_\theta \left( \bar{F}_T(T|\theta) < \frac{\alpha}{2} \right) \leq \frac{\alpha}{2},$$

showing that  $[\theta_L(T), \theta_U(T)]$  is a  $1 - \alpha$  confidence interval.  $\square$

We close this section with an example to illustrate the construction of Theorem 9.2.3. Notice that an alternative interval can be constructed by inverting an LRT (Exercise 9.23), but the method of pivotal inference fails.

**Example 9.2.10:** Let  $X_1, \dots, X_n$  be a random sample from a Poisson population with parameter  $\lambda$  and define  $Y = \sum X_i$ .  $Y$  is sufficient for  $\lambda$  and  $Y \sim \text{Poisson}(n\lambda)$ . Applying the above method, if  $Y = y_0$  is observed, we are led to solve for  $\lambda$  in the equations

$$(9.2.13) \quad \sum_{k=0}^{y_0} e^{-n\lambda} \frac{(n\lambda)^k}{k!} = \frac{\alpha}{2} \quad \text{and} \quad \sum_{k=y_0}^{\infty} e^{-n\lambda} \frac{(n\lambda)^k}{k!} = \frac{\alpha}{2}.$$

Recall the identity, from Example 3.2.1, linking the Poisson and gamma families. Applying that identity to the sums in (9.2.13), we can write (remembering that  $y_0$  is the *observed* value of  $Y$ )

$$\frac{\alpha}{2} = \sum_{k=0}^{y_0} e^{-n\lambda} \frac{(n\lambda)^k}{k!} = P(Y \leq y_0 | \lambda) = P(\chi^2_{2(y_0+1)} > 2n\lambda),$$

where  $\chi^2_{2(y_0+1)}$  is a chi squared random variable with  $2(y_0 + 1)$  degrees of freedom. Thus, the solution to the above equation is to take

$$\lambda = \frac{1}{2n} \chi^2_{2(y_0+1), \alpha/2}.$$

Similarly, applying the identity to the other equation in (9.2.13) yields

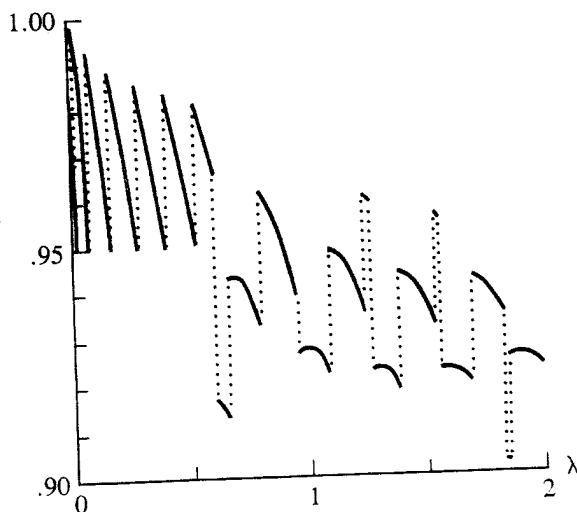
$$\frac{\alpha}{2} = \sum_{k=y_0}^{\infty} e^{-n\lambda} \frac{(n\lambda)^k}{k!} = P(Y \geq y_0 | \lambda) = P(\chi^2_{2y_0} < 2n\lambda).$$

Doing some algebra, we obtain the  $1 - \alpha$  confidence interval for  $\lambda$  as

$$(9.2.14) \quad \left\{ \lambda : \frac{1}{2n} \chi^2_{2y_0, 1-\alpha/2} \leq \lambda \leq \frac{1}{2n} \chi^2_{2(y_0+1), \alpha/2} \right\}.$$

(At  $y_0 = 0$  we define  $\chi^2_{0, 1-\alpha/2} = 0$ .)

These intervals were first derived by Garwood (1936). A graph of the coverage probabilities is given in Figure 9.2.3. Notice that the graph is quite jagged. The jumps occur at the endpoints of the different confidence intervals, where terms are added or subtracted from the sum that makes up the coverage probability. (See Exercise 9.24.)



**FIGURE 9.2.3** Coverage probability for 90% confidence interval from Example 9.2.10 ( $n = 5$ )

For a numerical example, consider  $n = 10$  and observing  $y_0 = \sum x_i = 6$ . A 90% confidence interval for  $\lambda$  is given by

$$\frac{1}{20} \chi^2_{12, .95} \leq \lambda \leq \frac{1}{20} \chi^2_{14, .05},$$

and from Table 3 we get

$$.262 \leq \lambda \leq 1.184.$$

Similar derivations, involving the negative binomial and binomial distributions, are given in the exercises. ||

#### 9.2.4 Bayesian Intervals

Thus far, when describing the interactions between the confidence interval and the parameter, we have carefully said that the interval *covers* the parameter, not that the parameter *is inside* the interval. This was done on purpose. We wanted to stress that the random quantity is the interval, not the parameter. Therefore, we tried to make the action verbs apply to the interval and not the parameter.

In Example 9.2.10 we saw that if  $y_0 = \sum_{i=1}^{10} x_i = 6$ , then a 90% confidence interval for  $\lambda$  is  $.262 \leq \lambda \leq 1.184$ . It is tempting to say (and many experimenters do) that “the probability is 90% that  $\lambda$  is in the interval [.262, 1.184].” Within classical statistics, however, such a statement is invalid since the parameter is assumed fixed. Formally, the interval [.262, 1.184] is one of the possible *realized values* of the random

interval  $[\frac{1}{2n}\chi^2_{2Y, .95}, \frac{1}{2n}\chi^2_{2(Y+1), .05}]$  and, since the parameter  $\lambda$  does not move,  $\lambda$  is in the *realized interval* [.262, 1.184] with probability either 0 or 1. When we say that the realized interval [.262, 1.184] has a 90% chance of coverage, we only mean that we know that 90% of the sample points of the random interval cover the true parameter.

In contrast, the Bayesian set-up allows us to say that  $\lambda$  is *inside* [.262, 1.184] with some probability, not 0 or 1. This is because, under the Bayesian model,  $\lambda$  is a random variable with a probability distribution. All Bayesian claims of coverage are made with respect to the posterior distribution of the parameter.

To keep the distinction between Bayesian and classical sets clear, since the sets make quite different probability assessments, the Bayesian set estimates are referred to as *credible sets* rather than confidence sets.

Thus, if  $\pi(\theta|x)$  is the posterior distribution of  $\theta$  given  $X = x$ , then for any set  $A \subset \Theta$  the credible probability of  $A$  is

$$P(\theta \in A|x) = \int_A \pi(\theta|x) d\theta,$$

and  $A$  is a *credible set* for  $\theta$ . If  $\pi(\theta|x)$  is a pmf, we replace integrals with sums in the above expressions.

Notice that both the interpretation and construction of the Bayes credible set is more straightforward than that of a classical confidence set. However, remember that nothing comes free. The ease of construction and interpretation comes with additional assumptions. The Bayesian model requires more input than the classical model.

**Example 9.2.11:** We now construct a credible set for the problem of Example 9.2.10. Let  $X_1, \dots, X_n$  be iid Poisson( $\lambda$ ) and assume that  $\lambda$  has a gamma prior pdf,  $\lambda \sim \text{gamma}(a, b)$ . The posterior pdf of  $\lambda$  (Exercise 7.25) is

$$\pi(\lambda|\sum X = \sum x) = \text{gamma}(a + \sum x, [n + (1/b)]^{-1}).$$

If we take  $a = b = 1$  we have that the posterior distribution of  $\lambda$  given  $\sum X = \sum x$  can be expressed as  $2(n+1)\lambda \sim \chi^2_{2(\sum x+1)}$ . We can form a credible set for  $\lambda$  in many different ways. One way would be to take upper and lower endpoints defined in a manner similar to that used in Example 9.2.10:

$$L(\sum x) = \chi^2_{2(\sum x+1), 1-\alpha/2} \quad \text{and} \quad U(\sum x) = \chi^2_{2(\sum x+1), \alpha/2},$$

giving the  $1 - \alpha$  Bayes credible set

$$\left\{ \lambda : \frac{\chi^2_{2(\sum x+1), 1-\alpha/2}}{2(n+1)} \leq \lambda \leq \frac{\chi^2_{2(\sum x+1), \alpha/2}}{2(n+1)} \right\}.$$

As in Example 9.2.10, assume  $n = 10$  and  $\sum x = 6$ . Since  $\chi^2_{14, .95} = 6.571$  and  $\chi^2_{14, .05} = 23.685$ , a 90% credible set for  $\lambda$  is given by [.299, 1.077].

We can also form a credible set by taking the *highest posterior density* (HPD) region of the parameter space, which is similar to the construction used in likelihood sets. The  $1 - \alpha$  HPD region is given by  $\{\lambda : \pi(\lambda | \sum x) \geq c\}$ , where  $c$  is chosen so that

$$1 - \alpha = \int_{\{\lambda : \pi(\lambda | \sum x) \geq c\}} \pi(\lambda | \sum x) d\lambda.$$

Such a construction is optimal in the sense of giving the shortest interval for a given credible probability (as will be seen in Theorem 9.3.1). As before, if  $n = 10$  and  $\sum x = 6$ , the 90% HPD credible set for  $\lambda$  is given by [.253, 1.005].

In Figure 9.2.4 we show three  $1 - \alpha$  intervals for  $\lambda$ , the two  $1 - \alpha$  Bayes credible sets derived here and the classical  $1 - \alpha$  confidence set of Example 9.2.10.

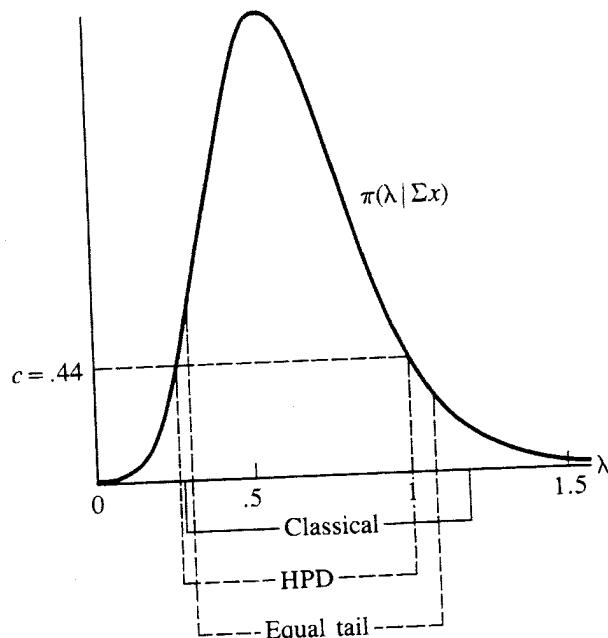


FIGURE 9.2.4 Three interval estimators from Example 9.2.11

According to the above definition, the HPD region for a parameter  $\theta$  is a set of the form  $\{\theta : \pi(\theta | \mathbf{x}) \geq c\}$ . The constant  $c$  is chosen to give the set the desired probability content. In general, the HPD region is not symmetric about a Bayes point estimator but, like the likelihood region, is rather asymmetric. For the Poisson distribution this is clearly true, as the above example shows. Although it will not always happen, we can usually expect asymmetric HPD regions for scale parameter problems and symmetric HPD regions for location parameter problems.

**Example 9.2.12:** Let  $X_1, \dots, X_n$  be iid  $n(\theta, \sigma^2)$  and let  $\theta$  have the prior pdf  $n(\mu, \tau^2)$ , where  $\mu$ ,  $\sigma$ , and  $\tau$  are all known. In Example 7.2.10 we saw that

$$\pi(\theta|\bar{x}) \sim n(\delta^B(\bar{x}), \text{Var}(\theta|\bar{x})),$$

where

$$\delta^B(\bar{x}) = \frac{\sigma^2}{\sigma^2 + n\tau^2}\mu + \frac{n\tau^2}{\sigma^2 + n\tau^2}\bar{x} \quad \text{and} \quad \text{Var}(\theta|\bar{x}) = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}.$$

It therefore follows that under the posterior distribution,

$$\frac{\theta - \delta^B(\bar{x})}{\sqrt{\text{Var}(\theta|\bar{x})}} \sim n(0, 1).$$

A  $1 - \alpha$  credible set for  $\theta$ , which is the  $1 - \alpha$  HPD region, is given by

$$(9.2.15) \quad \{\theta : \pi(\theta|\bar{x}) \geq c\} = \{\theta : |\theta - \delta^B(\bar{x})| \leq z_{\alpha/2} \sqrt{\text{Var}(\theta|\bar{x})}\}.$$

This is symmetric about  $\delta^B(\bar{x})$ . ||

It is important not to confuse credible probability (the Bayes posterior probability) with coverage probability (the classical probability). The probabilities are very different entities, with different meanings and interpretations. Credible probability comes from the posterior distribution, which in turn gets its probability from the prior distribution. Thus, credible probability reflects the experimenter's subjective beliefs, as expressed in the prior distribution and updated with the data to the posterior distribution. A Bayesian assertion of 90% coverage means that the experimenter, upon combining prior knowledge with data, is 90% sure of coverage.

Coverage probability, on the other hand, reflects the uncertainty in the sampling procedure, getting its probability from the objective mechanism of repeated experimental trials. A classical assertion of 90% coverage means that in a long sequence of identical trials, 90% of the realized confidence sets will cover the true parameter.

Statisticians sometimes argue as to which is the better way to do statistics, classical or Bayesian. We do not want to argue or even defend one over another. In fact, we believe that there is no one best way to do statistics; some problems are best solved with classical statistics and some are best solved with Bayesian statistics. The important point to realize is that the solutions may be quite different. A Bayes solution is often not reasonable under classical evaluations and vice versa.

**Example 9.2.12 (Continued):** Consider the classical coverage probability of the HPD region in (9.2.15). Under the classical model  $\bar{X}$  is the random variable and  $\theta$  is fixed. We have, under the classical model, that  $\bar{X} \sim n(\theta, \sigma^2/n)$  and we want to calculate

$$P_\theta \left( |\theta - \delta^B(\bar{X})| \leq z_{\alpha/2} \sqrt{\text{Var}(\theta|\bar{X})} \right).$$

For ease of notation define  $\gamma = \sigma^2/(n\tau^2)$ . Using the definitions of  $\delta^B(\bar{X})$  and  $\text{Var}(\theta|\bar{X})$  and a little algebra, we have

$$\begin{aligned} P_\theta(|\theta - \delta^B(\bar{X})| \leq z_{\alpha/2} \sqrt{\text{Var}(\theta|\bar{X})}) \\ = P_\theta\left(\left|\theta - \left(\frac{\gamma}{1+\gamma}\mu + \frac{1}{1+\gamma}\bar{X}\right)\right| \leq z_{\alpha/2} \sqrt{\frac{\sigma^2}{n(1+\gamma)}}\right) \\ = P_\theta\left(-\sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\theta-\mu)}{\sigma/\sqrt{n}} \leq Z \leq \sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\theta-\mu)}{\sigma/\sqrt{n}}\right), \end{aligned}$$

where the last equality used the fact that  $\sqrt{n}(\bar{X} - \theta)/\sigma = Z \sim N(0, 1)$ .

Although we started with a  $1 - \alpha$  credible set, we do not have a  $1 - \alpha$  confidence set, as can be seen by considering the following parameter configuration. Fix  $\theta \neq \mu$  and let  $\tau = \sigma/\sqrt{n}$ , so that  $\gamma = 1$ . Also, let  $\sigma/\sqrt{n}$  be very small ( $\rightarrow 0$ ). Then it is easy to see that the above probability goes to zero since if  $\theta > \mu$  the lower bound goes to infinity and if  $\theta < \mu$  the upper bound goes to  $-\infty$ . If  $\theta = \mu$ , however, the coverage probability is bounded away from zero.

On the other hand, the usual  $1 - \alpha$  confidence set for  $\theta$  is  $\{\theta : |\theta - \bar{x}| \leq z_{\alpha/2}\sigma/\sqrt{n}\}$ . The credible probability of this set (now  $\theta \sim \pi(\theta|\bar{x})$ ) is given by

$$\begin{aligned} P_{\bar{x}}\left(|\theta - \bar{x}| \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \\ = P_{\bar{x}}\left(|[\theta - \delta^B(\bar{x})] + [\delta^B(\bar{x}) - \bar{x}]| \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \\ = P_{\bar{x}}\left(-\sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\bar{x}-\mu)}{\sqrt{1+\gamma}\sigma/\sqrt{n}} \leq Z \leq \sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\bar{x}-\mu)}{\sqrt{1+\gamma}\sigma/\sqrt{n}}\right), \end{aligned}$$

where the last equality used the fact that  $(\theta - \delta^B(\bar{x}))/\sqrt{\text{Var}(\theta|\bar{x})} = Z \sim N(0, 1)$ . Again, it is fairly easy to show that this probability is not bounded away from zero, showing that the confidence set is also not, in general, a credible set. Details are in Exercise 9.30. ||

### 9.2.5 Invariant Intervals

We have already encountered invariance considerations in a variety of situations: as a data reduction device, to construct point estimators, and to construct hypothesis tests. Not surprisingly, we can also use invariance considerations to help construct confidence sets. Invariant confidence sets are also interesting because consideration of invariant confidence sets leads us to a situation where Bayesian sets and classical frequentist sets meet.

We will only consider cases of location and scale invariance, as a complete treatment of the subject requires more group theory than we want to pursue. A more

complete and advanced treatment of invariant confidence sets is given in Berger (1985) and Lehmann (1986).

We now define an invariant confidence set. The definition is similar to that of an invariant point estimate (see Definition 7.2.3 and the preceding discussion).

**DEFINITION 9.2.2:** Let  $\mathcal{F} = \{f(\mathbf{x}|\theta) : \theta \in \Theta\}$  be invariant under the group  $\mathcal{G}$ . A confidence set for  $\theta$ ,  $C(\mathbf{x})$ , is *invariant under the group  $\mathcal{G}$*  if, for every  $\mathbf{x} \in \mathcal{X}$ ,  $\theta \in \Theta$ , and  $g \in \mathcal{G}$ ,  $\theta \in C(\mathbf{x})$  if and only if  $\bar{g}(\theta) \in C(g(\mathbf{x}))$ .

**Example 9.2.13:** Suppose that  $X_1, \dots, X_n$  is a random sample from a location family. That is, the pdf of  $X_i$  is of the form  $f(x_i - \theta)$  where  $-\infty < \theta < \infty$ . By an argument like that used in Example 6.3.2, this family is invariant under the group  $\mathcal{G} = \{g_a(\mathbf{x}) : -\infty < a < \infty\}$ , where  $g_a(x_1, \dots, x_n) = (x_1 + a, \dots, x_n + a)$ . Consider a confidence interval of the form

$$C(\mathbf{x}) = \{\theta : \bar{x} - k_1 \leq \theta \leq \bar{x} + k_2\},$$

where  $k_1$  and  $k_2$  are constants. For each  $g_a \in \mathcal{G}$  and  $\mathbf{x} \in \mathcal{X}$  we have

$$\begin{aligned} C(g_a(\mathbf{x})) &= \left\{ \theta : \frac{\sum(x_i + a)}{n} - k_1 \leq \theta \leq \frac{\sum(x_i + a)}{n} + k_2 \right\} \\ &= \{\theta : \bar{x} + a - k_1 \leq \theta \leq \bar{x} + a + k_2\}. \end{aligned}$$

It is also clear that

$$\bar{x} - k_1 \leq \theta \leq \bar{x} + k_2 \Leftrightarrow \bar{x} + a - k_1 \leq \theta + a \leq \bar{x} + a + k_2.$$

For this group,  $\bar{g}(\theta) = \theta + a$  and we have

$$\theta \in C(\mathbf{x}) \Leftrightarrow \bar{g}(\theta) \in C(g_a(\mathbf{x})),$$

showing that  $C$  is an invariant confidence interval. ||

One method of constructing invariant confidence sets is by inverting the acceptance regions of invariant tests. However, some explanation is necessary. Recall Definition 8.2.3, where we first encountered invariant tests. That definition requires an invariant test to make the *same* inference whether a sample point  $\mathbf{x}$  or a transformed point  $g(\mathbf{x})$  is observed. But Definition 9.2.2 requires that an invariant confidence set *change* in a specified way when the sample point  $\mathbf{x}$  is transformed to  $g(\mathbf{x})$ . The connection between these seemingly disparate ideas is not as mystifying as it may at first seem.

When discussing the relationship between confidence sets and hypothesis tests, we are discussing only *one* confidence set  $C(\mathbf{x})$ , but an *entire collection* of tests, a different test for each  $H_0: \theta = \theta_0, \theta_0 \in \Theta$ . Now, when discussing invariance, we have a different group of transformations,  $\mathcal{G}_{\theta_0}$ , for each testing problem, with  $H_0: \theta = \theta_0$ .

The invariant test of  $H_0: \theta = \theta_0$  need be invariant only with respect to the group  $\mathcal{G}_{\theta_0}$ , but the invariant confidence set must be invariant with respect to a much larger group  $\mathcal{G}$ . This larger group  $\mathcal{G}$  must be a group containing all the groups  $\mathcal{G}_{\theta_0}, \theta_0 \in \Theta$ . Often,  $\mathcal{G}$  is just the union of all the groups  $\mathcal{G}_{\theta_0}$ . If all the invariant tests have a similar structure, then the confidence set constructed from these tests will be invariant with respect to  $\mathcal{G}$ . These ideas are illustrated in the following example.

**Example 9.2.14 (Continuation of Example 8.2.6):** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population, where both parameters are unknown. We will consider tests and confidence intervals based on the sufficient statistic  $(\bar{X}, S^2)$ . In Example 8.2.6 we saw that for testing  $H_0: \mu = 0$  versus  $H_1: \mu > 0$ , tests that depend only on the statistic  $\bar{X}/S$  are invariant with respect to the group

$$\mathcal{G}_0 = \{g_c(\bar{x}, s^2) : g_c(\bar{x}, s^2) = (c\bar{x}, c^2 s^2), c > 0\}.$$

To construct an invariant set estimator, however, we would want to start with an invariant test of the hypotheses  $H_0: \mu = \mu_0$  versus  $H_1: \mu > \mu_0$ , where  $\mu_0$  is arbitrary. (Notice that this hypothesis testing problem is invariant with respect to  $\mathcal{G}_0$  only if  $\mu_0 = 0$ .) A group that leaves the problem of testing  $H_0: \mu = \mu_0$  versus  $H_1: \mu > \mu_0$  invariant is

$$\mathcal{G}_{\mu_0} = \{g_{c,\mu_0}(\bar{x}, s^2) : g_{c,\mu_0}(\bar{x}, s^2) = (c(\bar{x} - \mu_0) + \mu_0, c^2 s^2), c > 0\}.$$

Tests that depend only on the statistic  $(\bar{X} - \mu_0)/S$  are invariant and verification of these facts is left as Exercise 9.32.

Consider, in particular, a size  $\alpha$  test of  $H_0: \mu = \mu_0$  versus  $H_1: \mu > \mu_0$  that has the acceptance region

$$A(\mu_0) = \left\{ \mathbf{x} : \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \leq t_{n-1, \alpha} \right\}.$$

Then  $\mathbf{x} \in A(\mu_0)$  if and only if  $\mu_0 \geq \bar{x} - t_{n-1, \alpha} s / \sqrt{n}$ . Thus

$$C(\mathbf{x}) = \left\{ \mu : \mu \geq \bar{x} - t_{n-1, \alpha} \frac{s}{\sqrt{n}} \right\}$$

is a  $1 - \alpha$  confidence interval for  $\mu$ . This confidence interval is invariant with respect to the group

$$\mathcal{G} = \{g_{a,b}(\bar{x}, s^2) : g_{a,b}(\bar{x}, s^2) = (a\bar{x} + b, a^2 s^2), a > 0, -\infty < b < \infty\}.$$

Observe that  $\mathcal{G}_{\mu_0} \subset \mathcal{G}$  for every  $\mu_0$  since  $g_{c,\mu_0} = g_{a,b}$  if  $a = c$  and  $b = -c\mu_0 + \mu_0$ . To see that  $C(\mathbf{x})$  is invariant with respect to  $\mathcal{G}$ , note that for a given  $g_{a,b}$  and  $\theta = (\mu, \sigma^2)$ ,  $\bar{g}(\theta) = (a\mu + b, a^2 \sigma^2)$ . So we have

$$\begin{aligned}
 \theta \in C(\mathbf{x}) &\Leftrightarrow \mu \geq \bar{x} - t_{n-1,\alpha} \frac{s}{\sqrt{n}} \\
 &\Leftrightarrow a\mu + b \geq a\left(\bar{x} - t_{n-1,\alpha} \frac{s}{\sqrt{n}}\right) + b \\
 &\Leftrightarrow a\mu + b \geq a\bar{x} + b - t_{n-1,\alpha} a \frac{s}{\sqrt{n}} \\
 &\Leftrightarrow \bar{g}(\theta) \in C(g_{a,b}(\mathbf{x})).
 \end{aligned}$$

Thus  $C(\mathbf{x})$  is invariant with respect to  $\mathcal{G}$ . ||

## 9.3 Methods of Evaluating Interval Estimators

We now have seen many methods for deriving confidence sets and, in fact, we can derive different confidence sets for the same problem. In such situations we would, of course, want to choose a best one. Therefore, we now examine some methods and criteria for evaluating set estimators.

In set estimation two quantities vie against one another, size and coverage probability. Naturally, we want our set to have small size and large coverage probability, but such sets are usually difficult to construct. (Clearly, we can have a large coverage probability by increasing the size of our set. The interval  $(-\infty, \infty)$  has coverage probability one!) Before we can optimize a set with respect to size and coverage probability, we must decide how to measure these quantities.

The coverage probability of a confidence set will, except in special cases, be a function of the parameter so there is not one value to consider, but an infinite number of values. For the most part, however, we will measure coverage probability performance by the *confidence coefficient*, the infimum of the coverage probabilities. This is one way, but not the only available way of summarizing the coverage probability information. (For example, we could calculate an average coverage probability.)

When we speak of the *size* of a confidence set we will usually mean the *length* of the confidence set, if the set is an interval. If the set is not an interval, or if we are dealing with a multidimensional set, then length will usually become *volume*. We will also see some cases where a size measure other than length is considered. Sometimes a different measure of size is more natural and sometimes it is more convenient. We will look into this in some detail in Section 9.3.3. (Also see Berger, 1985, Chapter 6.)

### 9.3.1 Size and Coverage Probability

We now consider what appears to be a simple, constrained minimization problem. For a given, specified coverage probability find the confidence interval with the shortest length. We first consider an example.

**Example 9.3.1:** Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$ , where  $\sigma$  is known. Using the method of Section 9.2.2 and the fact that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is a pivot with a standard normal distribution, any  $a$  and  $b$  that satisfy

$$P(a \leq Z \leq b) = 1 - \alpha$$

will give the  $1 - \alpha$  confidence interval

$$\left\{ \mu : \bar{x} - b \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} - a \frac{\sigma}{\sqrt{n}} \right\}.$$

Which choice of  $a$  and  $b$  is best? More formally, what choice of  $a$  and  $b$  will minimize the length of the confidence interval while maintaining  $1 - \alpha$  coverage? Notice that the length of the confidence intervals is equal to  $(b - a)\sigma/\sqrt{n}$  but, since the factor  $\sigma/\sqrt{n}$  is part of each interval length, it can be ignored and length comparisons can be based on the value of  $b - a$ . Thus, we want to find a pair of numbers  $a$  and  $b$  that satisfy  $P(a \leq Z \leq b) = 1 - \alpha$  and minimize  $b - a$ .

In Example 9.2.1 we took  $a = -z_{\alpha/2}$  and  $b = z_{\alpha/2}$ , but no mention was made of optimality. If we take  $1 - \alpha = .90$ , then any of the following pairs of numbers give 90% intervals:

$a$	$b$	Probability	$b - a$
-1.34	2.33	$P(Z < a) = .09, P(Z > b) = .01$	3.67
-1.44	1.96	$P(Z < a) = .075, P(Z > b) = .025$	3.40
-1.65	1.65	$P(Z < a) = .05, P(Z > b) = .05$	3.30

This numerical study suggests that the choice  $a = -1.65$  and  $b = 1.65$  gives the best interval and, in fact, it does. In this case splitting the probability  $\alpha$  equally is an optimal strategy. ||

The strategy of splitting  $\alpha$  equally, which is optimal in the above case, is not always optimal. What makes the equal  $\alpha$  split optimal in the above case is the fact that the height of the pdf is the same at  $-z_{\alpha/2}$  and  $z_{\alpha/2}$ . We now prove a theorem that will demonstrate this fact, a theorem that is applicable in some generality, needing only the assumption that the pdf is unimodal. Recall the definition of unimodal: A pdf  $f(x)$  is *unimodal* if there exists  $x^*$  such that  $f(x)$  is nondecreasing for  $x \leq x^*$  and  $f(x)$  is nonincreasing for  $x \geq x^*$ . (This is a rather weak requirement.)

**THEOREM 9.3.1:** Let  $f(x)$  be a unimodal pdf. If the interval  $[a, b]$  satisfies

- a.  $\int_a^b f(x) dx = 1 - \alpha$ ,
- b.  $f(a) = f(b) > 0$ , and
- c.  $a \leq x^* \leq b$ , where  $x^*$  is a mode of  $f(x)$ ,

then  $[a, b]$  is the shortest among all intervals that satisfy (a).

*Proof:* Let  $[a', b']$  be any interval with  $b' - a' < b - a$ . We will show that this implies  $\int_{a'}^{b'} f(x) dx < 1 - \alpha$ . The result will be proved only for  $a' \leq a$ , the proof being similar if  $a < a'$ . Also, two cases need to be considered,  $b' \leq a$  and  $b' > a$ .

If  $b' \leq a$ , then  $a' \leq b' \leq a \leq x^*$  and

$$\begin{aligned} \int_{a'}^{b'} f(x) dx &\leq f(b')(b' - a') && (x \leq b' \leq x^* \Rightarrow f(x) \leq f(b')) \\ &\leq f(a)(b' - a') && (b' \leq a \leq x^* \Rightarrow f(b') \leq f(a)) \\ &< f(a)(b - a) && (b' - a' < b - a \text{ and } f(a) > 0) \\ &\leq \int_a^b f(x) dx && ((b), (c), \text{ and unimodality} \\ &&& \Rightarrow f(x) \geq f(a) \text{ for } a \leq x \leq b) \\ &= 1 - \alpha, && ((a)) \end{aligned}$$

completing the proof in the first case.

If  $b' > a$ , then  $a' \leq a < b' < b$  for, if  $b'$  were greater than or equal to  $b$ , then  $b' - a'$  would be greater than or equal to  $b - a$ . In this case, we can write

$$\begin{aligned} \int_{a'}^{b'} f(x) dx &= \int_a^b f(x) dx + \left[ \int_{a'}^a f(x) dx - \int_{b'}^b f(x) dx \right] \\ &= (1 - \alpha) + \left[ \int_{a'}^a f(x) dx - \int_{b'}^b f(x) dx \right] \end{aligned}$$

and the theorem will be proved if we show that the expression in square brackets is negative. Now, using the unimodality of  $f$ , the ordering  $a' \leq a < b' < b$ , and (b), we have

$$\int_{a'}^a f(x) dx \leq f(a)(a - a')$$

and

$$\int_{b'}^b f(x) dx \geq f(b)(b - b').$$

Thus,

$$\begin{aligned} \int_{a'}^a f(x) dx - \int_{b'}^b f(x) dx &\leq f(a)(a - a') - f(b)(b - b') \\ &= f(a) [(a - a') - (b - b')] && (f(a) = f(b)) \\ &= f(a) [(b' - a') - (b - a)], \end{aligned}$$

which is negative if  $(b' - a') < (b - a)$  and  $f(a) > 0$ . □

Recall how we defined HPD regions (Section 9.2.4). Also recall how likelihood regions come about (see Example 9.2.2). Theorem 9.3.1 shows how both HPD regions and likelihood regions can be optimal. Also, we can see now that the equal  $\alpha$  split, which is optimal in Example 9.3.1, will be optimal for any symmetric unimodal pdf (Exercise 9.39). Theorem 9.3.1 may even apply when the optimality criterion is somewhat different from minimum length.

**Example 9.3.2:** For normal intervals based on the pivot  $\frac{\bar{X}-\mu}{S/\sqrt{n}}$  we know that the shortest length  $1 - \alpha$  confidence interval of the form

$$\bar{x} - b \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} - a \frac{s}{\sqrt{n}}$$

has  $a = -t_{n-1,\alpha/2}$  and  $b = t_{n-1,\alpha/2}$ . The interval length is a function of  $s$ , with general form

$$\text{Length}(s) = (b - a) \frac{s}{\sqrt{n}}$$

It is easy to see that if we had considered the criterion of *expected length* and wanted to find a  $1 - \alpha$  interval to minimize

$$E_\sigma(\text{Length}(S)) = (b - a) \frac{E_\sigma S}{\sqrt{n}} = (b - a)c(n) \frac{\sigma}{\sqrt{n}},$$

then Theorem 9.3.1 applies and the choice  $a = -t_{n-1,\alpha/2}$  and  $b = t_{n-1,\alpha/2}$  again gives the optimal interval. (The quantity  $c(n)$  is a constant dependent only on  $n$ . See Exercise 7.46.) ||

In some cases, especially when working outside of the location problem, we must be careful in the application of Theorem 9.3.1. In scale cases in particular, the theorem may not be directly applicable, but a variant may be.

**Example 9.3.3:** Suppose  $X \sim \text{gamma}(k, \beta)$ . The quantity  $Y = X/\beta$  is a pivot, with  $Y \sim \text{gamma}(k, 1)$ , so we can get a confidence interval by finding constants  $a$  and  $b$  to satisfy

$$(9.3.1) \quad P(a \leq Y \leq b) = 1 - \alpha.$$

However, blind application of Theorem 9.3.1 will not give the shortest confidence interval. That is, choosing  $a$  and  $b$  to satisfy (9.3.1) and also  $f_Y(a) = f_Y(b)$  is not optimal. This is because, based on (9.3.1), the interval on  $\beta$  is of the form

$$\left\{ \beta : \frac{x}{b} \leq \beta \leq \frac{x}{a} \right\},$$

so the length of the interval is  $(\frac{1}{a} - \frac{1}{b})x$ , that is, it is proportional to  $(1/a) - (1/b)$  and not to  $b - a$ .

Although Theorem 9.3.1 is not directly applicable here, the proof of the theorem can be modified to solve this problem. Condition (a) in Theorem 9.3.1 defines  $b$  as a function of  $a$ , say  $b(a)$ . We must solve the following constrained minimization problem:

$$\begin{aligned} \text{Minimize, with respect to } a: \quad & \frac{1}{a} - \frac{1}{b(a)} \\ \text{subject to:} \quad & \int_a^{b(a)} f_Y(y) dy = 1 - \alpha. \end{aligned}$$

The solution (Exercise 9.42) is to choose  $a$  and  $b$  to satisfy both the integral constraint and also  $f_Y(a)a^2 = f_Y(b)b^2$ . Equations like these arise in interval estimation of the variance of a normal distribution; see Example 9.2.7 and Exercise 9.55. Also, note that the above equations define not the shortest *overall* interval, but the shortest *pivotal* interval, that is, the shortest interval based on the pivot  $X/\beta$ . ||

### 9.3.2 Test-Related Optimality

Since there is a one-to-one correspondence between confidence sets and tests of hypotheses (Theorem 9.2.1), there is some correspondence between optimality of tests and optimality of confidence sets. Usually, test-related optimality properties of confidence sets do not directly relate to the size of the set but rather to the probability of the set covering false values.

The probability of covering false values, or the *probability of false coverage*, indirectly measures the size of a confidence set. Intuitively, smaller sets cover fewer values and, hence, are less likely to cover false values. Moreover, we will later see an equation that links size and probability of false coverage.

We first consider the general situation, where  $\mathbf{X} \sim f(\mathbf{x}|\theta)$ , and we construct a  $1 - \alpha$  confidence set for  $\theta$ ,  $C(\mathbf{x})$ , by inverting an acceptance region,  $A(\theta)$ . The probability of coverage of  $C(\mathbf{x})$ , that is, the probability of *true coverage*, is the function of  $\theta$  given by  $P_\theta(\theta \in C(\mathbf{X}))$ . The probability of *false coverage* is the function of  $\theta$  and  $\theta'$  defined by

$$(9.3.2) \quad \begin{aligned} P_\theta(\theta' \in C(\mathbf{X})), \theta \neq \theta', & \text{ if } C(\mathbf{X}) = [L(\mathbf{X}), U(\mathbf{X})], \\ P_\theta(\theta' \in C(\mathbf{X})), \theta' < \theta, & \text{ if } C(\mathbf{X}) = [L(\mathbf{X}), \infty), \\ P_\theta(\theta' \in C(\mathbf{X})), \theta' > \theta, & \text{ if } C(\mathbf{X}) = (-\infty, U(\mathbf{X})], \end{aligned}$$

the probability of covering  $\theta'$  when  $\theta$  is the true parameter.

It makes sense to define the probability of false coverage differently for one-sided and two-sided intervals. For example, if we have a lower confidence bound, we are asserting that  $\theta$  is greater than a certain value and false coverage would occur only if we cover values of  $\theta$  that are too small. A similar argument leads us to the definitions used for upper confidence bounds and two-sided bounds.

A  $1 - \alpha$  confidence set that minimizes the probability of false coverage over a class of  $1 - \alpha$  confidence sets is called a *uniformly most accurate* (UMA) confidence

set. Thus, for example, we would consider looking for a UMA confidence set among sets of the form  $[L(\mathbf{x}), \infty)$ . UMA confidence sets are constructed by inverting the acceptance regions of UMP tests, as we will prove below. Unfortunately, although a UMA confidence set is a desirable set, it exists only in rather rare circumstances (as do UMP tests). In particular, since UMP tests are generally one-sided, so are UMA intervals. They make for elegant theory, however. In the next theorem we see that a UMP test of  $H_0: \theta = \theta_0$  versus  $H_1: \theta > \theta_0$  yields a UMA lower confidence bound.

**THEOREM 9.3.2:** Let  $\mathbf{X} \sim f(\mathbf{x}|\theta)$  where  $\theta$  is a real-valued parameter. For each  $\theta_0 \in \Theta$ , let  $A^*(\theta_0)$  be the UMP level  $\alpha$  acceptance region of a test of  $H_0: \theta = \theta_0$  versus  $H_1: \theta > \theta_0$ . Let  $C^*(\mathbf{x})$  be the  $1 - \alpha$  confidence set formed by inverting the UMP acceptance regions. Then for any other  $1 - \alpha$  confidence set  $C$ ,

$$P_\theta(\theta' \in C^*(\mathbf{X})) \leq P_\theta(\theta' \in C(\mathbf{X})), \text{ for all } \theta' < \theta.$$

*Proof:* Let  $\theta'$  be any value less than  $\theta$ . Let  $A(\theta')$  be the acceptance region of the level  $\alpha$  test of  $H_0: \theta = \theta'$  obtained by inverting  $C$ . Since  $A^*(\theta')$  is the UMP acceptance region for testing  $H_0: \theta = \theta'$  versus  $H_1: \theta > \theta'$  and since  $\theta > \theta'$ , we have

$$\begin{aligned} P_\theta(\theta' \in C^*(\mathbf{X})) &= P_\theta(\mathbf{X} \in A^*(\theta')) \quad (\text{invert the confidence set}) \\ &\leq P_\theta(\mathbf{X} \in A(\theta')) \quad \left( \begin{array}{l} \text{true for any } A \\ \text{since } A^* \text{ is UMP} \end{array} \right) \\ &= P_\theta(\theta' \in C(\mathbf{X})). \quad \left( \begin{array}{l} \text{invert } A \text{ to} \\ \text{obtain } C \end{array} \right) \end{aligned}$$

Notice that the above inequality is " $\leq$ " because we are working with probabilities of acceptance regions. This is  $1 - \text{power}$ , so UMP tests will minimize these acceptance region probabilities. Therefore, we have established that for  $\theta' < \theta$ , the probability of false coverage is minimized by the interval obtained from inverting the UMP test. (A similar theorem can be proved for sets based on a test of  $H_0: \theta = \theta_0$  versus  $H_1: \theta < \theta_0$ ; see Exercise 9.43.)  $\square$

Recall our discussion in Section 9.2.1. The UMA confidence set in the above theorem is constructed by inverting the family of tests for the hypotheses

$$H_0: \theta = \theta_0 \quad \text{versus} \quad H_1: \theta > \theta_0,$$

where the form of the confidence set is governed by the alternative hypothesis. The above alternative hypotheses, which specify that  $\theta_0$  is less than a particular value, lead to *lower* confidence bounds, that is, if the sets are intervals they are of the form  $[L(\mathbf{X}), \infty)$ .

**Example 9.3.4:** Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$ , where  $\sigma^2$  is known. The interval

$$C(\bar{x}) = \left\{ \mu : \mu \geq \bar{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}} \right\}$$

is a  $1 - \alpha$  UMA lower confidence bound since it can be obtained by inverting the UMP test of  $H_0: \mu = \mu_0$  versus  $H_1: \mu > \mu_0$ .

The more common two-sided interval,

$$C(\bar{x}) = \left\{ \mu : \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\},$$

is not UMA, since it is obtained by inverting the two-sided acceptance region from the test of  $H_0: \mu = \mu_0$  versus  $H_1: \mu \neq \mu_0$ , hypotheses for which no UMP test exists. ||

In the testing problem, when considering two-sided tests, we found the property of *unbiasedness* to be both compelling and useful. In the confidence interval problem similar ideas apply. When dealing with two-sided confidence intervals it is reasonable to restrict consideration to unbiased confidence sets. Remember that an unbiased test is one in which the power in the alternative is always greater than the power in the null. Keep that in mind when reading the following definition.

**DEFINITION 9.3.1:** A  $1 - \alpha$  confidence set  $C(x)$  is *unbiased* if  $P_\theta(\theta' \in C(X)) \leq 1 - \alpha$  for all  $\theta \neq \theta'$ .

Thus, for an unbiased confidence set, the probability of false coverage is never more than the minimum probability of true coverage. Within the class of unbiased  $1 - \alpha$  confidence sets we can now look for a uniformly most accurate confidence set.

**DEFINITION 9.3.2:** A  $1 - \alpha$  confidence set  $C^*(x)$  is *uniformly most accurate unbiased* (UMA unbiased) if

$$\text{i. } P_\theta(\theta' \in C^*(X)) \leq 1 - \alpha \text{ for all } \theta \neq \theta' \quad (\text{unbiased})$$

and

$$\text{ii. } P_\theta(\theta' \in C^*(X)) \leq P_\theta(\theta' \in C(X)) \text{ for all } \theta \neq \theta', \quad (\text{most accurate})$$

where  $C(x)$  is any unbiased  $1 - \alpha$  confidence set.

The correspondence between UMP tests and UMA sets carries over to UMP unbiased tests and UMA unbiased sets. A theorem similar to Theorem 9.3.2 can be proved for unbiased sets.

**THEOREM 9.3.3:** Let  $X \sim f(x|\theta)$ . For each  $\theta_0 \in \Theta$ , let  $A^*(\theta_0)$  be the UMP unbiased level  $\alpha$  acceptance region of a test of  $H_0: \theta = \theta_0$  versus  $H_1: \theta \neq \theta_0$ . Let  $C^*(x)$  be the  $1 - \alpha$  confidence set formed by inverting the UMP unbiased acceptance

regions. Then  $C^*(\mathbf{x})$  is a uniformly most accurate unbiased  $1 - \alpha$  confidence set, that is,  $C^*(\mathbf{x})$  is unbiased and, for any other unbiased  $1 - \alpha$  confidence set  $C(\mathbf{x})$ ,

$$P_\theta(\theta' \in C^*(\mathbf{X})) \leq P_\theta(\theta' \in C(\mathbf{X})), \quad \text{for all } \theta' \neq \theta.$$

*Proof:* Exercise 9.47. □

**Example 9.3.4 (Continued):** The two-sided normal interval

$$C(\bar{x}) = \left\{ \mu : \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\},$$

is a UMA unbiased interval. It can be obtained by inverting the uniformly most powerful unbiased test of  $H_0: \mu = \mu_0$  versus  $H_1: \mu \neq \mu_0$ , given in Example 8.3.9. Similarly, the interval based on the  $t$  distribution,

$$C(\bar{x}, s) = \left\{ \mu : \bar{x} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \right\},$$

is also a UMA unbiased interval, since it also can be obtained by inverting a UMP unbiased test (Exercise 8.48). ||

Sets that minimize probability of false coverage are also called *Neyman-shortest*. The fact that there is a length connotation to this name is somewhat justified by the following theorem, due to Pratt (1961).

**THEOREM 9.3.4 (Pratt):** Let  $X$  be a real-valued random variable with  $X \sim f(x|\theta)$  where  $\theta$  is a real-valued parameter. Let  $C(x) = [L(x), U(x)]$  be a confidence interval for  $\theta$ . If  $L(x)$  and  $U(x)$  are both increasing functions of  $x$ , then for any value  $\theta^*$ ,

$$(9.3.3) \quad E_{\theta^*}(\text{Length}[C(X)]) = \int_{\theta \neq \theta^*} P_{\theta^*}(\theta \in C(X)) d\theta.$$

Theorem 9.3.4 says that the expected length of  $C(x)$  is equal to a sum (integral) of the probabilities of false coverage, the integral being taken over all false values of the parameter.

*Proof:* From the definition of expected value we can write

$$\begin{aligned} E_{\theta^*}(\text{Length}[C(X)]) &= \int_{\mathcal{X}} \text{Length}[C(x)] f(x|\theta^*) dx \\ &= \int_{\mathcal{X}} [U(x) - L(x)] f(x|\theta^*) dx && \text{(definition of length)} \\ &= \int_{\mathcal{X}} \left[ \int_{L(x)}^{U(x)} d\theta \right] f(x|\theta^*) dx && \left( \begin{array}{l} \text{using } \theta \text{ as a} \\ \text{dummy variable} \end{array} \right) \end{aligned}$$

$$\begin{aligned}
&= \int_{\Theta} \left[ \int_{U^{-1}(\theta)}^{L^{-1}(\theta)} f(x|\theta^*) dx \right] d\theta \quad \left( \begin{array}{l} \text{invert the order of} \\ \text{integration—see below} \end{array} \right) \\
&= \int_{\Theta} [P_{\theta^*}(U^{-1}(\theta) \leq X \leq L^{-1}(\theta))] d\theta \quad (\text{definition}) \\
&= \int_{\Theta} [P_{\theta^*}(\theta \in C(X))] d\theta \quad \left( \begin{array}{l} \text{invert the} \\ \text{acceptance region} \end{array} \right) \\
&= \int_{\theta \neq \theta^*} [P_{\theta^*}(\theta \in C(X))] d\theta. \quad \left( \begin{array}{l} \text{one point does} \\ \text{not change value} \end{array} \right)
\end{aligned}$$

The string of equalities establishes the identity and proves the theorem. The interchange of integrals is formally justified by Fubini's Theorem (Rudin, 1976), but is easily seen to be justified as long as all of the integrands are finite. The inversion of the confidence interval is standard, where we use the relationship

$$\theta \in \{\theta : L(x) \leq \theta \leq U(x)\} \Leftrightarrow x \in \{x : U^{-1}(\theta) \leq x \leq L^{-1}(\theta)\},$$

which is valid because of the assumption that  $L$  and  $U$  are increasing. Note that the theorem could be modified to apply to an interval with decreasing endpoints.  $\square$

Theorem 9.3.4 shows that there is a formal relationship between the length of a confidence interval and its probability of false coverage. In the two-sided case, this implies that minimizing the probability of false coverage carries along some guarantee of length optimality. In the one-sided case, however, the analogy does not quite work. In that case, intervals that are set up to minimize probability of false coverage are concerned with parameters in only a portion of the parameter space and length optimality may not obtain. Madansky (1962) has given an example of a  $1 - \alpha$  UMA interval (one-sided) that can be beaten in the sense that another, shorter,  $1 - \alpha$  interval can be constructed. (See Exercise 9.45.) Also, Maatta and Casella (1987) have shown that an interval obtained by inverting a UMP test can be suboptimal when measured against other reasonable criteria.

### 9.3.3 Invariant Optimality

As mentioned before, invariant confidence sets can lead us to a situation where Bayesian sets and classical frequentist sets meet, and sets that are classically best in a class of invariant sets also turn out to be Bayes HPD sets based on improper priors. This is one path to optimality properties of invariant sets.

A feature of invariant confidence sets that is quite nice is that, in many common cases, the coverage probability of the set is constant. Thus, it is easy to make size comparisons. We can order the sets according to their (constant) coverage probability and then calculate sizes.

Suppose that  $\theta$  and  $\theta'$  are two parameter values such that there exists a  $g \in \mathcal{G}$  with the property that  $g(\mathbf{X}) = \mathbf{Y} \sim f(\mathbf{y}|\theta')$  if  $\mathbf{X} \sim f(\mathbf{x}|\theta)$ , that is,  $\theta' = \bar{g}(\theta)$ . Then

the coverage probability of an invariant confidence set is the same at  $\theta$  and  $\theta'$ . To see this write

$$\begin{aligned} P_\theta(\theta \in C(\mathbf{X})) &= P_\theta(\theta' \in C(g(\mathbf{X}))) && \text{(Definition 9.2.2)} \\ &= P_{\theta'}(\theta' \in C(\mathbf{Y})). && (\mathbf{Y} \sim f(\mathbf{y}|\theta')) \end{aligned}$$

In many common situations, in particular with location and scale invariance, for any pair  $\theta$  and  $\theta'$ , a  $g \in \mathcal{G}$  with the required property can be found. In this situation the coverage probability is constant throughout the entire parameter space. The formal property that is needed is a property of the group, called *transitivity*. More details can be found in Berger (1985).

When measuring the size of an invariant confidence set, it makes sense to do so in an invariant way. Thus, if  $\text{Size}(\cdot)$  is a measure of size of a confidence set and  $C$  is a confidence set invariant with respect to  $\mathcal{G}$ , then we would want

$$\text{Size}(\{\theta: \theta \in C(\mathbf{x})\}) = \text{Size}(\{\bar{g}(\theta): \bar{g}(\theta) \in C(g(\mathbf{x}))\}),$$

for all  $\mathbf{x} \in \mathcal{X}$  and  $g \in \mathcal{G}$ . That is, the group operation does not change the size of the set.

**Example 9.3.5 (Continuation of Example 9.2.13):** In location problems, length is an invariant measure of size. The confidence interval

$$C(\mathbf{x}) = \{\theta: \bar{x} - k_1 \leq \theta \leq \bar{x} + k_2\},$$

where  $k_1$  and  $k_2$  are constants, has constant coverage probability (Exercise 9.53). Furthermore, length is an invariant measure of size because

$$\begin{aligned} \text{Length}(C(\mathbf{x})) &= \text{Length}(\{\theta: \bar{x} - k_1 \leq \theta \leq \bar{x} + k_2\}) \\ &= k_1 + k_2 \\ &= \text{Length}(\{\theta: \bar{x} + a - k_1 \leq \theta \leq \bar{x} + a + k_2\}) \\ &= \text{Length}(\{\bar{g}(\theta): \bar{g}(\theta) \in C(g_a(\mathbf{x}))\}), && (\bar{g}(\theta) = \theta + a) \\ &= \text{Length}(C(g_a(\mathbf{x}))). \end{aligned}$$

We know from Theorem 9.3.1 that choosing  $k_1 = k_2$  minimizes the length if the pdf of  $\bar{X}$  is symmetric and unimodal. ||

The location problem acted as our intuition suggested and length was a reasonable measure of size. The scale problem is different, however, and length is no longer an invariant measure of size.

**Example 9.3.6:** Let  $X$  have pdf of the form  $\frac{1}{\tau} f(\frac{x}{\tau})$  and consider the scale group

$$\mathcal{G} = \{g_c(x): g_c(x) = cx, c > 0\}.$$

An invariant confidence interval is of the form

$$C(x) = \left\{ \tau : \frac{x}{k_2} \leq \tau \leq \frac{x}{k_1} \right\},$$

and has constant coverage probability. Measuring the size of the interval by its length, however, does not give an invariant measure of size since

$$\begin{aligned} \text{Length}(C(x)) &= \text{Length} \left( \left\{ \tau : \frac{x}{k_2} \leq \tau \leq \frac{x}{k_1} \right\} \right) \\ &= x \left( \frac{1}{k_1} - \frac{1}{k_2} \right) \\ &\neq cx \left( \frac{1}{k_1} - \frac{1}{k_2} \right) \\ &= \text{Length} \left( \left\{ \tau : \frac{cx}{k_2} \leq \tau \leq \frac{cx}{k_1} \right\} \right) \\ &= \text{Length}(C(g_c(x))). \end{aligned}$$

For the scale problem an invariant measure of size is given by the ratio of the interval endpoints. That is, if we define

$$\begin{aligned} \text{Size}(C(x)) &= \text{Size} \left( \left\{ \tau : \frac{x}{k_2} \leq \tau \leq \frac{x}{k_1} \right\} \right) \\ &= \frac{x/k_1}{x/k_2} \\ &= \frac{k_2}{k_1}, \end{aligned}$$

this measure of size is invariant. (Exercise 9.54.) ||

We close this section with a look at the correspondence between invariant intervals and Bayes intervals. In some situations, the best invariant interval corresponds to a Bayesian HPD region against an invariant prior. Invariant priors often turn out to be *improper* (they do not have a finite integral) as the next example shows.

**Example 9.3.7:** Suppose that  $X \sim f(x - \theta)$  and recall that the location family  $\mathcal{F} = \{f(x - \theta) : -\infty < \theta < \infty\}$  is invariant under the group  $\mathcal{G} = \{g_a(x) : -\infty < a < \infty\}$ , where  $g_a(x) = x + a$ . In order to find a prior distribution  $\pi(\theta)$  that is also invariant, we could argue in the following way. (Such arguments are often used to arrive at so-called *noninformative priors*, priors that seemingly impart no prior preference and, hence, can be used to make objective Bayes inferences.)

We would want the prior to be invariant under the transformation  $(x, \theta) \rightarrow (x + a, \theta + a) = (y, \eta)$ . That is, since these formulations are equivalent, we require

the prior probabilities to be equal. Recall the invariance arguments of Section 6.3 and apply them to the prior cdfs  $F_\Theta(\theta)$  and  $F_\eta(\eta)$ . We have for  $\eta_i = \theta_i + a$ ,  $i = 1, 2$ ,

*Measurement Invariance:*

$$F_\Theta(\theta_1) - F_\Theta(\theta_2) = F_\eta(\eta_1) - F_\eta(\eta_2), \quad \left( \begin{array}{l} \theta_i \text{s and } \eta_i \text{s are same,} \\ \text{just different scales} \end{array} \right)$$

*Formal Invariance:*

$$F_\Theta(\eta_1) - F_\Theta(\eta_2) = F_\eta(\eta_1) - F_\eta(\eta_2). \quad \left( \begin{array}{l} \text{problems are statis-} \\ \text{tically equivalent} \end{array} \right)$$

Combining these two equations and using the fact that  $\eta_i = \theta_i + a$ , we arrive at the identity

$$F_\Theta(\theta_1) - F_\Theta(\theta_2) = F_\Theta(\theta_1 + a) - F_\Theta(\theta_2 + a)$$

and taking  $a = -\theta_1$  gives  $F_\Theta(\theta_1) - F_\Theta(\theta_2) = F_\Theta(0) - F_\Theta(\theta_2 - \theta_1)$ . This equation can be satisfied by taking  $F_\Theta(\theta) = \theta$ , resulting in the prior “pdf”  $\pi(\theta) = 1$ . Of course  $\pi(\theta)$  is not really a pdf; it has an infinite integral. But using this *improper prior*, we obtain the posterior

$$\begin{aligned} \pi(\theta|x) &= \frac{f(x-\theta)\pi(\theta)}{\int f(x-\theta)\pi(\theta) d\theta} \\ &= \frac{f(x-\theta)}{\int f(x-\theta) d\theta} \\ &= f(x-\theta). \end{aligned}$$

Thus the Bayes HPD region is of the form  $C(x) = \{\theta: f(x-\theta) \geq k\}$ , which is also the best invariant region and the likelihood region. (See Exercises 9.57 and 9.58.)

For a particular application, if  $X_1, \dots, X_n$  are iid  $n(\mu, \sigma^2)$ ,  $\sigma^2$  known, then the interval

$$C(x) = \left\{ \mu: \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\},$$

is both best invariant and Bayes HPD against  $\pi(\mu) = 1$ . ||

## 9.4 Other Considerations

As we have done in the previous two chapters, we now explore some approximate and asymptotic versions of confidence sets. Our purpose is as before, to illustrate some methods that will be of use in more complicated situations, methods that will get *some* answer. The answers obtained here are almost certainly not the best but are certainly not the worst. In many cases, however, they are the best that we can do.

We start, as previously, with approximations based on MLEs.

### 9.4.1 Approximate Maximum Likelihood Intervals

In Section 7.4 we saw that we could always get an approximate variance for an MLE and in Section 8.4 we saw that we could, in general, get the asymptotic null distribution of the LRT statistic. What we would like now is some overall asymptotic distribution on which we could base a confidence interval.

If  $X_1, \dots, X_n$  are iid  $f(x|\theta)$  and  $\hat{\theta}$  is the MLE of  $\theta$ , then from (7.4.1) the variance of a function  $h(\hat{\theta})$  can be approximated by

$$\widehat{\text{Var}}(h(\hat{\theta})|\theta) \approx \frac{[h'(\theta)]^2|_{\theta=\hat{\theta}}}{-\frac{\partial^2}{\partial \theta^2} \log L(\theta|x)|_{\theta=\hat{\theta}}}.$$

Now, for a fixed but arbitrary value of  $\theta$ , we are interested in the asymptotic distribution of

$$\frac{h(\hat{\theta}) - h(\theta)}{\sqrt{\widehat{\text{Var}}(h(\hat{\theta})|\theta)}}.$$

As in Section 8.4, under suitable and quite general “regularity conditions,” we have

$$\frac{h(\hat{\theta}) - h(\theta)}{\sqrt{\widehat{\text{Var}}(h(\hat{\theta})|\theta)}} \rightarrow N(0, 1),$$

giving the approximate confidence interval

$$h(\hat{\theta}) - z_{\alpha/2} \sqrt{\widehat{\text{Var}}(h(\hat{\theta})|\theta)} \leq h(\theta) \leq h(\hat{\theta}) + z_{\alpha/2} \sqrt{\widehat{\text{Var}}(h(\hat{\theta})|\theta)}.$$

Details on this result are rather complicated and can be found in Cramér (1946) or Huber (1967). The approximation is also discussed by Kendall and Stuart (1979).

**Example 9.4.1 (Continuation of Example 7.4.1):** We have a random sample  $X_1, \dots, X_n$  from a Bernoulli( $p$ ) population. We saw that we could estimate the odds ratio, which is given by  $p/(1-p)$ , by its MLE  $\hat{p}/(1-\hat{p})$  and that this estimate has approximate variance

$$\widehat{\text{Var}}\left(\frac{\hat{p}}{1-\hat{p}}\right) \approx \frac{\hat{p}}{n(1-\hat{p})^3}.$$

We therefore can construct the approximate confidence interval

$$\frac{\hat{p}}{1-\hat{p}} - z_{\alpha/2} \sqrt{\widehat{\text{Var}}\left(\frac{\hat{p}}{1-\hat{p}}\right)} \leq \frac{p}{1-p} \leq \frac{\hat{p}}{1-\hat{p}} + z_{\alpha/2} \sqrt{\widehat{\text{Var}}\left(\frac{\hat{p}}{1-\hat{p}}\right)}. \quad \|$$

A more restrictive form of the likelihood approximation, but one that, when applicable, gives better intervals, is the following. The random quantity

$$(9.4.1) \quad Q(\mathbf{X}|\theta) = \frac{\frac{\partial}{\partial\theta} \log L(\theta|\mathbf{X})}{\sqrt{-E_\theta\left(\frac{\partial^2}{\partial\theta^2} \log L(\theta|\mathbf{X})\right)}}$$

can also be shown to have a  $n(0, 1)$  distribution asymptotically as  $n \rightarrow \infty$ . Thus, the set

$$(9.4.2) \quad \{\theta : |Q(\mathbf{x}|\theta)| \leq z_{\alpha/2}\}$$

is an approximate  $1 - \alpha$  confidence set. Notice that, applying results from Section 7.3.2, we have

$$E_\theta(Q(\mathbf{X}|\theta)) = \frac{E_\theta\left(\frac{\partial}{\partial\theta} \log L(\theta|\mathbf{X})\right)}{\sqrt{-E_\theta\left(\frac{\partial^2}{\partial\theta^2} \log L(\theta|\mathbf{X})\right)}} = 0$$

and

$$\text{Var}_\theta(Q(\mathbf{X}|\theta)) = \frac{\text{Var}_\theta\left(\frac{\partial}{\partial\theta} \log L(\theta|\mathbf{X})\right)}{-E_\theta\left(\frac{\partial^2}{\partial\theta^2} \log L(\theta|\mathbf{X})\right)} = 1$$

and so this approximation matches the first two moments of a  $n(0, 1)$  random variable. Wilks (1938) proved that these intervals have an asymptotic optimality property; they are, asymptotically, the shortest in a certain class of intervals.

Of course, these intervals are not totally general and may not always be applicable to a function  $h(\theta)$ . We must be able to express (9.4.2) as a function of  $h(\theta)$ .

**Example 9.4.2:** Again using a binomial example, if  $Y = \sum_{i=1}^n X_i$ , where each  $X_i$  is an independent Bernoulli( $p$ ) random variable, we have

$$\begin{aligned} Q(Y|p) &= \frac{\frac{\partial}{\partial p} \log L(p|Y)}{\sqrt{-E_p\left(\frac{\partial^2}{\partial p^2} \log L(p|Y)\right)}} \\ &= \frac{\frac{y}{p} - \frac{n-y}{1-p}}{\sqrt{\frac{n}{p(1-p)}}} \\ &= \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}, \end{aligned}$$

where  $\hat{p} = y/n$ . Using (9.4.2), an approximate  $1 - \alpha$  confidence interval is given by

$$(9.4.3) \quad \left\{ p: \left| \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \right| \leq z_{\alpha/2} \right\}.$$

This interval is the same as that which results from inverting one of the approximate tests given in Example 8.4.2. (Also see Example 9.4.5 for details.) ||

### 9.4.2 Other Approximate Intervals

Most approximate confidence intervals are based on either finding approximate (or asymptotic) pivots or inverting approximate level  $\alpha$  test statistics. If we have any statistics  $W$  and  $V$  and a parameter  $\theta$  such that, as  $n \rightarrow \infty$ ,

$$\frac{W - \theta}{V} \rightarrow n(0, 1),$$

then we can form the approximate confidence interval for  $\theta$  given by

$$W - z_{\alpha/2}V \leq \theta \leq W + z_{\alpha/2}V.$$

In particular, direct application of the Central Limit Theorem, together with Slutsky's Theorem, will usually give an approximate confidence interval. (Note that the approximate maximum likelihood intervals of the previous section are a special case of this strategy.)

**Example 9.4.3:** If  $X_1, \dots, X_n$  are iid with mean  $\mu$  and variance  $\sigma^2$  then, from the Central Limit Theorem,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow n(0, 1).$$

Moreover, from Slutsky's Theorem, if  $S^2 \rightarrow \sigma^2$  in probability then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \rightarrow n(0, 1).$$

Either of these facts can be used to construct an approximate confidence interval. ||

In the above example, we could get an approximate confidence interval without specifying the form of the sampling distribution. We should be able to do better when we do specify the form.

**Example 9.4.4:** If  $X_1, \dots, X_n$  are iid  $\text{Poisson}(\lambda)$ , then we know that

$$\frac{\bar{X} - \lambda}{S/\sqrt{n}} \rightarrow n(0, 1).$$

However, this is true even if we did not sample from a Poisson population. Using the Poisson assumption, we know that  $\text{Var}(X) = \lambda = E\bar{X}$  and  $\bar{X}$  is a good estimator (Chapter 7) of  $\lambda$ . Thus, using the Poisson assumption, we could also get an approximate confidence interval from the fact that

$$\frac{\bar{X} - \lambda}{\sqrt{\bar{X}/n}} \rightarrow n(0, 1).$$

We can use that Poisson assumption in another way. Since  $\text{Var}(X) = \lambda$ , it follows that

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \rightarrow n(0, 1).$$

Of these approximations, the last gives the best interval (according to Wilks, 1938). The interval based on the last approximation is the likelihood interval of (9.4.2) (see Exercise 9.59). ||

Generally speaking, a reasonable rule of thumb is to use as few estimates and as many parameters as possible in an approximation. This is sensible for a very simple reason. Parameters are fixed and do not introduce any added variability into an approximation while each statistic brings more variability along with it.

**Example 9.4.5 (Continuation of Examples 8.4.2 and 9.4.2):** For a random sample  $X_1, \dots, X_n$  from a Bernoulli( $p$ ) population, we saw in Example 8.4.2 that, as  $n \rightarrow \infty$ , both

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}} \quad \text{and} \quad \frac{\hat{p} - p}{\sqrt{p(1 - p)/n}}$$

converge in distribution to a standard normal random variable, where  $\hat{p} = \sum x_i/n$ . In Example 8.4.2 we saw that we could base tests on either approximation. We also know that we can use either approximation to form a confidence interval for  $p$ . However, the second approximation (the one with fewer statistics and more parameter values) will give the interval (9.4.3) from Example 9.4.2, which is the asymptotically optimal one. That is,

$$\left\{ p: \left| \frac{\hat{p} - p}{\sqrt{p(1 - p)/n}} \right| \leq z_{\alpha/2} \right\},$$

is the better approximate interval.

It is not immediately clear what this interval looks like, but we can explicitly solve for the set of values. If we square both sides and rearrange terms, we are looking for the set of values of  $p$  that satisfy

$$\left\{ p: (\hat{p} - p)^2 \leq z_{\alpha/2}^2 \frac{p(1-p)}{n} \right\}.$$

This inequality is a quadratic in  $p$ , which can be put in a more familiar form through some further rearrangement:

$$\left\{ p: \left( 1 + \frac{z_{\alpha/2}^2}{n} \right) p^2 - \left( 2\hat{p} + \frac{z_{\alpha/2}^2}{n} \right) p + \hat{p}^2 \leq 0 \right\}.$$

Since the coefficient of  $p^2$  in the quadratic is positive, the quadratic opens upward and, thus, the inequality is satisfied if  $p$  lies between the two roots of the quadratic. These two roots are

$$\frac{2\hat{p} + z_{\alpha/2}^2/n \pm \sqrt{(2\hat{p} + z_{\alpha/2}^2/n)^2 - 4\hat{p}^2(1 + z_{\alpha/2}^2/n)}}{2(1 + z_{\alpha/2}^2/n)}$$

and the roots define the endpoints of the confidence interval for  $p$ . Although the expressions for the roots are somewhat nasty, the interval is, in fact, a very good interval for  $p$ . The interval can be further improved, however, by using a continuity correction (Example 3.2.2). To do this, we would solve two separate quadratics (see Exercise 9.61),

$$\begin{aligned} \left| \frac{\hat{p} + \frac{1}{2n} - p}{\sqrt{p(1-p)/n}} \right| &\leq z_{\alpha/2}, & \text{(larger root = upper interval endpoint)} \\ \left| \frac{\hat{p} - \frac{1}{2n} - p}{\sqrt{p(1-p)/n}} \right| &\leq z_{\alpha/2}. & \text{(smaller root = lower interval endpoint)} \end{aligned}$$

At the endpoints there are obvious modifications. If  $\sum x_i = 0$ , then the lower interval endpoint is taken to be 0 while, if  $\sum x_i = n$ , then the upper interval endpoint is taken to be 1. See Blyth (1986) for some good approximations. ||

Thus far, all of the approximations mentioned have been based on letting  $n \rightarrow \infty$ . However, there are other situations where we might use approximate intervals. For example, in Example 2.3.6 we saw that for certain parameter configurations, the Poisson distribution can be used to approximate the binomial. This suggests that, if such a parameter configuration is believed to be likely, then an approximate binomial interval can be based on the Poisson distribution. In that spirit we illustrate the following somewhat unusual case.

**Example 9.4.6:** Let  $X_1, \dots, X_n$  be iid negative binomial( $r, p$ ). We assume that  $r$  is known and we are interested in a confidence interval for  $p$ . Using the fact that  $Y = \sum X_i \sim \text{negative binomial}(nr, p)$ , we can form intervals in a number of ways. Using a variation of the binomial- $F$  distribution relationship, we can form an exact

confidence interval (Exercise 9.22) or we can use a normal approximation (Exercise 9.60). There is another approximation, which does not rely on large  $n$ , but rather small  $p$ .

In Exercise 2.38 it is established that, as  $p \rightarrow 0$ ,

$$2pY \rightarrow \chi^2_{2nr} \text{ in distribution.}$$

So, for small  $p$ ,  $2pY$  is a pivot! Using this fact, we can construct a pivotal  $1 - \alpha$  confidence interval, valid for small  $p$ :

$$\left\{ p: \frac{\chi^2_{2nr,1-\alpha/2}}{2y} \leq p \leq \frac{\chi^2_{2nr,\alpha/2}}{2y} \right\}.$$

Details are in Exercise 9.62. ||

## EXERCISES

---

- 9.1** If  $L(x)$  and  $U(x)$  satisfy  $P_\theta(L(X) \leq \theta) = 1 - \alpha_1$  and  $P_\theta(U(X) \geq \theta) = 1 - \alpha_2$ , and  $L(x) \leq U(x)$  for all  $x$ , show that  $P_\theta(L(X) \leq \theta \leq U(X)) = 1 - \alpha_1 - \alpha_2$ .
- 9.2** Let  $X_1, \dots, X_n$  be iid  $n(\theta, 1)$ . A 95% confidence interval for  $\theta$  is  $\bar{x} \pm 1.96/\sqrt{n}$ . Let  $p$  denote the probability that an additional independent observation,  $X_{n+1}$ , will fall in this interval. Is  $p$  greater than, less than, or equal to .95? Prove your answer.
- 9.3** The independent random variables  $X_1, \dots, X_n$  have the common distribution

$$P(X_i \leq x) = \begin{cases} 0 & \text{if } x \leq 0 \\ (x/\beta)^\alpha & \text{if } 0 < x < \beta \\ 1 & \text{if } x \geq \beta \end{cases}.$$

- a. In Exercise 7.10 the MLEs of  $\alpha$  and  $\beta$  were found. If  $\alpha$  is a known constant,  $\alpha_0$ , find an upper confidence limit for  $\beta$  with confidence coefficient .95.
- b. Use the data of Exercise 7.10 to construct an interval estimate for  $\beta$ . Assume that  $\alpha$  is known and equal to its MLE.
- 9.4** Let  $X_1, \dots, X_n$  be a random sample from a  $n(0, \sigma_X^2)$  and let  $Y_1, \dots, Y_m$  be a random sample from a  $n(0, \sigma_Y^2)$ , independent of the  $X$ s. Define  $\lambda = \sigma_Y^2/\sigma_X^2$ .
- Find the level  $\alpha$  LRT of  $H_0: \lambda = \lambda_0$  versus  $H_1: \lambda \neq \lambda_0$ .
  - Express the rejection region of the LRT of part (a) in terms of an  $F$  random variable.
  - Find a  $1 - \alpha$  confidence interval for  $\lambda$ .
- 9.5** In Example 9.2.4 a lower confidence bound was put on  $p$ , the success probability from a sequence of Bernoulli trials. This exercise will derive an upper confidence bound. That is, observing  $X_1, \dots, X_n$ , where  $X_i \sim \text{Bernoulli}(p)$ , we want an interval of the form  $[0, U(x_1, \dots, x_n)]$ , where  $P_p(p \in [0, U(x_1, \dots, x_n)]) \geq 1 - \alpha$ .
- Show that inversion of the acceptance region of the test

$$H_0: p = p_0 \quad \text{versus} \quad H_1: p < p_0$$

will give a confidence interval of the desired confidence level and form.

- b. Find equations, similar to those given in (9.2.7), that can be used to construct the confidence interval.
- 9.6** a. Derive a confidence interval for a binomial  $p$  by inverting the LRT of  $H_0: p = p_0$  versus  $H_1: p \neq p_0$ .  
 b. Show that the interval is a highest density region from  $p^y(1-p)^{n-y}$  and is not equal to the interval in (9.4.3).
- 9.7** a. Find the  $1 - \alpha$  confidence set for  $a$  that is obtained by inverting the LRT of  $H_0: a = a_0$  versus  $H_1: a \neq a_0$  based on a sample  $X_1, \dots, X_n$  from a  $n(\theta, a\theta)$  family, where  $\theta$  is unknown.  
 b. A similar question can be asked about the related family, the  $n(\theta, a\theta^2)$  family. If  $X_1, \dots, X_n$  are iid  $n(\theta, a\theta^2)$ , where  $\theta$  is unknown, find the  $1 - \alpha$  confidence set based on inverting the LRT of  $H_0: a = a_0$  versus  $H_1: a \neq a_0$ .
- 9.8** Given a sample  $X_1, \dots, X_n$  from a pdf of the form  $\frac{1}{\sigma} f((x - \theta)/\sigma)$ , list at least five different pivotal quantities.
- 9.9** Show that each of the three quantities listed in Example 9.2.5 is a pivot.
- 9.10** Suppose that  $T$  is a real-valued statistic. Suppose that  $Q(t, \theta)$  is a monotone function of  $t$  for each value of  $\theta \in \Theta$ . Show that if the pdf of  $T$ ,  $f(t|\theta)$ , can be expressed in the form

$$f(t|\theta) = g(Q(t, \theta)) \left| \frac{\partial}{\partial t} Q(t, \theta) \right|,$$

for some function  $g$ , then  $Q(T, \theta)$  is a pivot.

- 9.11** Find a pivotal quantity based on a random sample of size  $n$  from a  $n(\theta, \theta)$  population where  $\theta > 0$ . Use the pivotal quantity to set up a  $1 - \alpha$  confidence interval for  $\theta$ .
- 9.12** Let  $X$  be a single observation from the beta( $\theta, 1$ ) pdf.  
 a. Let  $Y = -(\log X)^{-1}$ . Evaluate the confidence coefficient of the set  $[y/2, y]$ .  
 b. Find a pivotal quantity and use it to set up a confidence interval having the same confidence coefficient as the interval in part (a).  
 c. Compare the two confidence intervals.
- 9.13** Let  $X_1, \dots, X_n$  be iid  $n(\mu, \sigma^2)$ , where both parameters are unknown. Simultaneous inference on both  $\mu$  and  $\sigma$  can be made using the Bonferroni Inequality in a number of ways.  
 a. Using the Bonferroni Inequality, combine the two confidence sets

$$\left\{ \mu: \bar{x} - \frac{ks}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{ks}{\sqrt{n}} \right\} \quad \text{and} \quad \left\{ \sigma^2: \frac{(n-1)s^2}{b} \leq \sigma^2 \leq \frac{(n-1)s^2}{a} \right\}$$

into one confidence set for  $(\mu, \sigma)$ . Show how to choose  $a$ ,  $b$ , and  $k$  to make the simultaneous set a  $1 - \alpha$  confidence set.

- b. Using the Bonferroni Inequality, combine the two confidence sets

$$\left\{ \mu: \bar{x} - \frac{k\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{k\sigma}{\sqrt{n}} \right\} \quad \text{and} \quad \left\{ \sigma^2: \frac{(n-1)s^2}{b} \leq \sigma^2 \leq \frac{(n-1)s^2}{a} \right\}$$

into one confidence set for  $(\mu, \sigma)$ . Show how to choose  $a$ ,  $b$ , and  $k$  to make the simultaneous set a  $1 - \alpha$  confidence set.

- c. Compare the confidence sets in parts (a) and (b).

**9.14** Solve for the roots of the quadratic equation that defines Fieller's confidence set for the ratio of normal means (see the *Miscellanea* section). Find conditions on the random variables for which

- The parabola opens upward (the confidence set is an interval).
- The parabola opens downward (the confidence set is the complement of an interval).
- The parabola has no real roots.

In each case, give an interpretation of the meaning of the confidence set. For example, what would you tell an experimenter if, for his data, the parabola had no real roots?

**9.15** Let  $X_1, \dots, X_n$  be iid  $n(\theta, \sigma^2)$ , where  $\sigma^2$  is known. For each of the following hypotheses, write out the acceptance region of a level  $\alpha$  test and the  $1 - \alpha$  confidence interval that results from inverting the test.

- $H_0: \theta = \theta_0$  versus  $H_1: \theta \neq \theta_0$ .
- $H_0: \theta \geq \theta_0$  versus  $H_1: \theta < \theta_0$ .
- $H_0: \theta \leq \theta_0$  versus  $H_1: \theta > \theta_0$ .

**9.16** Find a  $1 - \alpha$  confidence interval for  $\theta$ , given  $X_1, \dots, X_n$  iid with pdf

- $f(x|\theta) = 1, \theta - \frac{1}{2} < x < \theta + \frac{1}{2}$
- $f(x|\theta) = 2x/\theta^2, 0 < x < \theta, \theta > 0$ .

**9.17** In this exercise we will investigate some more properties of binomial confidence sets and the Sterne (1954) construction in particular. As in Example 9.2.8, we will again consider the binomial(3,  $p$ ) distribution.

- Draw, as a function of  $p$ , a graph of the four probability functions  $P_p(X = x)$ ,  $x = 0, \dots, 3$ . Identify the maxima of  $P_p(X = 1)$  and  $P_p(X = 2)$ .
- Show that for small  $\epsilon$ ,  $P_p(X = 0) > P_p(X = 2)$  for  $p = \frac{1}{3} + \epsilon$ .
- Show that the *most probable construction* is to blame for the difficulties with the Sterne sets by showing that the following acceptance regions can be inverted to obtain a  $1 - \alpha = .442$  confidence interval.

$p$	Acceptance region = $A(p)$
[.000, .238]	0
(.238, .305)	0, 1
[.305, .362]	1
(.362, .634)	1, 2
[.634, .695]	2
(.695, .762)	2, 3
[.762, 1.00]	3

(This is essentially Crow's (1956) modification of Sterne's construction. More recent exact binomial intervals are given by Blyth and Still (1983) and Casella (1986).)

**9.18** This is a slight generalization of the confidence interval construction method of Theorem 9.2.2. Suppose that expression (9.2.12) is changed to be

$$F_T(t|\theta_U(t)) = \alpha_1, \quad F_T(t|\theta_L(t)) = 1 - \alpha_2.$$

Show that Theorem 9.2.2 can be suitably modified to result in a  $1 - \alpha_1 - \alpha_2$  confidence interval for  $\theta$ .

**9.19** Prove part (b) of Theorem 9.2.2.

**9.20** Some of the details of the proof of Theorem 9.2.3 need to be filled in, and the second part of the theorem needs to be proved.

- If the cdf  $F_T(t|\theta)$  is a decreasing function of  $\theta$  for each  $t$ , show that the function  $\bar{F}_T(t|\theta)$  defined by  $\bar{F}_T(t|\theta) = P(T \geq t|\theta)$  is a nondecreasing function of  $\theta$  for each  $t$ .
  - Show that if  $F_T(T|\theta)$  is stochastically greater than or equal to a uniform random variable then so is  $\bar{F}_T(T|\theta)$ . That is, if  $P_\theta(F_T(T|\theta) \leq x) \leq x$  for every  $x, 0 \leq x \leq 1$ , then  $P_\theta(\bar{F}_T(T|\theta) \leq x) \leq x$  for every  $x, 0 \leq x \leq 1$ .
  - Prove part (b) of Theorem 9.2.3.
- 9.21** In Example 9.2.10 it was shown that a confidence interval for a Poisson parameter can be expressed in terms of chi squared cutoff points. Use a similar technique to show that if  $X \sim \text{binomial}(n, p)$ , then a  $1 - \alpha$  confidence interval for  $p$  is

$$\frac{1}{1 + \frac{n-x+1}{x} F_{2(n-x+1), 2x, \alpha/2}} \leq p \leq \frac{\frac{x+1}{n-x} F_{2(x+1), 2(n-x), \alpha/2}}{1 + \frac{x+1}{n-x} F_{2(x+1), 2(n-x), \alpha/2}},$$

where  $F_{\nu_1, \nu_2, \alpha}$  is the upper  $\alpha$  cutoff from an  $F$  distribution with  $\nu_1$  and  $\nu_2$  degrees of freedom, and we make the endpoint adjustment that the lower endpoint is 0 if  $x = 0$  and the upper endpoint is 1 if  $x = n$ . (Hint: Recall the following identity from Exercise 2.40, which can be interpreted in the following way. If  $X \sim \text{binomial}(n, \theta)$ , then  $P_\theta(X \geq x) = P(Y \leq \theta)$ , where  $Y \sim \text{beta}(x, n - x + 1)$ . Use the properties of the  $F$  and beta distributions from Chapter 5.)

- 9.22** If  $X \sim \text{negative binomial}(r, p)$  use the relationship between the binomial and negative binomial to show that a  $1 - \alpha$  confidence interval for  $p$  is given by

$$\frac{1}{1 + \frac{x+1}{r} F_{2(x+1), 2r, \alpha/2}} \leq p \leq \frac{\frac{r}{x} F_{2r, 2x, \alpha/2}}{1 + \frac{r}{x} F_{2r, 2x, \alpha/2}},$$

with a suitable modification if  $x = 0$ .

- 9.23** a. Let  $X_1, \dots, X_n$  be a random sample from a Poisson population with parameter  $\lambda$  and define  $Y = \sum X_i$ . In Example 9.2.10 a confidence interval for  $\lambda$  was found using the method of Section 9.2.3. Construct another interval for  $\lambda$ , by inverting an LRT, and compare the intervals.  
 b. The following data, the number of aphids per row in nine rows of a potato field, can be assumed to follow a Poisson distribution:

155, 104, 66, 50, 36, 40, 30, 35, 42.

Use these data to construct a 90% LRT confidence interval for the mean number of aphids per row. Also, construct an interval using the method of Example 9.2.10.

- 9.24** For  $X \sim \text{Poisson}(\lambda)$ , show that the coverage probability of the confidence interval  $[L(X), U(X)]$  in Example 9.2.10 is given by

$$P_\lambda(\lambda \in [L(X), U(X)]) = \sum_{x=0}^{\infty} I_{[L(x), U(x)]}(\lambda) \frac{e^{-\lambda} \lambda^x}{x!},$$

and that we can define functions  $x_l(\lambda)$  and  $x_u(\lambda)$  so that

$$P_\lambda(\lambda \in [L(X), U(X)]) = \sum_{x=x_1(\lambda)}^{x_u(\lambda)} \frac{e^{-\lambda} \lambda^x}{x!}.$$

Hence, explain why the graph of the coverage probability of the Poisson intervals given in Figure 9.2.3 has jumps occurring at the endpoints of the different confidence intervals.

- 9.25** If  $X_1, \dots, X_n$  are iid with pdf  $f(x|\mu) = e^{-(x-\mu)} I_{[\mu, \infty)}(x)$ , then  $Y = \min\{X_1, \dots, X_n\}$  is sufficient for  $\mu$  with pdf

$$f_Y(y|\mu) = n e^{-n(y-\mu)} I_{[\mu, \infty)}(y).$$

In Example 9.2.9 a  $1 - \alpha$  confidence interval for  $\mu$  was found using the method of Section 9.2.3. Compare that interval to  $1 - \alpha$  intervals obtained by likelihood and pivotal methods.

- 9.26** Let  $X_1, \dots, X_n$  be iid observations from a beta( $\theta, 1$ ) pdf and assume that  $\theta$  has a gamma( $r, \lambda$ ) prior pdf. Find a  $1 - \alpha$  Bayes credible set for  $\theta$ .
- 9.27** a. Let  $X_1, \dots, X_n$  be iid observations from an exponential( $\lambda$ ) pdf where  $\lambda$  has the conjugate IG( $a, b$ ) prior, an inverted gamma with pdf

$$\pi(\lambda|a, b) = \frac{1}{\Gamma(a)b^a} \left(\frac{1}{\lambda}\right)^{a+1} e^{-1/(b\lambda)}, \quad 0 < \lambda < \infty.$$

Show how to find a  $1 - \alpha$  Bayes HPD credible set for  $\lambda$ .

- b. Find a  $1 - \alpha$  Bayes HPD credible set for  $\sigma^2$ , the variance of a normal distribution, based on the sample variance  $s^2$  and using a conjugate IG( $a, b$ ) prior for  $\sigma^2$ .
- c. Starting with the interval from part (b), find the limiting  $1 - \alpha$  Bayes HPD credible set for  $\sigma^2$  obtained as  $a \rightarrow 0$  and  $b \rightarrow \infty$ .
- 9.28** Let  $X_1, \dots, X_n$  be iid  $n(\theta, \sigma^2)$ , where both  $\theta$  and  $\sigma^2$  are unknown, but there is only interest on inference about  $\theta$ . Consider the prior pdf

$$\pi(\theta, \sigma^2|\mu, \tau^2, a, b) = \frac{1}{\sqrt{2\pi\tau^2\sigma^2}} e^{-(\theta-\mu)^2/(2\tau^2\sigma^2)} \frac{1}{\Gamma(a)b^a} \left(\frac{1}{\sigma^2}\right)^{a+1} e^{-1/(b\sigma^2)},$$

a  $n(\mu, \tau^2\sigma^2)$  multiplied by an IG( $a, b$ ).

- a. Show that this prior is a conjugate prior for this problem.
- b. Find the posterior distribution of  $\theta$  and use it to construct a  $1 - \alpha$  credible set for  $\theta$ .
- c. The classical  $1 - \alpha$  confidence set for  $\theta$  can be expressed as

$$\left\{ \theta : |\theta - \bar{x}|^2 \leq F_{1, n-1, \alpha/2} \frac{s^2}{n} \right\}.$$

Is there any (limiting) configuration of  $\tau^2$ ,  $a$ , and  $b$  that would allow this set to be approached by a Bayes set from part (b)?

- 9.29** If  $X_1, \dots, X_n$  are a sequence of  $n$  Bernoulli( $p$ ) trials,
- a. Calculate a  $1 - \alpha$  credible set for  $p$  using the conjugate beta( $a, b$ ) prior.
- b. Using the relationship between the beta and  $F$  distributions, write the credible set in a form that is comparable to the form of the intervals in Exercise 9.21. Do the intervals match for any values of  $a$  and  $b$ ?

**9.30** In this exercise we will calculate the classical coverage probability of the HPD region in (9.2.15), that is, the coverage probability of the Bayes HPD region using the probability model  $\bar{X} \sim n(\theta, \sigma^2/n)$ .

- a. Using the definitions given in Example 9.2.12, prove that

$$\begin{aligned} P_\theta \left( |\theta - \delta^B(\bar{X})| \leq z_{\alpha/2} \sqrt{\text{Var}(\theta|\bar{X})} \right) \\ = P_\theta \left[ -\sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\theta-\mu)}{\sigma/\sqrt{n}} \leq Z \leq \sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\theta-\mu)}{\sigma/\sqrt{n}} \right]. \end{aligned}$$

- b. Show that the above set, although a  $1 - \alpha$  credible set, is not a  $1 - \alpha$  confidence set. (Fix  $\theta \neq \mu$ , let  $\tau = \sigma/\sqrt{n}$ , so that  $\gamma = 1$ . Prove that as  $\sigma^2/n \rightarrow 0$  the above probability goes to zero.)  
c. If  $\theta = \mu$ , however, prove that the coverage probability is bounded away from zero. Find the minimum and maximum of this coverage probability.  
d. Now we will look at the other side. The usual  $1 - \alpha$  confidence set for  $\theta$  is  $\{\theta : |\theta - \bar{x}| \leq z_{\alpha/2}\sigma/\sqrt{n}\}$ . Show that the credible probability of this set is given by

$$\begin{aligned} P_{\bar{x}} (|\theta - \bar{x}| \leq z_{\alpha/2}\sigma/\sqrt{n}) \\ = P_{\bar{x}} \left[ -\sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\bar{x}-\mu)}{\sqrt{1+\gamma}\sigma/\sqrt{n}} \leq Z \leq \sqrt{1+\gamma}z_{\alpha/2} + \frac{\gamma(\bar{x}-\mu)}{\sqrt{1+\gamma}\sigma/\sqrt{n}} \right] \end{aligned}$$

and that this probability is not bounded away from zero. Hence, the  $1 - \alpha$  confidence set is not a  $1 - \alpha$  credible set.

**9.31** Let  $X \sim n(\mu, 1)$  and consider the confidence interval,

$$C_a(x) = \{\mu : \min\{0, (x-a)\} \leq \mu \leq \max\{0, (x+a)\}\}.$$

- a. For  $a = 1.645$ , prove that the coverage probability of  $C_a(x)$  is exactly .95 for all  $\mu$ , with the exception of  $\mu = 0$ , where the coverage probability is 1.  
b. Now consider the noninformative prior  $\pi(\mu) = 1$ , discussed in Example 9.3.7. Using this prior and again taking  $a = 1.645$ , show that the posterior credible probability of  $C_a(x)$  is exactly .90 for  $-1.645 \leq x \leq 1.645$  and increases to .95 as  $|x| \rightarrow \infty$ .

(Is this interval just a theoretical oddity or does it have some statistical usefulness? Suppose that  $\mu = \mu_1 - \mu_2$ , the difference in means of two different populations. If the means are different, we would like to make such an inference conclusively (confidence interval does not cover zero), while if the means are the same, we would like a nice tight interval around zero (showing that there is little difference in population means). The above interval accomplishes these goals and, from a frequentist view, is better for this type of inference than the usual interval,  $\{\mu : x - a \leq \mu \leq x + a\}$ . *Communicated by Jason Hsu, Ohio State University.*)

**9.32** In this exercise some of the details of Example 9.2.14 will be completed.

- a. Show that the problem of testing  $H_0: \mu = \mu_0$  versus  $H_1: \mu > \mu_0$  is invariant with respect to the group  $\mathcal{G}_{\mu_0}$ .  
b. Show that any test based on the statistic  $(\bar{x} - \mu_0)/s$  is an invariant test of  $H_0: \mu = \mu_0$  versus  $H_1: \mu > \mu_0$  with respect to the group  $\mathcal{G}_{\mu_0}$ .  
c. Verify that  $\bar{g}(\mu, \sigma^2) = (a\mu + b, a^2\sigma^2)$  is the correct expression for the transformation of the parameter point  $\theta = (\mu, \sigma^2)$  under the transformation  $g_{a,b}$ .

**9.33** Consider the general location model from Examples 9.2.13 and 9.3.5.

- a. Show that the family of distributions of  $(X_1, \dots, X_n)$  is invariant under the group  $\mathcal{G} = \{g_a(\mathbf{x}), -\infty < a < \infty\}$ , where  $g_a(x_1, \dots, x_n) = (x_1 + a, \dots, x_n + a)$ .
- b. Let  $T(\mathbf{x})$  be any invariant estimate. That is, for all  $\mathbf{x} \in \mathcal{X}$  and all  $a$ ,  $T(x_1 + a, \dots, x_n + a) = T(x_1, \dots, x_n) + a$ . Show that any confidence interval of the form

$$C(\mathbf{x}) = \{\theta : T(\mathbf{x}) - k_1 \leq \theta \leq T(\mathbf{x}) + k_2\},$$

where  $k_1$  and  $k_2$  are constants, is an invariant confidence interval.

- c. Name three invariant estimators for this problem.

**9.34** Suppose that  $X_1, \dots, X_n$  is a random sample from a  $n(\mu, \sigma^2)$  population.

- a. If  $\sigma^2$  is known, find a minimum value for  $n$  to guarantee that a .95 confidence interval for  $\mu$  will have length no more than  $\sigma/4$ .
- b. If  $\sigma^2$  is unknown, find a minimum value for  $n$  to guarantee, with probability .90, that a .95 confidence interval for  $\mu$  will have length no more than  $\sigma/4$ .

**9.35** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population. Compare expected lengths of  $1 - \alpha$  confidence intervals for  $\mu$  that are computed assuming

- a.  $\sigma^2$  is known.
- b.  $\sigma^2$  is unknown.

**9.36** Let  $X_1, \dots, X_n$  be independent with pdfs  $f_{X_i}(x|\theta) = e^{i\theta-x} I_{[i\theta, \infty)}(x)$ . Prove that  $T = \min_i(X_i/i)$  is a sufficient statistic for  $\theta$ . Based on  $T$ , find the  $1 - \alpha$  confidence interval for  $\theta$  of the form  $[T + a, T + b]$  which is of minimum length.

**9.37** Let  $X_1, \dots, X_n$  be iid uniform( $0, \theta$ ). Let  $Y$  be the largest order statistic. Prove that  $Y/\theta$  is a pivotal quantity and show that the interval

$$\left\{ \theta : y \leq \theta \leq \frac{y}{\alpha^{1/n}} \right\}$$

is the shortest  $1 - \alpha$  pivotal interval.

**9.38** Condition (b) in Theorem 9.3.1 need not be satisfied for a shortest interval. Give an example, using a discontinuous pdf, of a shortest interval that does not satisfy (b).

**9.39** Prove a special case of Theorem 9.3.1. Let  $X \sim f(x)$ , where  $f$  is a *symmetric unimodal* pdf. For a fixed value of  $1 - \alpha$ , of all intervals  $[a, b]$  that satisfy  $\int_a^b f(x) dx = 1 - \alpha$ , the shortest is obtained by choosing  $a$  and  $b$  so that  $\int_{-\infty}^a f(x) dx = \alpha/2$  and  $\int_b^{\infty} f(x) dx = \alpha/2$ .

**9.40** a. Prove the following, which is related to Theorem 9.3.1. Let  $X \sim f(x)$ , where  $f$  is a *strictly decreasing* pdf on  $[0, \infty)$ . For a fixed value of  $1 - \alpha$ , of all intervals  $[a, b]$  that satisfy  $\int_a^b f(x) dx = 1 - \alpha$ , the shortest is obtained by choosing  $a = 0$  and  $b$  so that  $\int_0^b f(x) dx = 1 - \alpha$ .

b. Use the result of part (a) to find the shortest  $1 - \alpha$  confidence interval in Example 9.2.9.

**9.41** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population, where both  $\mu$  and  $\sigma^2$  are unknown. Consider confidence intervals for  $\sigma^2$  of the form

$$\left\{ \sigma^2 : \frac{(n-1)s^2}{a} \leq \sigma^2 \leq \frac{(n-1)s^2}{b} \right\},$$

where  $s^2$  is the sample variance and  $a$  and  $b$  are constants.

- a. Find the shortest length  $1 - \alpha$  confidence interval of this form. (That is, for a given value of  $\alpha$ , find the equations that can be used to solve for  $a$  and  $b$ .)  
 b. For  $\alpha = .1$  and  $n = 3$ , find the numerical values of  $a$  and  $b$ . Compare the length of this interval to the one obtained by splitting  $\alpha$  equally.
- 9.42** With one observation from a gamma( $k, \beta$ ) pdf with known shape parameter  $k$ , find the shortest  $1 - \alpha$  (pivotal) confidence interval of the form

$$\left\{ \beta : \frac{x}{b} \leq \beta \leq \frac{x}{a} \right\}.$$

- 9.43** State and prove the analogue of Theorem 9.3.2 for inversion of the UMP test of the hypotheses  $H_0: \theta = \theta_0$  versus  $H_1: \theta < \theta_0$ , where  $\theta_0$  is arbitrary.

- 9.44** Let  $f(x|\theta)$  be the location pdf

$$f(x|\theta) = \frac{e^{(x-\theta)}}{(1 + e^{(x-\theta)})^2}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

Based on one observation,  $x$ , find the UMA one-sided  $1 - \alpha$  confidence interval of the form  $\{\theta: \theta \leq U(x)\}$ .

- 9.45** Let  $X_1, \dots, X_n$  be iid exponential( $\lambda$ ).

- a. Find a UMP size  $\alpha$  hypothesis test of  $H_0: \lambda = \lambda_0$  versus  $H_1: \lambda < \lambda_0$ .  
 b. Find a UMA  $1 - \alpha$  confidence interval based on inverting the test in part (a). Show that the interval can be expressed as

$$C^*(x_1, \dots, x_n) = \left\{ \lambda : 0 \leq \lambda \leq \frac{2 \sum x_i}{\chi_{2n,\alpha}^2} \right\}.$$

- c. Find the expected length of  $C^*(x_1, \dots, x_n)$ .  
 d. Madansky (1962) exhibited a  $1 - \alpha$  interval whose expected length is shorter than that of the UMA interval. In general, Madansky's interval is difficult to calculate, but in the following situation calculation is relatively simple. Let  $1 - \alpha = .3$  and  $n = 120$ . Madansky's interval is

$$C^M(x_1, \dots, x_n) = \left\{ \lambda : 0 \leq \lambda \leq -\frac{x_{(1)}}{\log(.99)} \right\},$$

which is a 30% confidence interval. Use the fact that  $\chi_{240,.7}^2 = 251.046$  to show that the 30% UMA interval satisfies

$$E[\text{Length}(C^*(x_1, \dots, x_n))] = .956\lambda > E[\text{Length}(C^M(x_1, \dots, x_n))] = .829\lambda$$

- 9.46** Let  $X_1, \dots, X_n$  be iid Poisson( $\lambda$ ). Find a UMA  $1 - \alpha$  confidence interval based on inverting the UMP level  $\alpha$  test of  $H_0: \lambda = \lambda_0$  versus  $H_1: \lambda > \lambda_0$ .

- 9.47** Prove Theorem 9.3.3.

- 9.48** Let  $X_1, \dots, X_n$  be iid  $n(\theta, \sigma^2)$ , where  $\sigma^2$  is known.

- a. Show that the usual one-sided  $1 - \alpha$  upper confidence bound for  $\theta$  of the form  $\{\theta: \theta \leq \bar{x} + z_\alpha \sigma / \sqrt{n}\}$  is UMA.  
 b. Show that the usual one-sided  $1 - \alpha$  lower confidence bound for  $\theta$  of the form  $\{\theta: \theta \geq \bar{x} - z_\alpha \sigma / \sqrt{n}\}$  is UMA.

- c. Show that the usual two-sided  $1 - \alpha$  interval for  $\theta$  of the form

$$\{\theta : \bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \leq \theta \leq \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}\}$$

is UMA unbiased.

- 9.49** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population, where  $\sigma^2$  is unknown.

- a. Show that the interval

$$\theta \leq \bar{x} + t_{n-1, \alpha} \frac{s}{\sqrt{n}}$$

can be derived by inverting the acceptance region of an LRT.

- b. Show that the interval in part (a) is a UMA unbiased interval.

- 9.50** Let  $X_1, \dots, X_n$  be iid  $n(\theta, \sigma^2)$ , where  $\sigma^2$  is unknown.

- a. Show that the two-sided interval in (9.2.11) can be derived by inverting an LRT.

- b. Show that the interval in part (a) is a UMA unbiased interval.

- 9.51** (Cox's Paradox) We are to test

$$H_0: \theta = \theta_0 \quad \text{versus} \quad H_1: \theta > \theta_0,$$

where  $\theta$  is the mean of one of two normal distributions and  $\theta_0$  is a fixed but arbitrary value of  $\theta$ . We observe the random variable  $X$  with distribution

$$X \sim \begin{cases} n(\theta, 100) & \text{with probability } p \\ n(\theta, 1) & \text{with probability } 1 - p \end{cases}$$

- a. Show that the test given by

$$\text{reject } H_0 \text{ if } X > \theta_0 + z_\alpha \sigma,$$

where  $\sigma = 1$  or  $10$  depending on which population is sampled, is a level  $\alpha$  test. Derive a  $1 - \alpha$  confidence set by inverting the acceptance region of this test.

- b. Show that a more powerful level  $\alpha$  test (for  $\alpha > p$ ) is given by

$$\text{reject } H_0 \text{ if } X > \theta_0 + z_{(\alpha-p)/(1-p)} \text{ and } \sigma = 1, \text{ otherwise always reject } H_0.$$

Derive a  $1 - \alpha$  confidence set by inverting the acceptance region of this test, and show that it is the empty set with positive probability. (Cox's Paradox states that classic optimal procedures sometimes ignore the information about conditional distributions, and provide us with a procedure that, while optimal, is somehow unreasonable.)

- 9.52** Let  $X \sim f(x|\theta)$ , and suppose that the interval  $\{\theta : a(X) \leq \theta \leq b(X)\}$  is a UMA confidence set for  $\theta$ .

- a. Find a UMA confidence set for  $1/\theta$ . Note that if  $a(x) < 0 < b(x)$ , this set is  $\{1/\theta : 1/b(x) \leq 1/\theta \leq 1/a(x)\} \cup \{1/\theta : 1/\theta \leq 1/a(x)\}$ . Hence it is possible for the UMA confidence set to be neither an interval nor bounded.

- b. Show that, if  $h$  is an increasing function, the set  $\{h(\theta) : h(a(X)) \leq h(\theta) \leq h(b(X))\}$  is a UMA confidence set for  $h(\theta)$ . Can the condition on  $h$  be relaxed?

- 9.53** If  $X_1, \dots, X_n$  are iid from a location pdf  $f(x - \theta)$ , show that the confidence set

$$C(x_1, \dots, x_n) = \{\theta : \bar{x} - k_1 \leq \theta \leq \bar{x} + k_2\},$$

where  $k_1$  and  $k_2$  are constants, has constant coverage probability. (*Hint:* The pdf of  $\bar{X}$  is of the form  $f_{\bar{X}}(\bar{x} - \theta)$ .)

- 9.54** Let  $X$  have pdf of the form  $\frac{1}{\tau} f(\frac{x}{\tau})$  and consider the scale group

$$\mathcal{G} = \{g_c(x) : g_c(x) = cx, c > 0\}.$$

- a. Prove that invariant confidence intervals are of the form

$$C(x) = \left\{ \tau : k_1 \leq \frac{x}{\tau} \leq k_2 \right\}.$$

- b. Show that  $C(x)$  has constant coverage probability for all values of  $\tau$ .

- c. Show that if we measure the size of  $C(x)$  by  $\text{Size}[C(x)] = k_2/k_1$ , this is an invariant measure of size.

- 9.55** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population, where both  $\mu$  and  $\sigma^2$  are unknown. Each of the following methods of finding confidence intervals for  $\sigma^2$  results in intervals of the form

$$\left\{ \sigma^2 : \frac{(n-1)s^2}{b} \leq \sigma^2 \leq \frac{(n-1)s^2}{a} \right\},$$

but in each case  $a$  and  $b$  will satisfy different constraints. The intervals given in this exercise are derived by Tate and Klett (1959), who also tabulate some cutoff points.

Define  $f_p(t)$  to be the pdf of a  $\chi_p^2$  random variable with  $p$  degrees of freedom. In order to have a  $1 - \alpha$  confidence interval,  $a$  and  $b$  must satisfy

$$\int_a^b f_{n-1}(t) dt = 1 - \alpha,$$

but additional constraints are required to define  $a$  and  $b$  uniquely. Verify that each of the following constraints can be derived as stated.

- a. *The likelihood ratio interval:* The  $1 - \alpha$  confidence interval obtained by inverting the LRT of  $H_0: \sigma = \sigma_0$  versus  $H_1: \sigma \neq \sigma_0$  is of the above form where  $a$  and  $b$  also satisfy  $f_{n+2}(a) = f_{n+2}(b)$ .
- b. *The minimum length interval:* For intervals of the above form, the  $1 - \alpha$  confidence interval obtained by minimizing the interval length constrains  $a$  and  $b$  to satisfy  $f_{n+3}(a) = f_{n+3}(b)$ .
- c. *The shortest unbiased interval:* For intervals of the above form, the  $1 - \alpha$  confidence interval obtained by minimizing the probability of false coverage among all unbiased intervals constrains  $a$  and  $b$  to satisfy  $f_{n+1}(a) = f_{n+1}(b)$ .
- d. *The equal-tail interval:* For intervals of the above form, the  $1 - \alpha$  confidence interval obtained by requiring that the probability above and below the interval be equal constrains  $a$  and  $b$  to satisfy

$$\int_0^a f_{n-1}(t) dt = \frac{\alpha}{2}, \int_b^\infty f_{n-1}(t) dt = \frac{\alpha}{2}.$$

(This interval, although very common, is clearly nonoptimal no matter what length criterion is used.)

- 9.56** We return to the problem of finding a confidence interval for a normal variance. Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population, where both  $\mu$  and  $\sigma^2$  are unknown.

- a. Show that intervals of the form

$$C(s^2) = \left\{ \sigma^2 : \frac{(n-1)s^2}{a} \leq \sigma^2 \leq \frac{(n-1)s^2}{b} \right\}$$

are invariant with respect to the scale group

$$\mathcal{G} = \{g_c(s^2) : g_c(s^2) = cs^2, c > 0\}.$$

- b. Using the invariant measure of size,  $\text{Size}[C(s^2)] = a/b$  (Example 9.3.6), find the best invariant  $1 - \alpha$  confidence interval for  $\sigma^2$ . (That is, find equations that uniquely determine  $a$  and  $b$ .)
- c. Show that the best invariant interval found in part (b) can also be derived as the Bayes  $1 - \alpha$  HPD region against the prior  $\pi(\sigma^2) = 1/\sigma^2$ .
- d. Show that the best invariant interval found in part (b) coincides with the *shortest unbiased* interval derived in Exercise 9.55. (Note that, in this situation, considerations of unbiasedness and invariance lead to the same interval.)
- 9.57** Prove that the Bayes HPD credible interval given in Example 9.3.7 coincides with the confidence interval derived through either likelihood or invariance considerations.
- 9.58** In Example 9.3.7 a noninformative prior was derived for the location model. In this exercise a similar derivation will be presented for the scale model. Suppose that  $X \sim (1/\tau)f(x/\tau)$  and recall that the scale family  $\mathcal{F} = \{(1/\tau)f(x/\tau) : 0 < \tau < \infty\}$  is invariant under the group  $\mathcal{G} = \{g_c(x) : 0 < c < \infty\}$ , where  $g_c(x) = cx$ . To find a prior pdf  $\pi(\tau)$  that is invariant, an argument as in Example 9.3.7 could lead to considering prior cdfs that satisfy

*Measurement Invariance:*  $F_\tau(\tau_1) - F_\tau(\tau_2) = F_\gamma(\gamma_1) - F_\gamma(\gamma_2)$ ,

*Formal Invariance:*  $F_\tau(\tau_1) - F_\tau(\tau_2) = F_\gamma(\tau_1) - F_\gamma(\tau_2)$ ,

where  $\tau_i = c\gamma_i, i = 1, 2$ .

- a. Produce an argument that would lead to the above two invariance equations for the prior cdf.
- b. Combine these two equations to deduce that a solution is

$$F_\tau(\tau) = \log \tau,$$

and that  $\pi(\tau) = 1/\tau$  is a noninformative (invariant) prior pdf.

- c. Find the Bayes HPD credible set for  $\tau$  using the prior derived in part (b).  
d. Does the Bayes set from part (c) coincide with either of the classic confidence sets derived through likelihood or invariance considerations?
- 9.59** In Example 9.4.4 we saw that the Poisson assumption, together with the Central Limit Theorem, could be used to form an approximate interval based on the fact that

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \rightarrow n(0, 1).$$

Show that this approximation is optimal according to Wilks (1938). That is, show that

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} = \frac{\frac{\partial}{\partial \lambda} \log L(\lambda | \mathbf{X})}{\sqrt{-E_\lambda \left( \frac{\partial^2}{\partial \lambda^2} \log L(\lambda | \mathbf{X}) \right)}}.$$

- 9.60** Let  $X_1, \dots, X_n$  be iid negative binomial( $r, p$ ). We want to construct some approximate confidence intervals for the negative binomial parameters.
- a. Calculate the quantity

$$Q(\mathbf{x}|p) = \frac{\frac{\partial}{\partial p} \log L(p | \mathbf{x})}{\sqrt{-E_p \left( \frac{\partial^2}{\partial p^2} \log L(p | \mathbf{x}) \right)}}$$

(Wilks' approximation from Section 9.4) and show how to form confidence intervals with this expression.

- b. Find an approximate  $1 - \alpha$  confidence interval for the *mean* of the negative binomial distribution. Show how to incorporate the continuity correction into your interval.  
c. The aphid data of Exercise 9.23 can also be modeled using the negative binomial distribution. Construct an approximate 90% confidence interval for the aphid data using the results of part (b). Compare the interval to the Poisson-based intervals of Exercise 9.23.
- 9.61** Solve for the endpoints of the approximate binomial confidence interval, with continuity correction, given in Example 9.4.5. Show that this interval is wider than the corresponding interval without continuity correction and, also, the continuity corrected interval has a uniformly higher coverage probability. (In fact, the coverage probability of the uncorrected interval does not maintain  $1 - \alpha$ ; it dips below this level for some parameter values. The corrected interval does maintain a coverage probability greater than  $1 - \alpha$  for all parameter values.)
- 9.62** Let  $X_1, \dots, X_n$  be iid negative binomial( $r, p$ ).  
a. Complete the details of Example 9.4.6, that is, show that for small  $p$  the interval

$$\left\{ p: \frac{\chi_{2nr, 1-\alpha/2}^2}{2 \sum x} \leq p \leq \frac{\chi_{2nr, \alpha/2}^2}{2 \sum x} \right\},$$

is an approximate  $1 - \alpha$  confidence interval.

- b. Show how to choose the endpoints in order to obtain a minimum length  $1 - \alpha$  interval.

- 9.63** For the case of Fieller's confidence set (see the *Miscellanea* section), that is, given a random sample  $(X_1, Y_1), \dots, (X_n, Y_n)$  from a bivariate normal distribution with parameters  $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ , find an approximate confidence interval for  $\theta = \mu_Y/\mu_X$ . Use the approximate moment calculations in Example 7.4.3 and apply the Central Limit Theorem.

## Miscellanea

---

### Sufficient Statistics and Confidence Sets

In the Chapter 8 *Miscellanea* section we discussed the relationship between sufficient statistics and hypothesis tests. In light of a result like Theorem 9.2.1, we would expect a similar relationship between sufficient statistics and confidence sets. Again, such a result exists, but we need to resort to the concept of *randomized confidence sets* to implement the relationship in a mathematically elegant way. We stress that randomized confidence sets, like randomized tests, are theoretical tools; although we could establish a relationship between sufficiency and confidence sets without them, it would be cumbersome. In practice, however, we rely on sufficiency and avoid randomized procedures.

A theorem, similar to the one proved in the Chapter 8 *Miscellanea* section, can be proved in the following way. We define a (possibly randomized) confidence function  $\psi(\theta|\mathbf{x})$  as follows.  $\psi(\theta|\mathbf{x})$  is the probability of including  $\theta$  in the confidence set when  $\mathbf{x}$  is observed, and can have the form

$$\psi(\theta|\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in A(\theta) \\ \gamma & \text{if } \mathbf{x} \in B(\theta) \\ 0 & \text{if } \mathbf{x} \notin A(\theta) \cup B(\theta) \end{cases}$$

where  $0 \leq \gamma \leq 1$ . It is now clear that if  $T$  is sufficient for  $\theta$ , we can take the conditional expectation of  $\psi$  and get a procedure as good, since

$$P_\theta(\theta \text{ covered}) = E_\theta(\psi(\theta|\mathbf{X})) = E_\theta [E(\psi(\theta|\mathbf{X})|T)].$$

The end result of all this is that, when looking for a confidence set we need to consider only sets based on sufficient statistics.

### Confidence Procedures

Confidence sets and tests can be related formally by defining an entity called a *confidence procedure* (Joshi, 1969). If  $X \sim f(x|\theta)$ , where  $x \in \mathcal{X}$  and  $\theta \in \Theta$ , then a confidence procedure is a set in the space  $\mathcal{X} \times \Theta$ , the Cartesian product space. It is defined as

$$\{(x, \theta) : (x, \theta) \in C\},$$

for a set  $C \in \mathcal{X} \times \Theta$ .

From the confidence procedure we can define two slices, or sections, obtained by holding one of the variables constant. For fixed  $x$ , we define the  $\theta$ -section or confidence set as

$$C(x) = \{\theta : (x, \theta) \in C\}.$$

For fixed  $\theta$  we define the *x-section* or acceptance region as

$$A(\theta) = \{x: (x, \theta) \in C\}.$$

Although this development necessitates working with the product space  $\mathcal{X} \times \Theta$ , which is one reason why we do not use it here, it does provide a more straightforward way of seeing the relationship between tests and sets. Figure 9.2.1 illustrates this correspondence in the normal case.

### **Fieller's Theorem**

Fieller's Theorem (Fieller, 1954) is a clever argument to get an exact confidence set on a ratio of normal means.

Given a random sample  $(X_1, Y_1), \dots, (X_n, Y_n)$  from a bivariate normal distribution with parameters  $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ , a confidence set on  $\theta = \mu_Y/\mu_X$  can be formed in the following way. For  $i = 1, \dots, n$ , define  $Z_{\theta i} = Y_i - \theta X_i$  and  $\bar{Z}_\theta = \bar{Y} - \theta \bar{X}$ . It can be shown that  $\bar{Z}_\theta$  is normal with mean 0 and variance

$$V_\theta = \frac{1}{n} (\sigma_Y^2 - 2\theta\rho\sigma_Y\sigma_X + \theta^2\sigma_X^2).$$

$V_\theta$  can be estimated with  $\hat{V}_\theta$ , given by

$$\begin{aligned}\hat{V}_\theta &= \frac{1}{n(n-1)} \sum_{i=1}^n (Z_{\theta i} - \bar{Z}_\theta)^2 \\ &= \frac{1}{n-1} (S_Y^2 - 2\theta S_{YX} + \theta^2 S_X^2),\end{aligned}$$

where

$$S_Y^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2, \quad S_X^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2, \quad S_{YX} = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}).$$

Furthermore, it can also be shown that  $E\hat{V}_\theta = V_\theta$ ,  $\hat{V}_\theta$  is independent of  $\bar{Z}_\theta$ , and  $(n-1)\hat{V}_\theta/V_\theta \sim \chi^2_{n-1}$ . Hence,  $\bar{Z}_\theta/\sqrt{\hat{V}_\theta} \sim t_{n-1}$  and the set

$$\left\{ \theta: \frac{\bar{z}_\theta^2}{\hat{v}_\theta} \leq t_{n-1, \alpha/2}^2 \right\}$$

defines a  $1 - \alpha$  confidence set for  $\theta$ , the ratio of the means. This set defines a parabola in  $\theta$  and the roots of the parabola give the endpoints of the confidence set. Writing the set in terms of the original variables, we get

$$\left\{ \theta: \left( \bar{x}^2 - \frac{t_{n-1, \alpha/2}^2}{n-1} S_x^2 \right) \theta^2 - 2 \left( \bar{x}\bar{y} - \frac{t_{n-1, \alpha/2}^2}{n-1} S_{yx} \right) \theta + \left( \bar{y}^2 - \frac{t_{n-1, \alpha/2}^2}{n-1} S_y^2 \right) \leq 0 \right\}.$$

One interesting feature of this set is that, depending on the roots of the parabola, it can be an interval, the complement of an interval, or the whole real line (see Exercise 9.14). Furthermore, to maintain  $1 - \alpha$  confidence, this interval must be infinite with positive probability. See Hwang (1990).

# 0 Decision Theory

*"For a mixture of the modern and the mediaeval, of the practical and the wildly fanciful, I think this is surely the limit," said he. "What do you make of it, Watson?"*

Sherlock Holmes

*The Adventure of the Sussex Vampire*

## 10.1 Introduction

All of the forms of inference we have discussed—point estimation, hypothesis testing, and interval estimation—often involve making decisions. *Decision theory* is a study of inference problems in which all parts of the decision making process are formally defined, including desired optimality criteria. These criteria are then used to compare alternative decision procedures. Hence, decision theory provides an alternative method of analyzing inference problems, oftentimes providing similar conclusions to analyses we have done in earlier chapters but sometimes providing surprising new insights.

In earlier chapters we have discussed all the elements of a decision problem, but we have not always formally named or defined them. In a decision theoretic formulation, all these elements must be specified. The *data* are described by a random vector  $\mathbf{X}$  with *sample space*  $\mathcal{X}$ . The *model* is the set of possible probability distributions for  $\mathbf{X}$ , indexed by a parameter  $\theta$ . The parameter  $\theta$  is the true but unknown state of nature about which we wish to make an inference. In all problems we consider,  $\theta$  will be real- or vector-valued. The set of possible values for  $\theta$  is called the *parameter space* and is denoted by  $\Theta$ . Thus the model is a set  $\{f(\mathbf{x}|\theta) : \theta \in \Theta\}$  where each  $f(\mathbf{x}|\theta)$  is a pdf or pmf on  $\mathcal{X}$ .

After the data  $\mathbf{X} = \mathbf{x}$  are observed, a decision regarding  $\theta$  is made. The set of allowable decisions is the *action space*, denoted by  $\mathcal{A}$ . The action space determines whether the inference problem at hand is a point estimation problem, a hypothesis testing problem, an interval estimation problem, or some other type of problem. In a point estimation problem, the allowable actions are point “guesses” at the value of  $\theta$ , where any possible value of  $\theta$  is a possible guess. Usually, in point estimation problems,  $\mathcal{A}$  is equal to  $\Theta$ . In a hypothesis testing problem, only two actions are allowable, “accept  $H_0$ ” or “reject  $H_0$ .” These two actions might be denoted  $a_0$  and  $a_1$ , respectively. The action space in hypothesis testing is the two-point set  $\mathcal{A} = \{a_0, a_1\}$ . In an interval estimation problem the actions are intervals or, more generally, subsets in the parameter space. Thus, in such a problem,  $\mathcal{A}$  might be the set of all subsets of  $\Theta$ . In general, elements of the action space will be denoted by  $a \in \mathcal{A}$ .

If  $\theta \in \Theta$  is the true state of nature, the action  $a \in \mathcal{A}$  may be correct, may be wrong but not too far wrong, or may be grossly incorrect. This relationship is

quantified by the *loss function*, denoted by  $L(\theta, a)$ , that gives the "loss" incurred if  $\theta$  is the true state of nature and action  $a$  is taken. Although vector-valued losses are sometimes appropriate, in all examples we will consider the loss will be real-valued. Thus the loss function is a function from  $\Theta \times \mathcal{A}$  into  $\mathbb{R}$ . Larger values of  $L(\theta, a)$  indicate that  $a$  is more incorrect and smaller values indicate that  $a$  is less incorrect. Usually  $L(\theta, a) = 0$  means that  $a$  is the correct decision if  $\theta$  is the true state of nature but this relationship is not required. If loss is measured in money, then  $L(\theta, a) = -20,000$  might indicate that the payoff is \$20,000 if the decision  $a$  is made. Instead of a loss function, one may speak of a *utility function*, a function that is usually used in economic theory or in some branches of Bayesian analysis. When using a utility function we speak of gain rather than loss, for the utility function is merely the negative of the loss function. (See Berger (1985) for a complete discussion of utility functions and *utility theory*, a prescription for behavior in which a person acts to maximize a personal utility function.)

A *decision rule* is a rule that specifies, for each  $x \in \mathcal{X}$ , what action  $a \in \mathcal{A}$  will be taken if  $X = x$  is observed. Thus a decision rule, denoted by  $\delta(x)$ , is just a function from  $\mathcal{X}$  into  $\mathcal{A}$ . For example, if the decision problem is a hypothesis testing problem, then

$$\delta(x) = a_0 \quad \text{for all } x \text{ that are in the acceptance region of the test}$$

and

$$\delta(x) = a_1 \quad \text{for all } x \text{ that are in the rejection region of the test.}$$

The set of all allowable decision rules will be denoted by  $\mathcal{D}$ . We will want  $\mathcal{D}$  to be as large as possible so that all reasonable decision rules are considered. For example,  $\mathcal{D}$  could be *all* functions from  $\mathcal{X}$  into  $\mathcal{A}$ . We will then use the optimality properties, to be developed, to determine which  $\delta(x) \in \mathcal{D}$  are good decision rules.

In a decision theoretic analysis, the quality of a decision rule is quantified in the *risk function* of the decision rule. For a decision rule  $\delta(x)$ , the risk function, a function of  $\theta$ , is defined to be

$$(10.1.1) \quad R(\theta, \delta) = E_\theta L(\theta, \delta(X)).$$

At a given  $\theta$ , the risk function is the average loss that will be incurred if the decision rule  $\delta(x)$  is used.

Since the true value of  $\theta$  is unknown, we would like to use a decision rule that has a small value of  $R(\theta, \delta)$  for all values of  $\theta$ . This would mean that, regardless of the true value of  $\theta$ , the decision rule will have a small expected loss. If the qualities of two different decision rules,  $\delta_1$  and  $\delta_2$ , are to be compared, then they will be compared by comparing their risk functions,  $R(\theta, \delta_1)$  and  $R(\theta, \delta_2)$ . If  $R(\theta, \delta_1) < R(\theta, \delta_2)$  for all  $\theta \in \Theta$ , then  $\delta_1$  is the preferred decision rule because  $\delta_1$  performs better for all  $\theta$ . More typically, the two risk functions will cross. Then the judgment as to which decision rule is better may not be so clear-cut. But still, in a decision theoretic analysis, the comparison of the decision rules must be based on the risk functions, and only on the

risk functions. This use of the loss and risk functions to entirely describe the quality of decision rules is the distinguishing characteristic of a decision theoretic analysis.

It may not be immediately clear why the risk function, rather than the loss function itself, is used to assess the quality of a decision rule. With a little reflection, however, we see that the risk is a very reasonable measure of the worth of a decision rule. Although we would like  $L(\theta, \delta(x))$  always to be small, we must consider the relationship between  $\theta$  and  $x$  as described by the pdf or pmf  $f(x|\theta)$ . If, for a given  $\theta$ ,  $x$  is a “likely” observation then we should use a decision rule,  $\delta(x)$ , that makes  $L(\theta, \delta(x))$  small. But, for a given  $\theta$ , if  $x$  is an “unlikely” observation then the value of  $L(\theta, \delta(x))$  is of lesser concern. Typically, there is some other  $\theta'$  for which  $x$  is a “likely” observation and we want  $L(\theta', \delta(x))$  to be small. These considerations are summarized in the risk function where the weighting of  $f(x|\theta)$  is used to calculate the average loss.  $R(\theta, \delta)$  will be small if  $L(\theta, \delta(x))$  is small for all “likely” values of  $x$ .

The above mentioned elements—parameter space, model, action space, and loss function—must be specified to carry out a decision theoretic analysis of a problem. One other element is sometimes used, a *prior distribution*, denoted by  $\pi(\theta)$ , a probability distribution on the parameter space. We have discussed prior distributions previously. In particular, we have used them in previous chapters to incorporate opinions about the parameter, held prior to sampling, into the analysis.

In a decision theoretic analysis prior distributions can be used as an opinion summary, but they can also have other uses. Sometimes prior distributions are simply used to summarize the information about a decision rule that is in the risk function, acting as a weight function in computing an average risk. The *Bayes risk* of a decision rule  $\delta$  with respect to a prior distribution  $\pi(\theta)$  is defined to be

$$(10.1.2) \quad B(\pi, \delta) = E_{\pi} R(\theta, \delta)$$

where  $E_{\pi}$  refers to the expectation in which  $\theta$  is the random variable with probability distribution  $\pi(\theta)$ .

The Bayes risk summarizes the risk function in a single number, small values indicating that the risk of the decision rule  $\delta$  is small on the average. In later sections we will see that *Bayes rules*, decision rules that minimize the Bayes risk, have certain optimality properties, regardless of whether we consider the prior distribution as a summary of subjective prior information or as a convenient method of summarizing the risk function.

In a mathematically rigorous description of a decision theoretic problem, care must be taken in the definitions to ensure that the various quantities mentioned above are well defined. In particular, a decision rule  $\delta$  must be a function from  $\mathcal{X}$  into  $\mathcal{A}$  such that the expectation defining the risk function in (10.1.1) is well defined. There sometimes exist functions for which this is not the case. So, strictly speaking,  $\mathcal{D}$  may not be *all* functions from  $\mathcal{X}$  into  $\mathcal{A}$ . But any function we could easily describe has a well-defined expectation. We will leave these technical considerations to a more advanced treatment of decision theory.

We have now seen the basic elements in a decision problem. In the next section, we consider some specific examples. In subsequent sections we will discuss different

criteria for comparing decision rules and general methods for finding optimal decision rules.

## 10.2 Common Decision Theoretic Analyses

In the next three subsections we will apply the ideas developed in the previous section to statistical problems we have already encountered. We will present some examples of the general decision theoretic concepts in the context of point estimation, hypothesis testing, and interval estimation.

### 10.2.1 Point Estimation

In a point estimation problem, the possible actions are the various possible values of  $\theta$ , so usually,  $\mathcal{A} = \Theta$ . Sometimes, it will be useful to have  $\mathcal{A}$  be a bigger set than  $\Theta$ , that is,  $\Theta \subset \mathcal{A}$ . For example, if  $\theta$  is a binomial success probability, we might know that  $0 < \theta < 1$ . That is, we know that a success has positive probability ( $\theta > 0$ ) but is not certain ( $\theta < 1$ ). However, we might wish to use  $\mathcal{A} = \{a : 0 \leq a \leq 1\}$  because otherwise, some reasonable estimators would not be decision rules. For example, the maximum likelihood estimator estimates  $\theta$  to be zero if no successes are observed, hence is in  $\mathcal{A}$  only if  $\mathcal{A}$  includes 0. Generally speaking, aside from these types of exceptions, the specification of  $\mathcal{A} = \Theta$  is what characterizes a decision problem as a point estimation problem.

The loss function in a point estimation problem reflects the fact that if an action  $a$  is close to  $\theta$ , then the decision  $a$  is reasonable and little loss is incurred. If  $a$  is far from  $\theta$  then a large loss is incurred. The loss function generally increases as the distance between  $a$  and  $\theta$  increases. If  $\theta$  is real-valued, two commonly used loss functions are

$$\text{absolute error loss, } L(\theta, a) = |a - \theta|$$

and

$$\text{squared error loss, } L(\theta, a) = (a - \theta)^2.$$

Both of these loss functions increase as the distance between  $\theta$  and  $a$  increases. Squared error loss gives relatively more penalty for large discrepancies and absolute error loss gives relatively more penalty for small discrepancies. A variation of squared error loss, one that penalizes overestimation more than underestimation, is

$$L(\theta, a) = \begin{cases} (a - \theta)^2 & \text{if } a < \theta \\ 10(a - \theta)^2 & \text{if } a \geq \theta \end{cases}$$

A loss that penalizes errors in estimation more if  $\theta$  is near zero than if  $|\theta|$  is large, a relative squared error loss, is

$$L(\theta, a) = \frac{(a - \theta)^2}{|\theta| + 1}.$$

Notice that both of these last variations of squared error loss could have been based instead on absolute error loss. In general, the experimenter must consider the consequences of various errors in estimation for different values of  $\theta$  and specify a loss function that reflects these consequences.

The risk function for a decision rule  $\delta$  is the expected loss, as defined in (10.1.1). For squared error loss, the risk function is a familiar quantity, the mean squared error (MSE) that was used in Chapter 7. In Chapter 7 the MSE of an estimator was defined as  $MSE(\theta) = E_\theta(\delta(\mathbf{X}) - \theta)^2$ , which is just  $E_\theta L(\theta, \delta(\mathbf{X})) = R(\theta, \delta)$  if  $L(\theta, a) = (a - \theta)^2$ . As in Chapter 7 we have that, for squared error loss,

$$(10.2.1) \quad R(\theta, \delta) = \text{Var}_\theta \delta(\mathbf{X}) + (E_\theta \delta(\mathbf{X}) - \theta)^2 = \text{Var}_\theta \delta(\mathbf{X}) + (\text{Bias}_\theta \delta(\mathbf{X}))^2.$$

This risk function for squared error loss clearly indicates that a good estimator should have both a small variance and a small bias. A decision theoretic analysis would judge how well an estimator succeeded in simultaneously minimizing these two quantities.

It would be an atypical decision theoretic analysis in which the set  $\mathcal{D}$  of allowable estimators was restricted to the set of unbiased estimators, as was done in Chapter 7. Then, minimizing the risk would just be minimizing the variance. A decision theoretic analysis would be more comprehensive in that both the variance and bias are in the risk and will be considered simultaneously. An estimator would be judged good if it had a small, but probably nonzero, bias combined with a small variance.

As mentioned above, an ideal decision theoretic analysis would let  $\mathcal{D}$  be as large as possible and would use the risk functions to compare estimators. We will compute some risk functions and compare some estimators in the following examples. Since any time we calculated the  $MSE(\theta)$  for an estimator in Chapter 7 we were actually calculating the risk function for the estimator using squared error loss, some of the calculations done there can be used in a decision theoretic analysis.

**Example 10.2.1:** In Example 7.3.2 we considered  $X_1, \dots, X_n$ , a random sample from a  $Bernoulli(p)$  population. We considered two estimators,

$$\hat{p}_B = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n + \sqrt{n}} \quad \text{and} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

The risk functions for these two estimators, for  $n = 4$  and  $n = 400$ , are graphed in Figure 7.3.1 and the comparisons of these risk functions are as stated in Example 7.3.2. On the basis of risk comparison, the estimator  $\hat{p}_B$  would be preferred for small  $n$  and the estimator  $\bar{X}$  would be preferred for large  $n$ . ||

**Example 10.2.2:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population. Consider estimating  $\sigma^2$  using squared error loss. We will consider estimators of the form  $\delta_b(\mathbf{X}) = bS^2$  where  $S^2$  is the sample variance and  $b$  can be any nonnegative constant. Recall that  $ES^2 = \sigma^2$  and, for a normal sample,  $\text{Var } S^2 = 2\sigma^4/(n-1)$ . Using (10.2.1), we can compute the risk function for  $\delta_b$  as

$$\begin{aligned}
 R((\mu, \sigma^2), \delta_b) &= \text{Var } bS^2 + (\text{E}bS^2 - \sigma^2)^2 \\
 &= b^2 \text{Var } S^2 + (b\text{E}S^2 - \sigma^2)^2 \\
 &= \frac{b^2 2\sigma^4}{n-1} + (b-1)^2 \sigma^4 \quad (\text{using } \text{Var } S^2) \\
 &= \left[ \frac{2b^2}{n-1} + (b-1)^2 \right] \sigma^4.
 \end{aligned}$$

The risk function for  $\delta_b$  does not depend on  $\mu$  and is a quadratic function of  $\sigma^2$ . This quadratic function is of the form  $c_b(\sigma^2)^2$  where  $c_b$  is a positive constant. To compare two risk functions, and hence the worth of two estimators, note that if  $c_b < c_{b'}$  then

$$R((\mu, \sigma^2), \delta_b) = c_b(\sigma^2)^2 < c_{b'}(\sigma^2)^2 = R((\mu, \sigma^2), \delta_{b'})$$

for all values of  $(\mu, \sigma^2)$ . Thus  $\delta_b$  would be a better estimator than  $\delta_{b'}$ . The value of  $b$  that gives the overall minimum value of

$$(10.2.2) \quad c_b = \frac{2b^2}{n-1} + (b-1)^2$$

yields the best estimator  $\delta_b$  in this class. Standard calculus methods show that  $b = (n-1)/(n+1)$  is the minimizing value. Thus, at every value of  $(\mu, \sigma^2)$ , the estimator

$$\tilde{S}^2 = \frac{n-1}{n+1} S^2 = \frac{1}{n+1} \sum (X_i - \bar{X})^2$$

has the smallest risk among all estimators of the form  $bS^2$ . For  $n = 5$ , the risk functions for this estimator and two other estimators in this class are shown in Figure 10.2.1. The other estimators are  $S^2$ , the unbiased estimator, and  $\hat{\sigma}^2 = \frac{n-1}{n} S^2$ , the MLE of  $\sigma^2$ . It is clear that the risk function for  $\tilde{S}^2$  is smallest, everywhere.

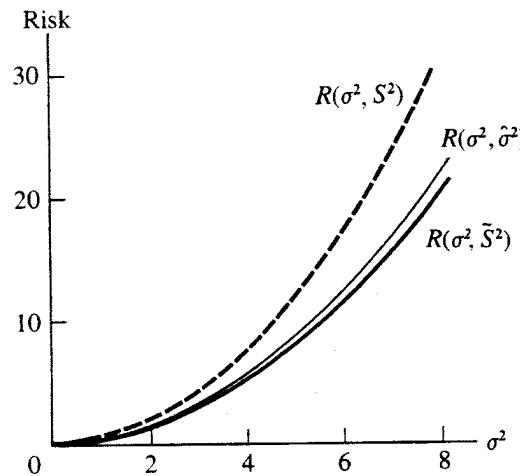


FIGURE 10.2.1 Risk functions for three variance estimators in Example 10.2.2

**Example 10.2.3:** Again we consider estimating a population variance  $\sigma^2$  with an estimator of the form  $bS^2$ . In this analysis we can be quite general and assume only that  $X_1, \dots, X_n$  is a random sample from some population with positive, finite variance  $\sigma^2$ . Now we will use the loss function

$$L(\sigma^2, a) = \frac{a}{\sigma^2} - 1 - \log \frac{a}{\sigma^2}.$$

This loss is more complicated than squared error loss but it has some reasonable properties. Note that if  $a = \sigma^2$ , the loss is zero. Also, for any fixed value of  $\sigma^2$ ,  $L(\sigma^2, a) \rightarrow \infty$  as  $a \rightarrow 0$  or  $a \rightarrow \infty$ . That is, gross underestimation is penalized just as heavily as gross overestimation. (A criticism of squared error loss in a variance estimation problem is that underestimation has only a finite penalty while overestimation has an infinite penalty.) The loss function, sometimes known as Stein's loss, also arises out of the likelihood function for  $\sigma^2$ , if this is a sample from a normal population, and thus ties together good decision theoretic properties with good likelihood properties. (See Exercise 10.4.)

For the estimator  $\delta_b = bS^2$ , the risk function is

$$\begin{aligned} R(\sigma^2, \delta_b) &= E \left( \frac{bS^2}{\sigma^2} - 1 - \log \frac{bS^2}{\sigma^2} \right) \\ &= bE \frac{S^2}{\sigma^2} - 1 - E \log \frac{bS^2}{\sigma^2} \\ &= b - \log b - 1 - E \log \frac{S^2}{\sigma^2}. \quad \left( E \frac{S^2}{\sigma^2} = 1 \right) \end{aligned}$$

$E \log(S^2/\sigma^2)$  may be a function of  $\sigma^2$  and other population parameters but it is not a function of  $b$ . Thus  $R(\sigma^2, \delta_b)$  is minimized in  $b$ , for all  $\sigma^2$ , by the value of  $b$  that minimizes  $b - \log b$ , that is,  $b = 1$ . Therefore the estimator of the form  $bS^2$  that has the smallest risk for all values of  $\sigma^2$  is  $\delta_1 = S^2$ . ||

## 10.2.2 Hypothesis Testing

A hypothesis testing problem is characterized by the fact that the action space consists of only two elements,  $\mathcal{A} = \{a_0, a_1\}$ . The decision  $a_0$  is the decision to accept that  $H_0: \theta \in \Theta_0$  is true, and the decision  $a_1$  is the decision to accept that  $H_1: \theta \in \Theta_0^c$  is true. A decision rule  $\delta(x)$  is a function on  $\mathcal{X}$  that takes on only two values,  $a_0$  and  $a_1$ . The set  $\{x : \delta(x) = a_0\}$  is the acceptance region for the test and the set  $\{x : \delta(x) = a_1\}$  is the rejection region, just as in Definition 8.1.3.

The loss function in a hypothesis testing problem should reflect the fact that, if  $\theta \in \Theta_0$  and decision  $a_1$  is made, or if  $\theta \in \Theta_0^c$  and decision  $a_0$  is made, a mistake has been made. But in the other two possible cases, the correct decision has been made. Since there are only two possible actions, the loss function  $L(\theta, a)$  in a hypothesis testing problem is composed of only two parts. The function  $L(\theta, a_0)$  is the loss

incurred for various values of  $\theta$  if the decision to accept  $H_0$  is made and  $L(\theta, a_1)$  is the loss incurred for various values of  $\theta$  if the decision to reject  $H_0$  is made.

The simplest kind of loss in a testing problem is called *0–1 loss*, and is defined by

$$L(\theta, a_0) = \begin{cases} 0 & \theta \in \Theta_0 \\ 1 & \theta \in \Theta_0^c \end{cases} \quad \text{and} \quad L(\theta, a_1) = \begin{cases} 1 & \theta \in \Theta_0 \\ 0 & \theta \in \Theta_0^c \end{cases}.$$

With 0–1 loss, the value 0 is lost if a correct decision is made and the value 1 is lost if an incorrect decision is made. This is a particularly simple situation in which both types of error have the same consequence. A slightly more realistic loss, one that gives different costs to the two types of error, is *generalized 0–1 loss*,

$$(10.2.3) \quad L(\theta, a_0) = \begin{cases} 0 & \theta \in \Theta_0 \\ c_{II} & \theta \in \Theta_0^c \end{cases} \quad \text{and} \quad L(\theta, a_1) = \begin{cases} c_I & \theta \in \Theta_0 \\ 0 & \theta \in \Theta_0^c \end{cases}.$$

In this loss,  $c_I$  is the cost of a Type I Error, the error of falsely rejecting  $H_0$ , and  $c_{II}$  is the cost of a Type II Error, the error of falsely accepting  $H_0$ . (Actually, in comparing tests, all that really matters is the ratio  $c_{II}/c_I$ , not the two individual values. If  $c_I = c_{II}$ , we essentially have 0–1 loss.)

In Chapter 8, a hypothesis testing procedure was evaluated through its power function; however, in a decision theoretic analysis, the risk function (the expected loss) is used to evaluate a hypothesis testing procedure. These two functions are closely related, as the following analysis shows.

Let  $\beta(\theta)$  be the power function of the test based on the decision rule  $\delta$ . That is, if  $R = \{\mathbf{x}: \delta(\mathbf{x}) = a_1\}$  denotes the rejection region of the test, then

$$\beta(\theta) = P_\theta(\mathbf{X} \in R) = P_\theta(\delta(\mathbf{X}) = a_1).$$

The risk function associated with (10.2.3) and, in particular, 0–1 loss is very simple. For any value of  $\theta \in \Theta$ ,  $L(\theta, a)$  takes on only two values, 0 and  $c_I$  if  $\theta \in \Theta_0$  and 0 and  $c_{II}$  if  $\theta \in \Theta_0^c$ . Thus the risk is

$$(10.2.4) \quad \begin{aligned} R(\theta, \delta) &= 0P_\theta(\delta(\mathbf{X}) = a_0) + c_I P_\theta(\delta(\mathbf{X}) = a_1) = c_I \beta(\theta), & \text{if } \theta \in \Theta_0, \\ &\text{and} \\ R(\theta, \delta) &= c_{II} P_\theta(\delta(\mathbf{X}) = a_0) + 0P_\theta(\delta(\mathbf{X}) = a_1) = c_{II}(1 - \beta(\theta)), & \text{if } \theta \in \Theta_0^c. \end{aligned}$$

This similarity between a decision theoretic approach and a more traditional power approach is due, in part, to the form of the loss function. But in all hypothesis testing problems, as we shall see below, the power function plays an important role in the risk function.

**Example 10.2.4:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population,  $\sigma^2$  known. The UMP level  $\alpha$  test of  $H_0: \theta \geq \theta_0$  versus  $H_1: \theta < \theta_0$  is the test that

rejects  $H_0$  if  $(\bar{X} - \theta_0)/(\sigma/\sqrt{n}) < -z_\alpha$  (Example 8.3.7). The power function for this test is

$$\beta(\theta) = P_\theta \left( Z < -z_\alpha - \frac{\theta - \theta_0}{\sigma/\sqrt{n}} \right)$$

where  $Z$  has a  $n(0, 1)$  distribution. For  $\alpha = .10$ , the risk function (10.2.4) for  $c_I = 8$  and  $c_{II} = 3$  is shown in Figure 10.2.2. Notice the discontinuity in the risk function at  $\theta = \theta_0$ . This is due to the fact that at  $\theta_0$  the expression in the risk function changes from  $\beta(\theta)$  to  $1 - \beta(\theta)$  as well as to the difference between  $c_I$  and  $c_{II}$ .

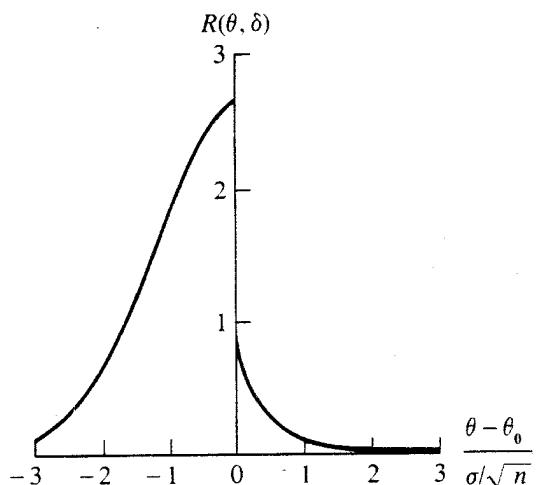


FIGURE 10.2.2 Risk function for test in Example 10.2.4

The 0–1 loss judges only whether a decision is right or wrong. It may be the case that some wrong decisions are more serious than others and the loss function should reflect this. When testing  $H_0: \theta \geq \theta_0$  versus  $H_1: \theta < \theta_0$ , it is a Type I Error to reject  $H_0$  if  $\theta$  is slightly bigger than  $\theta_0$ , but it may not be a very serious mistake. The adverse consequences of rejecting  $H_0$  may be much worse if  $\theta$  is much larger than  $\theta_0$ . A loss function that reflects this is

$$(10.2.5) \quad L(\theta, a_0) = \begin{cases} 0 & \theta \geq \theta_0 \\ b(\theta_0 - \theta) & \theta < \theta_0 \end{cases} \quad \text{and} \quad L(\theta, a_1) = \begin{cases} c(\theta - \theta_0)^2 & \theta \geq \theta_0 \\ 0 & \theta < \theta_0 \end{cases}$$

where  $b$  and  $c$  are positive constants. For example, if an experimenter is testing whether a drug lowers cholesterol level,  $H_0$  and  $H_1$  might be set up like this with  $\theta_0 =$  standard acceptable cholesterol level. Since a high cholesterol level is associated with heart disease, the consequences of rejecting  $H_0$  when  $\theta$  is large are quite serious. A loss function like (10.2.5) reflects such a consequence. A similar type of loss function is advocated by Vardeman (1987).

Even for a general loss function like (10.2.5), the risk function and the power function are closely related. For any fixed value of  $\theta$ , the loss is either  $L(\theta, a_0)$  or  $L(\theta, a_1)$ . Thus the expected loss is

$$(10.2.6) \quad \begin{aligned} R(\theta, \delta) &= L(\theta, a_0)P_\theta(\delta(\mathbf{X}) = a_0) + L(\theta, a_1)P_\theta(\delta(\mathbf{X}) = a_1) \\ &= L(\theta, a_0)(1 - \beta(\theta)) + L(\theta, a_1)\beta(\theta). \end{aligned}$$

The power function of a test is always important when evaluating a hypothesis test. But in a decision theoretic analysis, the weights given by the loss function are also important.

### 10.2.3 Interval Estimation

In an interval estimation problem, the action space  $\mathcal{A}$  will consist of subsets of the parameter space  $\Theta$ . Thus, more formally we should talk of “set estimation,” since an optimal rule may not necessarily be an interval. Following the strict rules of a decision theoretic development, we should not restrict the form of the answer to necessarily be an interval. However, practical considerations lead us to mainly consider set estimators that are intervals and, happily, many optimal procedures turn out to be intervals. Thus, if  $\Theta$  is the real line, the elements in  $\mathcal{A}$  are usually taken to be intervals and if  $\Theta \subset \mathbb{R}^p$ , then  $\mathcal{A}$  might be restricted to contain only squares, circles, ellipses, or rectangles.

We will use the more standard symbol  $C$  (for confidence interval) to denote elements of  $\mathcal{A}$ , with the meaning of the action  $C$  being that the interval estimate “ $\theta \in C$ ” is made. A decision rule  $\delta(\mathbf{x})$  simply specifies, for each  $\mathbf{x} \in \mathcal{X}$ , which set  $C \in \mathcal{A}$  will be used as an estimate of  $\theta$  if  $\mathbf{X} = \mathbf{x}$  is observed. Thus we will use the notation  $C(\mathbf{x})$ , as in Chapter 9.

The loss function in an interval estimation problem usually includes two quantities, a measure of whether the set estimate correctly includes the true value  $\theta$ , and a measure of the size of the set estimate. We will, for the most part, consider only sets  $C$  that are intervals, so a natural measure of size is  $\text{Len}(C) = \text{length of } C$ . To express the correctness measure, it is common to use

$$I_C(\theta) = \begin{cases} 1 & \theta \in C \\ 0 & \theta \notin C \end{cases}.$$

That is,  $I_C(\theta) = 1$  if the estimate is correct and zero otherwise. In fact,  $I_C(\theta)$  is just the indicator function for the set  $C$ . But realize that  $C$  will be a random set determined by the value of the data  $\mathbf{X}$ .

The loss function should reflect the fact that a good estimate would have  $\text{Len}(C)$  small and  $I_C(\theta)$  large. One such loss function is

$$(10.2.7) \quad L(\theta, C) = b \text{Len}(C) - I_C(\theta)$$

where  $b$  is a positive constant that reflects the relative weight that we want to give to the two criteria, a necessary consideration since the two quantities are very different. If there is more concern with correct estimates then  $b$  should be small, while a large  $b$  should be used if there is more concern with interval length.

The risk function associated with (10.2.7) is particularly simple, given by

$$\begin{aligned} R(\theta, C) &= bE_{\theta} [\text{Len}(C(X))] - E_{\theta} I_{C(X)}(\theta) \\ &= bE_{\theta} [\text{Len}(C(X))] - P_{\theta}(I_{C(X)}(\theta) = 1) \\ &= bE_{\theta} [\text{Len}(C(X))] - P_{\theta}(\theta \in C(X)). \end{aligned}$$

The risk has two components, the expected length of the interval and the coverage probability of the interval estimator. The risk reflects the fact that, simultaneously, we want the expected length to be small and the coverage probability to be high, just as in Chapter 9. But, unlike the development in Chapter 9, the specific functional tradeoff between these two quantities is now specified in the risk function. The approach in Chapter 9 was to fix the confidence coefficient, the infimum of the coverage probability, at  $(1 - \alpha)$  and try to minimize the length. In the present decision theoretic analysis, both quantities are viewed through the risk function. Perhaps a smaller coverage probability will be acceptable if it results in a greatly decreased length.

By varying the size of  $b$  in the loss (10.2.7), we can vary the relative importance of size and coverage probability of interval estimators, something that could not be done in Chapter 9. The situation here can be compared to that of finding an unrestricted maximum while the Chapter 9 development was most like finding a restricted maximum (finding the smallest confidence set subject to a coverage probability restriction). As an example of the flexibility of the present set up, consider some limiting cases. If  $b = 0$ , then size does not matter, only coverage probability, so the interval estimator  $C = (-\infty, \infty)$ , which has coverage probability 1, is the best decision rule. Similarly, if  $b = \infty$ , then coverage probability does not matter, so point sets are optimal. Hence, an entire range of decision rules are possible candidates. In the next example, for a specified finite range of  $b$ , choosing a good rule amounts to using the risk function to decide the confidence coefficient while, if  $b$  is outside this range, the optimal decision rule is a point estimator.

**Example 10.2.5:** Let  $X \sim n(\mu, \sigma^2)$  and assume  $\sigma^2$  is known.  $X$  would typically be a sample mean and  $\sigma^2$  would have the form  $\tau^2/n$  where  $\tau^2$  is the known population variance and  $n$  is the sample size. For each  $c \geq 0$ , define an interval estimator for  $\mu$  by  $C(x) = [x - c\sigma, x + c\sigma]$ . We will compare these estimators using the loss in (10.2.7). The length of an interval,  $\text{Len}(C(x)) = 2c\sigma$ , does not depend on  $x$ . Thus, the first term in the risk is  $b(2c\sigma)$ . The second term in the risk is

$$\begin{aligned} P_{\mu}(\mu \in C(X)) &= P_{\mu}(X - c\sigma \leq \mu \leq X + c\sigma) \\ &= P_{\mu}\left(-c \leq \frac{X - \mu}{\sigma} \leq c\right) \\ &= 2P(Z \leq c) - 1 \end{aligned}$$

where  $Z \sim n(0, 1)$ . Thus, the risk function for an interval estimator in this class is

$$(10.2.8) \quad R(\mu, C) = b(2c\sigma) - [2P(Z \leq c) - 1].$$

The risk function is constant, as it does not depend on  $\mu$ , and the best interval estimator in this class is the one corresponding to the value  $c$  that minimizes (10.2.8).

If  $b\sigma > 1/\sqrt{2\pi}$ , it can be shown that  $R(\mu, C)$  is minimized at  $c = 0$ . That is, the length portion completely overwhelms the coverage probability portion of the loss and the best *interval* estimator is the *point* estimator  $C(x) = [x, x]$ . But if  $b\sigma \leq 1/\sqrt{2\pi}$ , the risk is minimized at  $c = \sqrt{-2 \log(b\sigma\sqrt{2\pi})}$ . If we express  $c$  as  $z_{\alpha/2}$  for some  $\alpha$ , then the interval estimator that minimizes the risk is just the usual  $1 - \alpha$  confidence interval. (See Exercise 10.5 for details.)

It is important to understand the reasoning that led to the choice of this estimator. Here  $b$  is chosen to reflect the relative weight that is put on the two components of the risk and the best value of  $c$  determined. In Chapter 9, only one part of the risk, the confidence coefficient, was considered and specified first. Then only interval estimators with that confidence coefficient were considered. ||

The use of decision theory in interval estimation problems is not as widespread as in point estimation or hypothesis testing problems. One reason for this is the difficulty in choosing  $b$  in (10.2.7) (or in Example 10.2.5). We saw in the previous example that a choice that might seem reasonable could lead to unintuitive results, indicating that the loss in (10.2.7) may not be appropriate. Some who would use decision theoretic analysis for other problems still prefer to use only interval estimators with a fixed confidence coefficient  $(1 - \alpha)$ . They then use the risk function to judge other qualities like the size of the set.

Another difficulty is in the restriction of the shape of the allowable sets in  $\mathcal{A}$ . Ideally, the loss and risk functions would be used to judge which shapes are best. But one can always add isolated points to an interval estimator and get an improvement in coverage probability with no loss penalty regarding size. In the previous example we could have used the estimator

$$C(x) = [x - c\sigma, x + c\sigma] \cup \{\text{all integer values of } \mu\}.$$

The “length” of these sets is the same as before but now the coverage probability is one for all integer values of  $\mu$ . Some more sophisticated measure of size must be used to avoid such anomalies. (Joshi (1969) addressed this problem by defining equivalence classes of estimators.)

### 10.3 Decision Theoretic Bayes Rules

In a decision theoretic analysis, decision rules are to be compared through their risk functions. But because risk functions are *functions* of  $\theta$ , the comparison is not always clear-cut. The risk functions,  $R(\theta, \delta_1)$  and  $R(\theta, \delta_2)$ , may be such that  $R(\theta, \delta_1) > R(\theta, \delta_2)$  for some values of  $\theta \in \Theta$  but  $R(\theta, \delta_1) < R(\theta, \delta_2)$  for other values of  $\theta \in \Theta$ . So  $\delta_1$  has lower expected loss for some possible values of  $\theta$  but  $\delta_2$  has lower expected loss for other values. Since the true value of  $\theta$  is unknown, it is not clear which decision rule is better. One way to provide a clearer comparison is to summarize the

risk function in a single number. Then the decision rule with the smaller summary number would be preferred. In this section we discuss one such summary method and in Section 10.5 we discuss another.

### 10.3.1 Bayesian Decision Problems

Let  $\pi(\theta)$  denote the pdf or pmf of a probability distribution on the parameter space  $\Theta$ . Recall that this probability distribution is called the *prior distribution of  $\theta$* . To a *subjective Bayesian*, the prior reflects the beliefs of the experimenter about the value of  $\theta$  prior to data collection. To a *decision theoretic Bayesian*,  $\pi(\theta)$  is just a weight function. Values of  $\theta$  that are given high prior probability are values for which the experimenter would like to have particularly small risk, whereas values of  $\theta$  that are given lower prior probability are values for which the experimenter is not as concerned about the risk. But regardless of the interpretation, the risk function of a decision rule  $\delta$  is summarized by the *Bayes risk* defined as in (10.1.2) by  $B(\pi, \delta) = E_\pi R(\theta, \delta)$ . The subscript  $\pi$  indicates that the expectation is taken with  $\theta$  considered as a random variable with probability distribution  $\pi(\theta)$ . Thus for continuous and discrete probability distributions, the Bayes risks are

$$(10.3.1) \quad B(\pi, \delta) = \int_{\Theta} R(\theta, \delta) \pi(\theta) d\theta \quad \text{and} \quad B(\pi, \delta) = \sum_{\theta \in \Theta} R(\theta, \delta) \pi(\theta),$$

respectively. The risk function is summarized by the average risk,  $B(\pi, \delta)$ , and, according to this criterion, a rule with smaller Bayes risk is preferred to a rule with larger Bayes risk. The prior  $\pi(\theta)$  has been used to give more weight to more “important” values of  $\theta$  and less weight to others.

Note that this development of Bayes inference is different from our previous encounters with Bayes methods, as our development in Section 10.2 was different from previous classical solutions. Previously, a Bayes solution to a problem was complete once the posterior distribution was calculated. Point and interval estimators were obtained from the posterior in any way that the experimenter thought reasonable, rather than optimizing a formal quantity like that in (10.3.1). The decision theoretic Bayesian attempts to minimize the Bayes risk.

The *Bayes rule with respect to a prior  $\pi$*  is the decision rule  $\delta^\pi(x)$  that minimizes the Bayes risk among all possible decision rules. That is, the Bayes rule is defined to be the decision rule that satisfies

$$(10.3.2) \quad B(\pi, \delta^\pi) = \inf_{\delta \in \mathcal{D}} B(\pi, \delta).$$

There may be no  $\delta^\pi \in \mathcal{D}$  that satisfies (10.3.2) or there may be many decision rules that do. So there may be no Bayes rule or many. But, typically, there will be one Bayes rule. In such cases this criterion has succeeded in identifying one decision rule as the best. The Bayesian version of a decision problem is to find the rule  $\delta^\pi$  that minimizes  $B(\pi, \delta)$ .

### 10.3.2 Finding Bayes Rules

Finding the Bayes decision rule for a given prior  $\pi$  using definition (10.3.2) looks like a daunting task. We must find the function from  $\mathcal{X}$  into  $\mathcal{A}$ ,  $\delta^\pi(\mathbf{x})$ , that results in the function,  $R(\theta, \delta^\pi)$ , that minimizes the integral or sum in (10.3.1). But actually, finding the Bayes rule is a rather mechanical task, as the following theorem indicates. The technique of finding Bayes rules by the method given in Theorem 10.3.1 works in greater generality than presented here. A more general version is given in Brown and Purves (1973).

Throughout the rest of this section, we will use integrals to compute expectations, so for discrete distributions the integrals should be replaced by sums. Also, recall the definition of the *posterior distribution of  $\theta$  given the sample  $\mathbf{x}$* ,  $\pi(\theta|\mathbf{x})$ , given in (7.2.6).

**THEOREM 10.3.1:** For each  $\mathbf{x} \in \mathcal{X}$  and  $a \in \mathcal{A}$ , define

$$(10.3.3) \quad r(\mathbf{x}, a) = \int_{\Theta} L(\theta, a) \pi(\theta|\mathbf{x}) d\theta.$$

For each  $\mathbf{x} \in \mathcal{X}$ , suppose that there exists an  $a_x \in \mathcal{A}$  such that

$$(10.3.4) \quad r(\mathbf{x}, a_x) = \inf_{a \in \mathcal{A}} r(\mathbf{x}, a).$$

Let  $\delta^\pi$  be a function from  $\mathcal{X}$  into  $\mathcal{A}$  defined by  $\delta^\pi(\mathbf{x}) = a_x$ . If  $\delta^\pi \in \mathcal{D}$ , then  $\delta^\pi$  is the Bayes rule with respect to  $\pi$ .

*Proof:* From the definition of the posterior distribution, we know that  $\pi(\theta|\mathbf{x})m(\mathbf{x}) = f(\mathbf{x}|\theta)\pi(\theta)$  where  $m(\mathbf{x})$  is the marginal distribution of  $\mathbf{X}$ . The Bayes risk of a decision rule  $\delta$  can be written as

$$\begin{aligned}
 B(\pi, \delta) &= \int_{\Theta} R(\theta, \delta) \pi(\theta) d\theta \\
 &= \int_{\Theta} \left[ \int_{\mathcal{X}} L(\theta, \delta(\mathbf{x})) f(\mathbf{x}|\theta) d\mathbf{x} \right] \pi(\theta) d\theta \quad (\text{definition of } R(\theta, \delta)) \\
 &= \int_{\Theta} \int_{\mathcal{X}} L(\theta, \delta(\mathbf{x})) f(\mathbf{x}|\theta) \pi(\theta) d\mathbf{x} d\theta \quad (\text{move } \pi(\theta)) \\
 &= \int_{\Theta} \int_{\mathcal{X}} L(\theta, \delta(\mathbf{x})) \pi(\theta|\mathbf{x}) m(\mathbf{x}) d\mathbf{x} d\theta \quad \left( \begin{array}{c} f(\mathbf{x}|\theta)\pi(\theta) = \\ \pi(\theta|\mathbf{x})m(\mathbf{x}) \end{array} \right) \\
 &= \int_{\mathcal{X}} \int_{\Theta} L(\theta, \delta(\mathbf{x})) \pi(\theta|\mathbf{x}) m(\mathbf{x}) d\theta d\mathbf{x} \quad (\text{interchange integrals}) \\
 &= \int_{\mathcal{X}} \int_{\Theta} L(\theta, \delta(\mathbf{x})) \pi(\theta|\mathbf{x}) d\theta m(\mathbf{x}) d\mathbf{x} \quad (\text{move } m(\mathbf{x})) \\
 &= \int_{\mathcal{X}} r(\mathbf{x}, \delta(\mathbf{x})) m(\mathbf{x}) d\mathbf{x}. \quad (\text{expression (10.3.3)})
 \end{aligned} \tag{10.3.5}$$

But by (10.3.4) and the definition of  $\delta^\pi(x)$ , for every  $x \in \mathcal{X}$ ,  $r(x, \delta^\pi(x)) = r(x, a_x)$  is the smallest possible value. Thus  $\delta^\pi$  minimizes the integral (10.3.5) and, hence, the Bayes risk.  $\square$

Theorem 10.3.1 tells us exactly what the Bayes rule should do for each  $x \in \mathcal{X}$ . For a given observation  $x$ , the Bayes rule should take action  $a_x$  defined by (10.3.4). Hence, for each  $x$ , we need to perform the minimization indicated in (10.3.4) to determine the Bayes decision. This is quite unlike any prescription we have had in previous chapters. For example, consider the methods of finding best unbiased estimators discussed in Chapter 7. To use Theorem 7.3.5, first we need to find a complete sufficient statistic  $T$ . Then we need to find a function  $\phi(T)$  that is an unbiased estimator of the parameter. The Rao–Blackwell Theorem, Theorem 7.3.2, may be helpful if we know of some unbiased estimator of the parameter. But if we cannot dream up some unbiased estimator, then the method does not tell us how to construct one. Theorem 10.3.1 tells us exactly how to construct the Bayes rule, once we have decided on the prior distribution.

Even if the minimization in (10.3.4) cannot be done analytically, the integral can be evaluated and the minimization carried out numerically. In fact, having observed  $X = x$ , we need to solve (10.3.4) only for this particular  $x$ . We do not need to be concerned with what the value of  $\delta^\pi(x')$  is for any other  $x' \in \mathcal{X}$ . However, in many problems Theorem 10.3.1 can be used to explicitly describe the Bayes rule, as the next two theorems show.

**THEOREM 10.3.2:** Consider a point estimation problem for a real-valued parameter  $\theta$ . In each of the following two situations, if  $\delta^\pi \in \mathcal{D}$  then  $\delta^\pi$  is the Bayes rule (also called the *Bayes estimator*).

- a. For squared error loss,  $\delta^\pi(x) = E(\theta|x)$ .
- b. For absolute error loss,  $\delta^\pi(x) = \text{median of } \pi(\theta|x)$ .

*Proof:* For squared error loss,

$$\begin{aligned} r(x, a) &= \int_{\Theta} (\theta - a)^2 \pi(\theta|x) d\theta \\ &= E((\theta - a)^2 | X = x). \end{aligned}$$

Here  $\theta$  is the random variable with distribution  $\pi(\theta|x)$ . By Example 2.2.4, this expected value is minimized by  $a_x = E(\theta|x)$ . So, by Theorem 10.3.1, (a) is true. To prove (b), use Exercise 2.19 to see that  $E(|\theta - a| | X = x)$  is minimized by  $a_x = \text{median of } \pi(\theta|x)$ .  $\square$

In Chapter 7, the Bayes estimator we discussed was  $\delta^\pi(x) = E(\theta|x)$ , the posterior mean. We now see that this is the Bayes estimator with respect to squared error loss. If some other loss function is deemed more appropriate than squared error loss, the Bayes estimator might be a different statistic.

One way that the condition  $\delta^\pi \in \mathcal{D}$  in Theorem 10.3.2 might not be satisfied in a point estimation problem is if  $\Theta = \mathcal{A}$  is a finite set. Then  $E(\theta|x)$  might not be

in  $\mathcal{A}$  and, hence,  $\delta^\pi(\mathbf{x}) = E(\theta|\mathbf{x})$  would not be a function from  $\mathcal{X}$  into  $\mathcal{A}$  and  $\delta^\pi(\mathbf{x})$  would not be a legitimate estimator. If  $\Theta = \mathcal{A}$  is a convex set, it can be shown that  $E(\theta|\mathbf{x}) \in \mathcal{A}$  for any probability distribution  $\pi(\theta|\mathbf{x})$ . Then the only way that  $\delta^\pi \in \mathcal{D}$  would not be satisfied is if  $\mathcal{D}$  had been artificially restricted in some way so as not to include  $\delta^\pi$ . If the minimization in (10.3.4) can be carried out and if  $\mathcal{D}$  contains all functions from  $\mathcal{X}$  into  $\mathcal{A}$ , then  $\delta^\pi$  from Theorem 10.3.1 is the Bayes rule.

**Example 10.3.1:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population and let  $\pi(\theta)$  be  $n(\mu, \tau^2)$ . The values  $\sigma^2$ ,  $\mu$ , and  $\tau^2$  are known. In Example 7.2.10, as extended in Exercise 7.23, we found that the posterior distribution of  $\theta$  given  $\bar{X} = \bar{x}$  is normal with

$$E(\theta|\bar{x}) = \frac{\tau^2}{\tau^2 + (\sigma^2/n)} \bar{x} + \frac{\sigma^2/n}{\tau^2 + (\sigma^2/n)} \mu$$

and

$$\text{Var}(\theta|\bar{x}) = \frac{\tau^2 \sigma^2 / n}{\tau^2 + (\sigma^2/n)}.$$

For squared error loss, the Bayes estimator is  $\delta^\pi(\mathbf{x}) = E(\theta|\bar{x})$ . Since the posterior distribution is normal, it is symmetric about its mean and the median of  $\pi(\theta|\mathbf{x})$  is equal to  $E(\theta|\bar{x})$ . Thus, for absolute error loss, the Bayes estimator is also  $\delta^\pi(\mathbf{x}) = E(\theta|\bar{x})$ . ||

**Example 10.3.2:** Let  $X_1, \dots, X_n$  be iid Bernoulli( $p$ ) and let  $Y = \sum X_i$ . Suppose the prior on  $p$  is beta( $\alpha, \beta$ ). In Example 7.2.9 we found that the posterior distribution depends on the sample only through the observed value of  $Y = y$  and is beta  $(y + \alpha, n - y + \beta)$ . Hence,  $\delta^\pi(y) = E(p|y) = (y + \alpha)/(\alpha + \beta + n)$  is the Bayes estimator of  $p$  for squared error loss.

For absolute error loss, we need to find the median of  $\pi(p|y) = \text{beta}(y + \alpha, n - y + \beta)$ . In general, there is no simple expression for this median. The median is implicitly defined to be the number,  $m$ , that satisfies

$$\int_0^m \frac{\Gamma(\alpha + \beta + n)}{\Gamma(y + \alpha)\Gamma(n - y + \beta)} p^{y+\alpha-1} (1-p)^{n-y+\beta-1} dp = \frac{1}{2}.$$

This integral can be evaluated numerically to find (approximately) the value  $m$  that satisfies the equality. We have done this for  $n = 10$  and  $\alpha = \beta = 1$ , the uniform  $(0, 1)$  prior. The Bayes estimator for absolute error loss is given in Table 10.3.1. In the table we have also listed the Bayes estimator for squared error loss, derived above, and the MLE,  $\hat{p} = y/n$ .

Notice in Table 10.3.1 that, unlike the MLE, neither Bayes estimator estimates  $p$  to be 0 or 1, even if  $y$  is 0 or  $n$ . It is typical of Bayes estimators that they would not take on extreme values in the parameter space. No matter how large the sample size, the prior always has some influence on the estimator and tends to draw it away from the extreme values. In the above expression for  $E(p|y)$ , you can see that even if  $y = 0$  and  $n$  is large, the Bayes estimator is a positive number.

**Table 10.3.1** Three estimators for a binomial  $p$ 

$n = 10$		prior $\pi(p) \sim \text{uniform}(0, 1)$	
$y$	MLE	Bayes absolute error	Bayes squared error
0	.0000	.0611	.0833
1	.1000	.1480	.1667
2	.2000	.2358	.2500
3	.3000	.3238	.3333
4	.4000	.4119	.4167
5	.5000	.5000	.5000
6	.6000	.5881	.5833
7	.7000	.6762	.6667
8	.8000	.7642	.7500
9	.9000	.8520	.8333
10	1.0000	.9389	.9137

In Chapter 8 we considered some Bayesian hypothesis tests, and suggested that a Bayesian might compute the posterior probability,

$$P(\theta \in \Theta_0 | \mathbf{x}) = P(H_0 \text{ is true} | \mathbf{x}),$$

and reject  $H_0$  if this probability is too small. In the next theorem we see how this idea can be quantified in terms of the generalized 0–1 loss given in (10.2.3).

**THEOREM 10.3.3:** Consider a Bayesian hypothesis testing problem using generalized 0–1 loss. Any test of the form

$$\text{reject } H_0: \theta \in \Theta_0 \text{ if } P(\theta \in \Theta_0 | \mathbf{x}) < \frac{c_{II}}{c_I + c_{II}}$$

and

$$\text{accept } H_0: \theta \in \Theta_0 \text{ if } P(\theta \in \Theta_0 | \mathbf{x}) > \frac{c_{II}}{c_I + c_{II}}$$

is a Bayes rule (also called the Bayes test).

*Proof:* In a hypothesis testing problem,  $\mathcal{A} = \{a_0, a_1\}$  has only two elements. In this case application of Theorem 10.3.1 is particularly simple. The theorem says that the Bayes test compares  $r(\mathbf{x}, a_0)$  and  $r(\mathbf{x}, a_1)$  and takes the action  $a_i$  that gives the smaller value. If  $r(\mathbf{x}, a_0) = r(\mathbf{x}, a_1)$ , the Bayes test can take either action.

For generalized 0–1 loss,  $L(\theta, a_0)$  is zero if  $\theta \in \Theta_0$  and is  $c_{II}$  if  $\theta \in \Theta_0^c$ . Thus

$$r(\mathbf{x}, a_0) = \int_{\Theta} L(\theta, a_0) \pi(\theta | \mathbf{x}) d\theta = \int_{\Theta_0^c} c_{II} \pi(\theta | \mathbf{x}) d\theta = c_{II} P(\theta \in \Theta_0^c | \mathbf{x}).$$

In a similar way we find that  $r(\mathbf{x}, a_1) = c_I P(\theta \in \Theta_0 | \mathbf{x})$ . Thus the set of  $\mathbf{x}$  for which the Bayes test rejects  $H_0$ , that is, the set of  $\mathbf{x}$  for which  $r(\mathbf{x}, a_1) < r(\mathbf{x}, a_0)$ , is the set of  $\mathbf{x}$  satisfying

$$\{\mathbf{x} : c_I P(\theta \in \Theta_0 | \mathbf{x}) < c_{II} P(\theta \in \Theta_0^c | \mathbf{x})\}.$$

Since  $P(\theta \in \Theta_0^c | \mathbf{x}) = 1 - P(\theta \in \Theta_0 | \mathbf{x})$ , this set is equivalent to

$$\left\{ \mathbf{x} : P(\theta \in \Theta_0 | \mathbf{x}) < \frac{c_{II}}{c_I + c_{II}} \right\}.$$

In the same way, the inequality  $r(\mathbf{x}, a_1) > r(\mathbf{x}, a_0)$  is equivalent to  $P(\theta \in \Theta_0 | \mathbf{x}) > c_{II}/(c_I + c_{II})$ . A Bayes test accepts  $H_0$  on the set

$$\left\{ \mathbf{x} : P(\theta \in \Theta_0 | \mathbf{x}) > \frac{c_{II}}{c_I + c_{II}} \right\}. \quad \square$$

**Example 10.3.3:** Consider again the model from Example 10.3.1. Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population and let  $\pi(\theta)$  be  $n(\mu, \tau^2)$ . The values  $\sigma^2$ ,  $\mu$ , and  $\tau^2$  are known. To simplify notation, let

$$\eta = \frac{\sigma^2}{n\tau^2 + \sigma^2}.$$

The posterior distribution of  $\theta$  given  $\bar{X} = \bar{x}$  is normal with

$$E(\theta | \bar{x}) = (1 - \eta)\bar{x} + \eta\mu \quad \text{and} \quad \text{Var}(\theta | \bar{x}) = \tau^2\eta.$$

Consider testing  $H_0: \theta \geq \theta_0$  versus  $H_1: \theta < \theta_0$  using generalized 0–1 loss. Then

$$P(\theta \in \Theta_0 | \mathbf{x}) = P(\theta \geq \theta_0 | \bar{x}) = P\left(Z \geq \frac{\theta_0 - (1 - \eta)\bar{x} - \eta\mu}{\tau\sqrt{\eta}} \middle| \bar{x}\right)$$

where  $Z \sim n(0, 1)$ . Let  $\alpha = c_{II}/(c_I + c_{II})$ . Then

$$P(\theta \in \Theta_0 | \mathbf{x}) < \frac{c_{II}}{c_I + c_{II}} = \alpha$$

is true if and only if

$$\frac{\theta_0 - (1 - \eta)\bar{x} - \eta\mu}{\tau\sqrt{\eta}} > z_\alpha.$$

This inequality is equivalent to

$$\theta_0 - \frac{\eta(\mu - \theta_0) - z_\alpha \tau \sqrt{\eta}}{1 - \eta} > \bar{x}.$$

Thus the Bayes test rejects for all small values of  $\bar{X}$ , where the cutoff point depends on the losses for the two types of errors, as measured by  $\alpha$ , and the prior distribution.

For one situation, the cutoff point is particularly simple. Suppose both types of errors are judged to be equally bad so that  $c_I = c_{II}$ ; then  $z_\alpha = z_{.5} = 0$ . Further, suppose that  $\mu = \theta_0$ . This prior is centered at  $\theta_0$  and assigns equal weight to  $\Theta_0$  and  $\Theta_0^c$ . Then the cutoff point is  $\theta_0$ , that is, the Bayes test rejects  $H_0$  if  $\bar{X} > \theta_0$  and accepts  $H_0$  if  $\bar{X} \leq \theta_0$ .

In Example 8.3.7, we saw that the uniformly most powerful level  $\alpha'$  test of  $H_0$  versus  $H_1$  rejected  $H_0$  if

$$\theta_0 - z_{\alpha'} \frac{\sigma}{\sqrt{n}} > \bar{X}.$$

Here too the test rejects for all small values of  $\bar{X}$ . But for this test the cutoff point is determined from the specification of the Type I Error probability. ||

## 10.4 Admissibility of Decision Rules

In a decision problem, the class  $\mathcal{D}$  of allowable decision rules is usually very large. It may be all functions from  $\mathcal{X}$  into  $\mathcal{A}$ . Choosing which  $\delta \in \mathcal{D}$  to use may be a difficult task, but the task would be easier if a subclass  $\mathcal{C}$  of  $\mathcal{D}$  could be found that contained all the good rules. Since this new class is a smaller class, choosing a rule would be an easier task. In this section we describe how some classes of good rules are defined using the criterion of *admissibility* and investigate how rules in these classes might be identified.

Rigorous treatment of these ideas sometimes becomes rather difficult mathematically. Rather than getting involved in these details, at times we will only refer to more profound results.

### 10.4.1 Comparing Decision Rules

Various comparisons can be made between decision rules using their risk functions. These comparisons are formalized in the following definitions.

**DEFINITION 10.4.1:** Let  $\delta$  and  $\delta'$  be two decision rules with risk functions  $R(\theta, \delta)$  and  $R(\theta, \delta')$ , respectively.  $\delta$  is *as good as*  $\delta'$  if  $R(\theta, \delta) \leq R(\theta, \delta')$  for all  $\theta \in \Theta$ .  $\delta$  is *better than*  $\delta'$  if  $R(\theta, \delta) \leq R(\theta, \delta')$  for all  $\theta \in \Theta$  and  $R(\theta, \delta) < R(\theta, \delta')$  for some  $\theta \in \Theta$ .  $\delta$  is *equivalent to*  $\delta'$  if  $R(\theta, \delta) = R(\theta, \delta')$  for all  $\theta \in \Theta$ .

Given that we agree to compare decision rules only through their risk functions, then the comparisons described in Definition 10.4.1 are of a most fundamental nature. We have already noticed these kinds of relationships in some examples. In Example 10.2.2 we found that for estimating a normal variance with squared error loss, the estimator  $\tilde{S}^2 = \frac{n-1}{n+1} S^2$  is better than any other estimator of the form  $bS^2$ . We came to this conclusion because the risk function for  $\tilde{S}^2$  is everywhere smaller than any of

the other risk functions. It would also be correct to say that  $\tilde{S}^2$  is as good as any other estimator of the form  $bS^2$ , although this would not be as strong an assertion.

The relationships in Definition 10.4.1 do not provide a comparison between all pairs of decision rules; rather they define a *partial ordering* on  $\mathcal{D}$ . If the risk functions of two decision rules cross, then none of the terms in Definition 10.4.1 apply. For example, none of these terms apply to the estimators  $\hat{p}_B$  and  $\bar{X}$  of a Bernoulli parameter  $p$  that were examined in Example 10.2.1.

**DEFINITION 10.4.2:** A decision rule  $\delta$  is *admissible* if there is no  $\delta' \in \mathcal{D}$  that is better than  $\delta$ . A decision rule  $\delta$  is *inadmissible* if there is a  $\delta' \in \mathcal{D}$  that is better than  $\delta$ .

An admissible decision rule is good in that there is no other decision rule that is clearly superior to it. An inadmissible decision rule is bad in that there is some other decision rule that is clearly superior. Notice, however, that admissibility is not really a positive property, but rather the absence of a negative property. That is, an admissible estimator is not necessarily uniformly good, but it is not uniformly bad! A little reflection will make it clear that, although we want to restrict attention to admissible rules, we should not immediately equate admissibility with desirability.

In general, we would not want to use an inadmissible decision rule, knowing that something better exists. However, the situation is complicated somewhat by the fact that in some (rather artificial) problems, *every* decision rule is inadmissible! Furthermore, even if a rule is inadmissible, it may be close to admissible and we may have a hard time finding a rule that beats it. (A realistic example of this is the positive-part estimator discussed in Section 10.7.)

Knowing that a decision rule is admissible does not mean that this decision rule is obviously the rule to use since, in most cases, there are many admissible decision rules. Some of these rules will be reasonable, some will be difficult to find or compute, and some may not be intuitively appealing. The next example illustrates this.

**Example 10.4.1:** Let  $X$  have a binomial( $n, p$ ) distribution, where the sample size  $n$  is known. Consider estimating  $p$  using absolute error loss, and suppose that we use the estimator  $\delta(x) = \frac{1}{3}$  for  $x = 0, \dots, n$ . This estimator ignores the data, always estimates  $p$  to be  $\frac{1}{3}$ , and is admissible. The risk for  $\delta$  at  $p = \frac{1}{3}$  is

$$\begin{aligned} R\left(\frac{1}{3}, \delta\right) &= \sum_{x=0}^n \left| \delta(x) - \frac{1}{3} \right| P\left(X = x | p = \frac{1}{3}\right) \\ &= \sum_{x=0}^n \left| \frac{1}{3} - \frac{1}{3} \right| P\left(X = x | p = \frac{1}{3}\right) = 0. \end{aligned}$$

Let  $\delta'$  be an estimator that is as good as  $\delta$ . Then it must be the case that

$$\sum_{x=0}^n \left| \delta'(x) - \frac{1}{3} \right| P\left(X = x | p = \frac{1}{3}\right) = R\left(\frac{1}{3}, \delta'\right) \leq R\left(\frac{1}{3}, \delta\right) = 0.$$

Since  $P(X = x|p = \frac{1}{3}) > 0$  for all  $x = 0, \dots, n$ ,  $\delta'(x)$  must equal  $\frac{1}{3}$  for all  $x = 0, \dots, n$ . That is, the only estimator that is as good as  $\delta$  is  $\delta$  itself. There is no  $\delta' \in \mathcal{D}$  that is better than  $\delta$  and, hence,  $\delta$  is admissible.

The estimator  $\delta$  is admissible only because it performs well if  $p = \frac{1}{3}$ . The risk of  $\delta$  will be large compared to the risk of other estimators at values of  $p \neq \frac{1}{3}$ , as Exercise 10.3 shows. ||

**Example 10.4.2:** Consider again the problem of estimating a population variance  $\sigma^2$  using squared error loss based on a random sample  $X_1, \dots, X_n$  from a  $n(\mu, \sigma^2)$  population. Unlike in Example 10.2.2 where we considered only estimators of the form  $bS^2$ , now let  $\mathcal{D}$  be the class of all estimators of  $\sigma^2$ . In Example 10.2.2, we showed that if  $b \neq \frac{n-1}{n+1}$ , then the estimator  $\tilde{S}^2 = \frac{n-1}{n+1}S^2$  is better than the estimator  $\delta_b(\mathbf{X}) = bS^2$ . Hence,  $\delta_b$  is inadmissible. However, from the analysis in Example 10.2.2, we cannot say whether  $\tilde{S}^2$  is admissible or inadmissible. There may be an estimator of some other form that is better than  $\tilde{S}^2$ . (Brewster and Zidek (1974) find an estimator better than  $\tilde{S}^2$ ). ||

Using the comparisons in Definition 10.4.1, we can define the following classes of decision rules. These classes contain, in a sense, all good decision rules. By restricting attention to rules in these classes we are not overlooking any obviously desirable decision rules.

**DEFINITION 10.4.3:** Let  $\mathcal{C}$  be a class of decision rules that is a subclass of  $\mathcal{D}$ .  $\mathcal{C}$  is a *complete class* if for any  $\delta' \notin \mathcal{C}$  there is a  $\delta \in \mathcal{C}$  such that  $\delta$  is better than  $\delta'$ .  $\mathcal{C}$  is an *essentially complete class* if for any  $\delta' \notin \mathcal{C}$  there is a  $\delta \in \mathcal{C}$  such that  $\delta$  is as good as  $\delta'$ .

The following theorem gives the general relationship between admissible decision rules and complete classes. A similar result for essentially complete classes is given in Exercise 10.20.

**THEOREM 10.4.1:** If  $\mathcal{C}$  is a complete class of decision rules, then the class of admissible decision rules is contained in  $\mathcal{C}$ .

*Proof:* Let  $\delta'$  be an admissible decision rule. If  $\delta' \notin \mathcal{C}$  then, from the definition of complete class, there is a  $\delta \in \mathcal{C}$  such that  $\delta$  is better than  $\delta'$ . Since  $\delta'$  is admissible, no such  $\delta$  exists. Therefore,  $\delta' \in \mathcal{C}$ . □

Theorem 10.4.1 gives some description of what rules should be in a complete class. To be useful, we must be able to describe complete classes more completely. Some examples of how this may be done are given in the next subsection.

## 10.4.2 Finding Admissible Rules and Complete Classes

In this section we give some examples of how admissible rules and complete classes of rules can be characterized. The first result concerns the admissibility of rules that are Bayes rules using priors that do not exclude any values of  $\theta$ .

**THEOREM 10.4.2:** Consider a decision problem in which the parameter space  $\Theta$  is a subset of the real line. Suppose that for every decision rule  $\delta \in \mathcal{D}$ , the risk function  $R(\theta, \delta)$  is a continuous function of  $\theta$ . Let  $\pi(\theta)$  be a prior distribution on  $\theta$  with the property that for any  $\epsilon > 0$  and any  $\theta \in \Theta$ , the interval  $(\theta - \epsilon, \theta + \epsilon)$  has positive probability under  $\pi$ . Let  $\delta^\pi$  be a Bayes rule with respect to  $\pi$ . If  $-\infty < B(\pi, \delta^\pi) < \infty$ , then  $\delta^\pi$  is an admissible decision rule.

*Proof:* Suppose that  $\delta^\pi$  is inadmissible. Then there exists a rule  $\delta \in \mathcal{D}$  such that  $R(\theta, \delta) \leq R(\theta, \delta^\pi)$  for all  $\theta \in \Theta$  and for some  $\theta$ , say  $\theta'$ ,  $R(\theta', \delta) < R(\theta', \delta^\pi)$ . Let  $R(\theta', \delta^\pi) - R(\theta', \delta) = \nu > 0$ . Since  $R(\theta, \delta^\pi)$  and  $R(\theta, \delta)$  are both continuous, so is  $R(\theta, \delta^\pi) - R(\theta, \delta)$ . Thus there exists an  $\epsilon > 0$  such that

$$(10.4.1) \quad R(\theta, \delta^\pi) - R(\theta, \delta) > \frac{\nu}{2} \quad \text{for all } \theta \in (\theta' - \epsilon, \theta' + \epsilon).$$

Since  $-\infty < B(\pi, \delta^\pi) < \infty$ , the following expression is well defined (not of the form  $\infty - \infty$ ):

$$\begin{aligned} B(\pi, \delta^\pi) - B(\pi, \delta) &= \int_{-\infty}^{\infty} R(\theta, \delta^\pi) \pi(\theta) d\theta - \int_{-\infty}^{\infty} R(\theta, \delta) \pi(\theta) d\theta \\ &= \int_{-\infty}^{\infty} [R(\theta, \delta^\pi) - R(\theta, \delta)] \pi(\theta) d\theta \\ &\geq \int_{\theta' - \epsilon}^{\theta' + \epsilon} [R(\theta, \delta^\pi) - R(\theta, \delta)] \pi(\theta) d\theta \quad \left( \begin{array}{l} R(\theta, \delta^\pi) - R(\theta, \delta) \geq 0 \\ \text{for all } \theta, \text{ by assumption} \end{array} \right) \\ &\geq \frac{\nu}{2} \int_{\theta' - \epsilon}^{\theta' + \epsilon} \pi(\theta) d\theta \quad (\text{from 10.4.1}) \\ &> 0. \quad \left( \begin{array}{l} (\theta' - \epsilon, \theta' + \epsilon) \text{ has positive} \\ \text{probability under } \pi \end{array} \right) \end{aligned}$$

This strict inequality contradicts the fact that  $\delta^\pi$  is Bayes with respect to  $\pi$ . Hence  $\delta^\pi$  is admissible.  $\square$

This theorem gives one set of conditions under which Bayes rules are admissible. There are many other situations in which this is true. Exercises 10.14 and 10.15 give two more such situations.

Although not true in every instance, the general idea is that Bayes rules are admissible and, hence, Bayes rules are reasonable rules to consider. By Theorem 10.4.1, admissible rules are contained in any complete class. Thus, Bayes rules are usually contained in a complete class. It would be nice if the set of all Bayes rules formed a complete class because the structure of Bayes rules is explicitly defined by Theorem 10.3.1. Often, however, certain "limits" of Bayes rules must be added to the set of Bayes rules to form a complete class. Theorems of this sort can be found, for example, in the classic book by Wald (1950) or in the more recent book by Berger (1985).

The assumption in Theorem 10.4.2 that *all* risk functions are continuous is a bit hard to verify. General books on decision theory, such as Berger (1985) and Ferguson (1967), give conditions under which this is true. The conditions usually include that  $L(\theta, a)$  is continuous in  $\theta$  for each  $a \in \mathcal{A}$  and that  $f(\mathbf{x}|\theta)$  is continuous in  $\theta$  for each  $\mathbf{x} \in \mathcal{X}$ . Then additional conditions are put on the loss function or on the pdfs or pmfs to obtain the desired result.

The concept of a sufficient statistic plays an important role in a decision theoretic analysis, just as in earlier chapters. The idea remains the same. A sufficient statistic  $T(\mathbf{X})$  contains all the information about the parameter  $\theta$  that is in the sample  $\mathbf{X}$ . Thus, we would expect that, when making a decision, we need consider only the value of  $T(\mathbf{X})$ , not the actual value of  $\mathbf{X}$  that is observed. The following theorem is a generalization of the Rao–Blackwell Theorem and gives conditions under which restriction to decision rules that are based on a sufficient statistic does not cost us anything in terms of risk.

This result is true in much more generality if we consider *randomized decision rules* (see the *Miscellanea* sections of Chapters 8 and 9). The randomized hypothesis tests discussed in Chapter 8 are examples of randomized decision rules, but this theorem deals with the nonrandomized decision rules that we have been discussing.

**THEOREM 10.4.3:** Let the action space  $\mathcal{A}$  be an interval (possibly infinite) of the real line, and suppose that the loss function  $L(\theta, a)$  is a convex function of the action  $a$  for each  $\theta \in \Theta$ . Suppose that  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$  with sample space  $\mathcal{T}$ . If  $\delta(\mathbf{x}) \in \mathcal{D}$  is any decision rule, then the decision rule based only on  $T$  defined by

$$(10.4.2) \quad \delta'(t) = E(\delta(\mathbf{X})|T = t)$$

is a decision rule that is as good as  $\delta$ , provided the expectation exists for every  $t \in \mathcal{T}$ .

*Proof:* First we need to verify that  $\delta'$  is a decision rule, that is,  $\delta'$  defines a function from  $\mathcal{X}$  into  $\mathcal{A}$ . Since  $T$  is sufficient, the expectation in (10.4.2) does not depend on  $\theta$ . Since  $\delta$  is a decision rule,  $\delta(\mathbf{x}) \in \mathcal{A}$  for every  $\mathbf{x} \in \mathcal{X}$ . By Theorem 2.2.1, the expectation in (10.4.2) is also in the interval  $\mathcal{A}$ . Thus,  $\delta'(T(\mathbf{x}))$  does define a function from  $\mathcal{X}$  into  $\mathcal{A}$ .

Now it remains to verify that  $\delta'$  is as good as  $\delta$ . In the following, expectations that are not subscripted with  $\theta$  do not depend on  $\theta$  because of the sufficiency of  $T$ . For any  $\theta \in \Theta$  we have

$$\begin{aligned} R(\theta, \delta) &= E_\theta(L(\theta, \delta(\mathbf{X}))) \\ &= E_\theta E(L(\theta, \delta(\mathbf{X}))|T) && \text{(iterated expectations)} \\ &\geq E_\theta L(\theta, E(\delta(\mathbf{X})|T)) && \text{(Jensen's Inequality,)} \\ &= E_\theta L(\theta, \delta'(T)) \\ &= R(\theta, \delta'). \end{aligned}$$

Since this inequality is true for every  $\theta \in \Theta$ ,  $\delta'$  is as good as  $\delta$ . □

Assuming that the expectation in (10.4.2) defining  $\delta'$  exists for every  $\delta \in \mathcal{D}$ , this theorem says that the class of decision rules that depend on  $X$  only through the sufficient statistic  $T$  is an essentially complete class. Theorem 10.4.3, with  $\mathcal{A}$  an interval, easily applies to point estimation problems.

The reason why randomized rules could be avoided in Theorem 10.4.3 was due to the convexity of the loss function. When the loss function is convex and the action space is convex, it never pays to randomize in the sense that a nonrandomized rule can always be found that is as good as any randomized rule. (See Exercise 10.17.) In the previous chapters we needed to be somewhat concerned about randomized rules because we did not want to make the assumption that the loss was convex. Although this is not that restrictive in point estimation problems, losses for interval estimation and hypothesis testing are not usually convex (see Exercise 10.18).

The next theorem is an example of a *complete class theorem* for a hypothesis testing problem. It gives conditions under which the tests defined in the Neyman–Pearson Lemma (Theorem 8.3.1) constitute an essentially complete class.

**THEOREM 10.4.4:** Consider testing  $H_0: \theta = \theta_0$  versus  $H_1: \theta = \theta_1$  using any loss that satisfies  $L(\theta_1, a_1) \leq L(\theta_1, a_0)$ . Let  $f(x|\theta_i), i = 0, 1$ , denote the pdf. Suppose  $f(x|\theta_0) > 0$  for every  $x \in \mathcal{X}$ . Suppose that for any  $k \geq 0$ ,

$$(10.4.3) \quad P_{\theta_0}(f(X|\theta_1) = kf(X|\theta_0)) = 0.$$

Let  $\delta_k$  be the test with rejection region  $R_k = \{x : f(x|\theta_1) > kf(x|\theta_0)\}$ . Then the class of tests  $\{\delta_k : 0 \leq k \leq \infty\}$  is an essentially complete class.

*Proof:* Let  $\delta$  be a test with rejection region  $R$  and power function  $\beta(\theta)$ . Define the random variable  $Y = f(X|\theta_1)/f(X|\theta_0)$ . Since  $f(x|\theta_0)$  is never zero,  $Y$  is a well-defined, nonnegative random variable. If  $\theta = \theta_0$ , by (10.4.3)  $Y$  is a continuous random variable with cdf  $F_Y(y)$  a continuous function. Thus, there exists a  $k \geq 0$  such that  $F_Y(k) = 1 - \beta(\theta_0)$  ( $k = \infty$  can be used if  $\beta(\theta_0) = 0$ ). This value of  $k$  can be used to define a rule  $\delta_k$ , which we will show is as good as  $\delta$ .

Let  $\beta_k(\theta)$  denote the power function of  $\delta_k$ . The tests  $\delta$  and  $\delta_k$  have the same size since

$$\beta_k(\theta_0) = P_{\theta_0}(f(X|\theta_1) > kf(X|\theta_0)) = P_{\theta_0}(Y > k) = 1 - F_Y(k) = \beta(\theta_0).$$

Using (10.2.6) we have

$$(10.4.4) \quad \begin{aligned} R(\theta_0, \delta_k) &= L(\theta_0, a_0)(1 - \beta_k(\theta_0)) + L(\theta_0, a_1)\beta_k(\theta_0) \\ &= L(\theta_0, a_0)(1 - \beta(\theta_0)) + L(\theta_0, a_1)\beta(\theta_0) \\ &= R(\theta_0, \delta). \end{aligned}$$

Since  $\delta$  and  $\delta_k$  have the same size and  $\delta_k$  is a test of the form specified in the Neyman–Pearson Lemma,  $\delta_k$  is at least as powerful as  $\delta$  at  $\theta_1$ , that is,  $\beta_k(\theta_1) \geq \beta(\theta_1)$ . Thus we have

$$\begin{aligned}
 R(\theta_1, \delta_k) &= L(\theta_1, a_0)(1 - \beta_k(\theta_1)) + L(\theta_1, a_1)\beta_k(\theta_1) \\
 &= (L(\theta_1, a_1) - L(\theta_1, a_0))\beta_k(\theta_1) + L(\theta_1, a_0) \\
 (10.4.5) \quad &\leq (L(\theta_1, a_1) - L(\theta_1, a_0))\beta(\theta_1) + L(\theta_1, a_0) \quad \left( \begin{array}{l} L(\theta_1, a_1) \leq L(\theta_1, a_0) \\ \text{and } \beta_k(\theta_1) \geq \beta(\theta_1) \end{array} \right) \\
 &= R(\theta_1, \delta).
 \end{aligned}$$

From (10.4.4) and (10.4.5), we see that  $\delta_k$  is as good as  $\delta$ . Since  $\delta$  was arbitrary, the class of tests  $\delta_k$ ,  $0 \leq k \leq \infty$ , is an essentially complete class.  $\square$

More general theorems like this, for one-sided testing problems when the family of pdfs or pmfs  $f(x|\theta)$  has an MLR, can be found in Berger (1985).

### 10.4.3 Admissibility of the Sample Mean Under Normality

In this section we consider the special problem of estimating the mean of a normal population using squared error loss. Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population and consider estimating  $\theta$  with the sample mean  $\bar{X}$ . We will show that  $\bar{X}$  is an admissible estimator of  $\theta$ . This is an important result because the normal model is widely used and the sample mean is almost always used to estimate the population mean. Furthermore, inadmissibility of  $\bar{X}$  would somewhat shake our intuition, since we somehow expect  $\bar{X}$  to be an admissible estimator of  $\theta$ . We do not expect it to be uniformly dominated. The general technique of proof is also of interest in that it is applicable in other problems. Blyth (1951) was among the first to use the technique which has come to be known as Blyth's method. Finally, the result is interesting because, as we shall see in Section 10.7, it is not as obvious as it may first seem.

For now, let  $\sigma$  be fixed. In the end we will show that  $\bar{X}$  is an admissible estimator of  $\theta$  when  $\sigma$  is unknown, but first we will prove the result for known  $\sigma$ . In Chapter 7 we showed that  $\bar{X}$  is the uniformly minimum variance unbiased estimator of  $\theta$ , with risk given by

$$(10.4.6) \quad R(\theta, \bar{X}) = E_\theta (\theta - \bar{X})^2 = \frac{\sigma^2}{n}.$$

We also saw that  $\bar{X}$  is the MLE of  $\theta$  (which does not imply any optimality in itself).  $\bar{X}$  is not a Bayes estimator for any prior but it is the limit of Bayes estimators.

If the prior distribution for  $\theta$  is  $\pi(\theta) \sim n(\mu, \tau^2)$ , the Bayes estimator is

$$\delta^\pi(x) = \frac{n\tau^2}{n\tau^2 + \sigma^2} \bar{x} + \frac{\sigma^2}{n\tau^2 + \sigma^2} \mu.$$

In the limit, as  $\tau^2 \rightarrow \infty$ , it can be seen that  $\delta^\pi(x) \rightarrow \bar{x}$ . Given the comments after Theorem 10.4.2, it seems reasonable to question whether  $\bar{X}$  is an admissible estimator of  $\theta$ . We will now show that it is. Some of the calculations that we need of risk functions and Bayes risks are outlined in Exercise 10.8.

To simplify notation, again let

$$\eta = \frac{\sigma^2}{n\tau^2 + \sigma^2}$$

and suppose that  $\bar{X}$  is inadmissible, that is, that there exists an estimator  $\delta'(\mathbf{x})$  such that  $R(\theta, \delta') \leq R(\theta, \bar{X})$  for all  $\theta$  and  $R(\theta', \delta') < R(\theta', \bar{X})$  for some  $\theta'$ . Since the sample pdf is normal, the risk function for any estimator in this problem is continuous. Thus, we can use an argument similar to that in the proof of Theorem 10.4.2. If we let  $\pi(\theta)$  be a  $n(0, \tau^2)$  prior on  $\theta$ , we can find an  $\epsilon > 0$  and  $\nu > 0$  such that

$$(10.4.7) \quad \tau(B(\pi, \bar{X}) - B(\pi, \delta')) \geq \frac{\nu/2}{\sqrt{2\pi}} \int_{\theta' - \epsilon}^{\theta' + \epsilon} e^{-\theta^2/(2\tau^2)} d\theta.$$

Since the risk function of  $\bar{X}$  is constant with value  $\sigma^2/n$ , the Bayes risk of  $\bar{X}$  for any prior is also  $\sigma^2/n$ . The Bayes risk of  $\delta^\pi$ , which is the posterior variance, is  $B(\pi, \delta^\pi) = \tau^2 \eta$ . Thus,

$$(10.4.8) \quad \tau(B(\pi, \delta^\pi) - B(\pi, \bar{X})) = \tau\left(\tau^2 \eta - \frac{\sigma^2}{n}\right) = -\frac{\sigma^2}{n} \tau \eta. \quad \begin{matrix} \text{(definition)} \\ \text{of } \eta \end{matrix}$$

Combining (10.4.7) and (10.4.8), we have

$$(10.4.9) \quad \begin{aligned} 0 &\geq \tau(B(\pi, \delta^\pi) - B(\pi, \delta')) && \text{(since } \delta^\pi \text{ is Bayes)} \\ &= \tau(B(\pi, \delta^\pi) - B(\pi, \bar{X})) + \tau(B(\pi, \bar{X}) - B(\pi, \delta')) && (\pm B(\pi, \bar{X})) \\ &\geq -\frac{\sigma^2}{n} \tau \eta + \frac{\nu/2}{\sqrt{2\pi}} \int_{\theta' - \epsilon}^{\theta' + \epsilon} e^{-\theta^2/(2\tau^2)} d\theta \end{aligned}$$

The inequality in (10.4.9) is true for every  $\tau > 0$ . Taking the limit as  $\tau \rightarrow \infty$ , we see that  $\tau \eta \rightarrow 0$ , the integrand in the last term  $\rightarrow 1$ , and thus the last term  $\rightarrow \nu \epsilon / \sqrt{2\pi}$ . Therefore, as  $\tau \rightarrow \infty$ , the last line in (10.4.9) converges to a positive number, giving a contradiction which implies that  $\bar{X}$  is admissible, if  $\sigma$  is known.

The estimator  $\bar{X}$  is also admissible in the unknown  $\sigma$  model. If  $\bar{X}$  is inadmissible in this case, there is an estimator  $\delta'$  such that

$$R((\theta, \sigma), \delta') \leq R((\theta, \sigma), \bar{X}) \quad \text{for all } (\theta, \sigma)$$

and

$$R((\theta', \sigma'), \delta') < R((\theta', \sigma'), \bar{X}) \quad \text{for some } (\theta', \sigma').$$

The estimator  $\delta'$  is a function from  $\mathcal{X}$  into the real line and, hence, may be used as an estimator of  $\theta$  for the model in which  $\sigma$  is fixed to be the value  $\sigma'$ . But for any estimator, the risk in the known  $\sigma'$  case is the same as the risk in the unknown  $\sigma$  case evaluated at  $\sigma = \sigma'$ . That is,  $R(\theta, \delta) = R((\theta, \sigma'), \delta)$ . Thus,

$$R(\theta, \delta') = R((\theta, \sigma'), \delta') \leq R((\theta, \sigma'), \bar{X}) = R(\theta, \bar{X}) \quad \text{for all } \theta$$

and

$$R(\theta', \delta') = R((\theta', \sigma'), \delta') < R((\theta', \sigma'), \bar{X}) = R(\theta', \bar{X}).$$

These inequalities imply that  $\delta'$  is better than  $\bar{X}$  for estimating  $\theta$  when  $\sigma = \sigma'$  is known. This is a contradiction, since  $\bar{X}$  is admissible in the known  $\sigma$  problem. Hence,  $\bar{X}$  is also an admissible estimator of  $\theta$  in the unknown  $\sigma$  model.

In Section 10.7 we will see that in a seemingly very similar problem, in which  $\theta$  is a vector, an estimator analogous to  $\bar{X}$  is not admissible.

## 10.5 Minimax Rules

In Section 10.3, the information in a risk function was summarized using the Bayes risk. The Bayes risk is a measure of the average risk (average defined in terms of the prior  $\pi$ ) associated with a decision rule. An advantage of using the Bayes risk is that the summary was in terms of a single number and any two decision rules could be compared by comparing their respective Bayes risks. In this section we investigate another commonly used summary, the maximum (or supremum) value of the risk function. The use of this measure leads to the following definition.

**DEFINITION 10.5.1:** A decision rule  $\delta'$  is called a *minimax decision rule* if

$$\sup_{\theta \in \Theta} R(\theta, \delta') = \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta).$$

A minimax decision rule has the smallest possible maximum risk, so the maximum risk for any other decision rule is at least as big as the maximum for a minimax decision rule. For each decision rule, the minimax criterion looks at the “worst” value of  $\theta$  that could be true, and guards against this worst case. (The word *minimax* is derived from the above equality. If “inf” is replaced with *min* and “sup” is replaced with *max*, we have something that approaches the word *minimax*.)

Some find the minimax criterion too conservative, since it judges Nature, the chooser of  $\theta$ , as an adversary of the statistician. For each decision rule, the minimax criterion evaluates the worst thing that Nature could do to the statistician. The minimax rule guards against this worst choice. In a statistical problem, however, Nature is not usually considered to be an adversary. The value of  $\theta$  is considered to be a fixed, unknown value but not a value deliberately chosen to make the risk large. Thus, according to this view, the use of a criterion that guards against such a value is unnecessarily conservative. Fortunately, although minimax rules come from overly conservative considerations, in many situations they turn out to be reasonable rules in practice. For example, the sample mean,  $\bar{X}$ , is minimax in many situations and best invariant rules (Section 10.6) often turn out to be minimax. (See the Miscellanea section about the Hunt–Stein Theorem.)

The minimax criterion is a useful addition to the concept of admissibility. We saw in Example 10.4.1 that a decision rule may be admissible simply because it has

a small risk at some particular value of  $\theta$ . The risk may be large, even approaching infinity, at other values of  $\theta$ . The minimax criterion, on the other hand, ensures that the risk is not too large at any value of  $\theta$ . Hence, many find a decision rule that is both admissible and minimax to be very desirable.

Minimax decision rules are closely related to both Bayes decision rules and admissible decision rules. We now explore some of these relationships.

**THEOREM 10.5.1:** Suppose that  $\delta^\pi$  is a decision rule that is Bayes with respect to some prior  $\pi$ . If the risk function of  $\delta^\pi$  satisfies

$$(10.5.1) \quad R(\theta, \delta^\pi) \leq B(\pi, \delta^\pi), \quad \text{for all } \theta \in \Theta,$$

then  $\delta^\pi$  is a minimax decision rule.

*Proof:* Suppose  $\delta^\pi$  is not minimax. Then there is a decision rule  $\delta'$  such that

$$\sup_{\theta \in \Theta} R(\theta, \delta') < \sup_{\theta \in \Theta} R(\theta, \delta^\pi).$$

For this decision rule we have

$$\begin{aligned} B(\pi, \delta') &\leq \sup_{\theta \in \Theta} R(\theta, \delta') && \text{(Theorem 2.2.1)} \\ &< \sup_{\theta \in \Theta} R(\theta, \delta^\pi) && \text{(assumption)} \\ &\leq B(\pi, \delta^\pi), && \text{(from (10.5.1))} \end{aligned}$$

contradicting the fact that  $\delta^\pi$  is Bayes with respect to  $\pi$ . Hence  $\delta^\pi$  is minimax.  $\square$

The prior distribution in Theorem 10.5.1, whose Bayes rule is minimax, is called a *least favorable prior distribution*. It has the property that, if  $\pi'$  is any other prior and  $\delta^{\pi'}$  is a Bayes rule with respect to  $\pi'$ , then

$$(10.5.2) \quad B(\pi', \delta^{\pi'}) \leq B(\pi, \delta^\pi).$$

(See Exercise 10.28.) To use Theorem 10.5.1 to find a minimax rule, it is necessary to guess the least favorable prior, find the Bayes rules with respect to this prior, and verify that (10.5.1) is true.

The conditions in Theorem 10.5.1 can be relaxed somewhat and, in fact, the minimax rule need not be a Bayes rule. It may be the limit of Bayes rules. Exercise 10.30 provides a statement of such a result.

Suppose we have a decision rule  $\delta$  in mind. Think of "modifying"  $\delta$  to create a minimax decision rule. Since the minimax rule is quite concerned with the maximum risk, the modification will have to reduce the maximum value of the risk function. A modification that reduces the risk at some value of  $\theta$  often increases the risk at another value of  $\theta$  and the limit of this process often results in a decision rule that has constant risk. A decision rule with constant risk is called an *equalizer rule* and,

in many cases, an equalizer rule is minimax. The next two results specify conditions when this holds, the first result being a simple corollary of Theorem 10.5.1.

**COROLLARY 10.5.1:** Suppose that  $\delta$  is an equalizer rule and  $\delta$  is Bayes with respect to some prior  $\pi$ . Then  $\delta$  is minimax.

*Proof:* Exercise 10.32. □

When using Corollary 10.5.1, we often have an equalizer rule in mind. We then show it is Bayes with respect to some prior to show it is minimax, as illustrated in the next example.

**Example 10.5.1:** Let  $X_1, \dots, X_n$  be a random sample from a Bernoulli( $p$ ) population. Let  $Y = \sum X_i$ , and consider estimating  $p$  using the loss function  $L(p, a) = (p - a)^2/[p(1 - p)]$ . This loss penalizes mistakes more if  $p$  is near zero or one.

The MLE of  $p$ ,  $\hat{p} = Y/n$ , is an equalizer rule since

$$\begin{aligned} R(p, \hat{p}) &= E_p \left( \frac{(\hat{p} - p)^2}{p(1 - p)} \right) \\ &= \frac{1}{p(1 - p)} E_p (\hat{p} - E_p \hat{p})^2 && \left( \begin{array}{l} \hat{p} \text{ is an unbiased} \\ \text{estimate of } p \end{array} \right) \\ &= \frac{1}{p(1 - p)} \frac{p(1 - p)}{n} && \left( \text{Var } \hat{p} = \frac{p(1 - p)}{n} \right) \\ &= \frac{1}{n}. \end{aligned}$$

We will now show that  $\hat{p}$  is the Bayes estimator with respect to the uniform( $0, 1$ ) prior. Corollary 10.5.1 will then imply that  $\hat{p}$  is minimax.

The uniform prior is a beta( $1, 1$ ) distribution. By Example 7.2.9, the posterior distribution of  $p$  given  $Y = y$  is beta( $y + 1, n - y + 1$ ), and by Theorem 10.3.1, the Bayes estimate having observed  $Y = y$  is the value of  $a$  that minimizes

$$\begin{aligned} (10.5.3) \quad &\int_0^1 \frac{(p - a)^2}{p(1 - p)} \frac{\Gamma(n + 2)}{\Gamma(y + 1)\Gamma(n - y + 1)} p^y (1 - p)^{n-y} dp \\ &= \frac{(n + 1)n}{y(n - y)} \int_0^1 (p - a)^2 \frac{\Gamma(n)}{\Gamma(y)\Gamma(n - y)} p^{y-1} (1 - p)^{n-y-1} dp. \end{aligned}$$

(The equality is true only if  $y \neq 0$  and  $y \neq n$ . These cases will be treated later.) The last integral is equal to  $E(p - a)^2$  where  $p \sim \text{beta}(y, n - y)$ . (Note that we have absorbed the extra  $p(1 - p)$  factor into the posterior.) This is minimized by  $a = y/(y + n - y) = y/n$ , the mean for a beta( $y, n - y$ ) distribution. Thus, for  $y \neq 0$  and  $y \neq n$ ,  $\hat{p} = y/n$  is the Bayes estimator.

If  $y = 0$ , (10.5.3) simplifies to

$$\begin{aligned} \int_0^1 (p-a)^2 \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} p^{-1} (1-p)^{n-y-1} dp \\ = \int_0^1 (p-a)^2 (n+1)p^{-1} (1-p)^{n-1} dp. \end{aligned}$$

Due to the  $p^{-1}$  term, this integral is infinite if  $a \neq 0$ . But the integral is finite if  $a = 0$ . Thus if  $y = 0$ , the Bayes estimator is  $0 = y/n$ . A similar analysis for  $y = n$  yields the Bayes estimator as  $1 = y/n$ . Thus  $\hat{p} = y/n$  is the Bayes estimator for the uniform(0, 1) prior, and Corollary 10.5.1 tells us that  $\hat{p}$  is minimax for this loss. ||

Although minimax rules are often equalizer rules, this is certainly not always the case. The next example illustrates a minimax rule that is not an equalizer rule.

**Example 10.5.2:** The problem of estimating the mean of a normal distribution changes markedly if it is known that the mean lies in the interval  $[-m, m]$ , where  $m$  is a known, finite constant. To be specific, let  $X \sim N(\theta, 1)$ , and assume it is known that  $\theta \in [-m, m]$ , where  $0 < m < 1$ . The unique minimax estimator of  $\theta$ , using squared error loss, is

$$\delta^m(X) = m \tanh(mX),$$

where  $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$  is the hyperbolic tangent function (Casella and Strawderman, 1981). To prove that  $\delta^m$  is minimax, it can be shown that

1.  $\delta^m$  is Bayes against a prior  $\pi$  that gives probability  $\frac{1}{2}$  to the points  $\pm m$ .
2.  $R(\theta, \delta^m) \leq R(\theta, \delta^m)$  for all  $\theta \in [-m, m]$ .

But  $\delta^m$  is not an equalizer rule. Hence, Theorem 10.5.1 (but not Corollary 10.5.1) can be used to show that  $\delta^m$  is minimax. (See Exercise 10.29.) ||

The next theorem also addresses the question of minimaxity of equalizer rules, but is not as useful as Corollary 10.5.1. Usually it is harder to show a rule is admissible than it is to show a rule is minimax. (See Exercise 10.31.)

**THEOREM 10.5.2:** Suppose that  $\delta$  is an equalizer rule. If  $\delta$  is admissible then  $\delta$  is minimax.

*Proof:* Since  $\delta$  is an equalizer rule, there is a constant  $c$  such that  $R(\theta, \delta) = c$  for all  $\theta \in \Theta$ . Suppose that  $\delta$  is not minimax. This implies that there is a decision rule  $\delta'$  such that

$$\sup_{\theta \in \Theta} R(\theta, \delta') < \sup_{\theta \in \Theta} R(\theta, \delta),$$

and for every  $\theta \in \Theta$ ,

$$R(\theta, \delta') \leq \sup_{\theta \in \Theta} R(\theta, \delta') < \sup_{\theta \in \Theta} R(\theta, \delta) = c = R(\theta, \delta).$$

Therefore, it follows that  $\delta'$  is better than  $\delta$ , contradicting the fact that  $\delta$  is admissible and showing that  $\delta$  is minimax.  $\square$

**Example 10.5.3:** In Section 10.4.3 we showed that  $\bar{X}$  is an admissible estimator of a normal mean  $\theta$  when  $\sigma^2$  is known and squared error loss is used.  $\bar{X}$  is an equalizer rule since  $R(\theta, \bar{X}) = \text{Var } \bar{X} = \sigma^2/n$ . Thus, by Theorem 10.5.2,  $\bar{X}$  is minimax. However, this does not imply that  $\bar{X}$  is minimax in the unknown  $\sigma^2$  problem.

In the unknown  $\sigma^2$  problem  $\bar{X}$  is not an equalizer since  $R((\theta, \sigma), \bar{X}) = \sigma^2/n$  now depends on the parameter value. In fact, every estimator has  $\sup_{(\theta, \sigma)} R((\theta, \sigma), \delta) = \infty$  in this problem so *every* estimator is minimax! In Exercise 10.27, the loss  $L((\theta, \sigma), a) = (a - \theta)^2/\sigma^2$  is used and  $\bar{X}$  is both minimax (with finite maximum) and admissible for this loss.  $\parallel$

The previous results have described conditions under which other properties like equalizer or admissibility imply minimaxity. We now give one condition under which minimaxity implies admissibility. This result is not broadly applicable since verifying the conditions of the theorem is usually as difficult as proving admissibility by other methods.

**THEOREM 10.5.3:** Suppose that  $\delta$  is a unique minimax rule in that every minimax rule is equivalent to  $\delta$ . Then  $\delta$  is admissible.

*Proof:* Let  $\delta'$  be any other decision rule. If the risk function for  $\delta'$  is the same as the risk function for  $\delta$  then  $\delta'$  is not better than  $\delta$ . If the risk function for  $\delta'$  is different from the risk function for  $\delta$  then  $\delta'$  is not minimax. Since  $\delta$  is minimax,

$$\sup_{\theta \in \Theta} R(\theta, \delta') > \sup_{\theta \in \Theta} R(\theta, \delta).$$

Thus, for some  $\theta' \in \Theta$ ,

$$R(\theta', \delta') > \sup_{\theta \in \Theta} R(\theta, \delta) \geq R(\theta', \delta).$$

So also in this case,  $\delta'$  is not better than  $\delta$ . Since no  $\delta'$  is better than  $\delta$ , the decision rule  $\delta$  is admissible.  $\square$

**Example 10.5.4:** Let  $X \sim \text{binomial}(n, p)$  and consider estimating  $p$ . Suppose that  $\Theta = (0, 1)$  but  $\mathcal{A} = [0, 1]$ . Let the loss function be

$$L(\theta, a) = \left(1 - \frac{a}{p}\right)^2.$$

This loss function puts a high premium on getting a good estimate for  $p$  when  $p$  is near 0. The estimator  $\delta(x) = 0$  for all  $x = 0, \dots, n$  is the unique minimax rule. The risk of  $\delta$  is constant,  $R(p, \delta) = 1$  for all  $p$ , so if there is another minimax estimator, its maximum risk must be at most one. Let  $\delta'(x)$  be any other estimator. Then

$B = \{x : \delta'(x) > 0\}$  is nonempty. Let  $b = \min\{\delta'(x) : x \in B\} > 0$ . If  $p < b/2$  then  $L(p, \delta'(x)) > 1$  for all  $x \in B$ . Of course,  $L(p, \delta'(x)) = 1$  for all  $x \notin B$ . Thus, for any  $p < b/2$ ,

$$\begin{aligned} R(p, \delta') &= \sum_{x \in B} L(p, \delta'(x))f(x|p) + \sum_{x \notin B} L(p, \delta'(x))f(x|p) \\ &> \sum_{x \in B} 1f(x|p) + \sum_{x \notin B} 1f(x|p) && \left( \begin{array}{l} f(x|p) > 0 \\ \text{for all } x \in B \end{array} \right) \\ &= 1. \end{aligned}$$

Thus the maximum risk for  $\delta'$  is greater than one and  $\delta'$  is not minimax. Since  $\delta'$  was arbitrary, this implies  $\delta(x) = 0$  is the unique minimax estimator and, by Theorem 10.5.3,  $\delta(x) = 0$  is admissible. This is a simple example of the *control problem*. See Berliner (1983). ||

## 10.6 Invariant Decision Problems

We have discussed invariant decision procedures in each of the contexts of point estimation, hypothesis testing, and interval estimation. These ideas may have seemed a bit disparate. The invariant estimate in Example 6.3.3 changed in a prescribed way when the data point was transformed from  $x$  to  $g_a(x)$ . But in Definition 8.2.3 an invariant hypothesis test is required to make the same inference whether  $x$  or  $g(x)$  is observed. However, both of these ideas are special cases of a general definition of invariance for decision problems.

In previous sections we have considered various optimality criteria and classes of optimal rules. However, as explained in Chapter 6, invariance is a data reduction device, not an optimality criterion. Invariance is used to reduce the size of  $\mathcal{D}$ , the set of decision rules we are considering. The hope is that then one of the previously discussed optimality criteria might be applied to this reduced class to find a good decision rule. Invariance, in itself, is not the final goal.

The study of invariance in a decision theoretic context begins with the familiar Definition 6.3.2. Let  $\mathcal{G}$  be a group of transformations of the sample space  $\mathcal{X}$ . The model  $\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$  is invariant under the group  $\mathcal{G}$  if for every  $\theta \in \Theta$  and every  $g \in \mathcal{G}$  there exists a unique  $\theta' \in \Theta$  such that  $\mathbf{Y} = g(\mathbf{X}) \sim f(y|\theta')$  if  $\mathbf{X} \sim f(x|\theta)$ .

Recall that, if we fix  $g \in \mathcal{G}$  and allow  $\theta$  to vary, then this definition of invariance defines a function from  $\Theta$  into  $\Theta$ , denoted by  $\bar{g}(\theta)$ . For  $g$ ,  $\theta$ , and  $\theta'$  as in the definition, we have  $\bar{g}(\theta) = \theta'$  and if  $\mathbf{X} \sim f(x|\theta)$  then  $\mathbf{Y} = g(\mathbf{X}) \sim f(y|\bar{g}(\theta))$ .

**DEFINITION 10.6.1:** Let  $\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$  be a model that is invariant under the group  $\mathcal{G}$  of transformations of the sample space  $\mathcal{X}$ . The *decision problem is invariant under the group  $\mathcal{G}$*  if for every  $g \in \mathcal{G}$  and every  $a \in \mathcal{A}$  there exists a unique  $a' \in \mathcal{A}$  such that  $L(\theta, a) = L(\bar{g}(\theta), a')$  for all  $\theta \in \Theta$ . Since  $a'$  depends on  $g$  and  $a$ , the notation  $\tilde{g}(a) = a'$  will be used.

Since  $a'$  is unique, for fixed  $g \in \mathcal{G}$ ,  $\tilde{g}$  defines a function from  $\mathcal{A}$  into  $\mathcal{A}$  just as  $\bar{g}$  is a function from  $\Theta$  into  $\Theta$ . With all these definitions understood, the invariance relationship is summarized by the statement that  $L(\theta, a)$  must equal  $L(\bar{g}(\theta), \tilde{g}(a))$  for every  $g \in \mathcal{G}$ , every  $\theta \in \Theta$ , and every  $a \in \mathcal{A}$ .

In Chapter 6 we described how the invariance of a decision rule involved two separate kinds of invariance, measurement invariance and formal invariance. These two concepts are now embodied in the following definition.

**DEFINITION 10.6.2:** In an invariant decision problem, a decision rule  $\delta(x)$  is *invariant* if for every  $g \in \mathcal{G}$  and every  $x \in \mathcal{X}$

$$\delta(g(x)) = \tilde{g}(\delta(x)).$$

In Chapter 7 we saw that certain invariant estimates had constant mean squared error. In Chapter 9 we saw that invariant confidence sets had constant coverage probability. These phenomena are special cases of the following result.

**THEOREM 10.6.1:** Let  $\delta$  be an invariant decision rule in an invariant decision problem. Suppose that  $\theta$  and  $\theta'$  are two parameter points such that there is a  $g \in \mathcal{G}$  for which  $g(X) \sim f(y|\theta')$  if  $X \sim f(x|\theta)$ . Then  $R(\theta, \delta) = R(\theta', \delta)$ . In particular, if for every pair  $\theta$  and  $\theta'$  there is such a  $g$ , then  $R(\theta, \delta)$  is constant as a function of  $\theta$ .

*Proof:*

$$\begin{aligned} R(\theta', \delta) &= E_{\theta'} L(\theta', \delta(X)) && \text{(definition of risk)} \\ &= E_{\bar{g}(\theta)} L(\bar{g}(\theta), \delta(X)) && \left( \begin{array}{l} \text{definition of } \bar{g} \text{ and} \\ \text{assumption about } \theta \text{ and } \theta' \end{array} \right) \\ &= E_{\theta} L(\bar{g}(\theta), \delta(g(X))) && \left( \begin{array}{l} \text{if } X \sim f(x|\theta) \text{ then} \\ g(X) \sim f(y|\bar{g}(\theta)) \end{array} \right) \\ &= E_{\theta} L(\bar{g}(\theta), \tilde{g}(\delta(X))) && \text{(invariant decision rule)} \\ &= E_{\theta} L(\theta, \delta(X)) && \text{(invariant decision problem)} \\ &= R(\theta, \delta). && \text{(definition of risk)} \end{aligned}$$

The second assertion of the theorem follows immediately from the first.  $\square$

We now investigate how this general notion of invariance applies to some examples we have previously considered.

**Example 10.6.1:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population. Consider estimating  $\mu$  using squared error loss. Consider the translation group  $\mathcal{G} = \{g_c(x) : -\infty < c < \infty\}$  where

$$g_c(x_1, \dots, x_n) = (x_1 + c, \dots, x_n + c).$$

We saw in Example 6.3.2 that this model was invariant. Furthermore, since  $g_c(X_1, \dots, X_n)$  is a random sample from a  $n(\mu + c, \sigma^2)$  population, the transformation  $\bar{g}_c$  is  $\bar{g}_c(\mu, \sigma^2) = (\mu + c, \sigma^2)$ . Noting how the transformation  $\bar{g}_c$  affects the parameter, the corresponding transformation of the action is  $\tilde{g}_c(a) = a + c$ . To verify that this is the correct transformation, note that

$$L((\mu, \sigma^2), a) = (\mu - a)^2 = (\mu + c - (a + c))^2 = L(\bar{g}_c(\mu, \sigma^2), \tilde{g}_c(a)).$$

Thus Definition 10.6.1 is verified and this problem is invariant under this group.

Any estimator that satisfies  $T(x_1 + c, \dots, x_n + c) = T(x_1, \dots, x_n) + c$ , for all  $c$  and for all  $(x_1, \dots, x_n)$ , is an invariant estimator. To verify Definition 10.6.2 we note that

$$\begin{aligned} T(g_c(\mathbf{x})) &= T(x_1 + c, \dots, x_n + c) \\ &= T(x_1, \dots, x_n) + c \\ &= \tilde{g}_c(T(\mathbf{x})). \end{aligned}$$

Estimators that are invariant in this problem include the sample mean and the sample median. ||

**Example 10.6.2:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population. Consider estimating  $\sigma^2$  using squared error loss. Use the scale group  $\mathcal{G} = \{g_c(\mathbf{x}) : 0 < c < \infty\}$  where

$$g_c(x_1, \dots, x_n) = (cx_1, \dots, cx_n).$$

Then  $g_c(X_1, \dots, X_n)$  is a random sample from a  $n(c\mu, c^2\sigma^2)$  population. Thus the model is invariant under this group and  $\bar{g}_c(\mu, \sigma^2) = (c\mu, c^2\sigma^2)$ . If this problem were invariant under this group, then there would be a  $\tilde{g}_c$  such that

$$(\sigma^2 - a)^2 = L((\mu, \sigma^2), a) = L(\bar{g}_c(\mu, \sigma^2), \tilde{g}_c(a)) = (c^2\sigma^2 - \tilde{g}_c(a))^2.$$

This is true only if  $\tilde{g}_c(a) = c^2\sigma^2 + \sigma^2 - a$  or  $\tilde{g}_c(a) = c^2\sigma^2 - \sigma^2 + a$ . But according to Definition 10.6.1,  $\tilde{g}_c(a)$  can depend only on  $g_c$  and  $a$ , not  $\sigma^2$ . Thus this is not an invariant decision problem. ||

**Example 10.6.3:** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population. Consider testing  $H_0: \mu \leq 0$  versus  $H_1: \mu > 0$  using 0–1 loss. This problem is invariant under the scale transformation group defined in Example 10.6.2. Notice that according to the general decision theoretic definition of invariance, we need check only that the whole model is invariant under the group. We do not need to check that each subset of distributions,  $H_0$  and  $H_1$ , is invariant as was required in Definition 8.2.4. The hypothesis testing invariance we discussed in Chapter 8 is a special case of the general notion of invariance we are now discussing.

Recall that  $\overline{g_c}(\mu, \sigma^2) = (c\mu, c^2\sigma^2)$ . Since  $c > 0$ , if  $(\mu, \sigma^2) \in \Theta_0$ , that is,  $\mu \leq 0$ , then  $\overline{g_c}(\mu, \sigma^2) \in \Theta_0$  and if  $(\mu, \sigma^2) \in \Theta_0^c$  then  $\overline{g_c}(\mu, \sigma^2) \in \Theta_0^c$ . Thus action  $a_i, i = 0$  or 1, is correct or incorrect for  $(\mu, \sigma^2)$  and  $\overline{g_c}(\mu, \sigma^2)$  simultaneously. This suggests that the only transformation of the sample space needed is the identity transformation,  $\tilde{g}(a_i) = a_i, i = 0$  or 1. Unlike in the previous point estimation examples, here  $\tilde{g}$  does not depend on which  $g_c \in \mathcal{G}$  we are considering. To verify that this  $\tilde{g}$  is correct and Definition 10.6.1 is satisfied, note that if  $\mu \leq 0$  then  $c\mu \leq 0$  so that

$$L((\mu, \sigma^2), a_0) = 0 = L(\overline{g_c}(\mu, \sigma^2), a_0) = L(\overline{g_c}(\mu, \sigma^2), \tilde{g}(a_0))$$

and

$$L((\mu, \sigma^2), a_1) = 1 = L(\overline{g_c}(\mu, \sigma^2), a_1) = L(\overline{g_c}(\mu, \sigma^2), \tilde{g}(a_1)).$$

Similar equalities hold if  $\mu > 0$ . Thus the conditions of Definition 10.6.1 are satisfied and the decision problem is invariant.

Let  $\phi(\mathbf{x})$  be the test function for a test. Since  $\tilde{g}$  is just the identity transformation, Definition 10.6.2 says that a test is invariant if

$$\phi(g_c(\mathbf{x})) = \tilde{g}(\phi(\mathbf{x})) = \phi(\mathbf{x}).$$

This matches Definition 8.2.3. Thus the tests considered in Example 8.2.6, tests that depend on the sample only through the statistic  $\bar{X}/\sqrt{S^2/n}$ , are invariant decision rules. But there are other hypothesis testing problems in which tests that are invariant according to Definition 10.6.2 are not invariant according to Definition 8.2.4. (See Exercise 10.38.) Thus the type of invariance introduced in this section provides a more general concept of invariance for hypothesis testing problems than that considered in Chapter 8. ||

## 10.7 Stein's Paradox

In this section we will consider a special multivariate estimation problem, one that has some rather counterintuitive features. The problem to be considered is that of estimating several normal means simultaneously and is actually a special case of the statistical problem considered in Chapter 11. An excellent introduction to Stein's paradox is given in Efron and Morris (1977).

Let  $X_i \sim n(\theta_i, 1), i = 1, \dots, p$ , where  $p \geq 3$ . (This restriction will be addressed later.) Assume  $X_1, \dots, X_p$  are mutually independent. Notice that the  $X_i$ 's are not iid. They come from normal populations with possibly different means, but the problems will be tied together in that there will be one loss function for the  $p$  problems. Formally, we want to estimate  $\theta = (\theta_1, \dots, \theta_p)$ , using an estimator  $\delta(\mathbf{X}) = (\delta_1(\mathbf{X}), \dots, \delta_p(\mathbf{X}))$ . The loss function is

$$(10.7.1) \quad L(\theta, \delta(\mathbf{X})) = \sum_{i=1}^p (\theta_i - \delta_i(\mathbf{X}))^2.$$

This loss function is the sum of squared error loss functions, but there is one important point to see. Each  $\delta_i$  can be a function of  $(X_1, \dots, X_p)$ , so all of the data can be used in estimating each mean. Since the  $X_i$ 's are independent, we might think that restricting  $\delta_i$  just to be a function of  $X_i$  would be enough. However, the  $X_i$ 's are tied together in the loss function, and we will see that this matters.

The situation described here is not too farfetched and can be used as a model for a number of situations. For example, suppose a company needs to estimate average crop yield  $\theta_i$  based on data  $X_i$  for a number of different crops in different places. Although each estimation problem is separate, they all affect the company. So it is reasonable for the loss function to tie them together. Realize that good estimation overall takes precedence over doing well in any particular problem. The results obtained in the simple model considered here have been obtained in much more generality; see Berger (1985) or Lehmann (1983) for some generalizations.

Using Exercise 10.30, we will now show that the estimator  $\mathbf{X} = (X_1, \dots, X_p)$  is minimax. (In more generality, the sample mean is minimax in the combined problem. See Exercise 10.39.) An estimator  $\delta(\mathbf{X}) = (\delta_1(\mathbf{X}), \dots, \delta_p(\mathbf{X}))$  of the parameter  $\theta = (\theta_1, \dots, \theta_p)$ , using the loss function (10.7.1), has risk function

$$(10.7.2) \quad R(\theta, \delta) = E_\theta \left( \sum_{i=1}^p (\theta_i - \delta_i(\mathbf{X}))^2 \right).$$

Furthermore, if  $\pi(\theta)$  is a prior on  $\theta$ , then the Bayes risk of an estimator is

$$(10.7.3) \quad B(\pi, \delta) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} R(\theta, \delta) \pi(\theta) d\theta_1 \cdots d\theta_p.$$

Suppose now that we take a prior that is a product of independent priors, that is,

$$\pi(\theta) = \prod_{i=1}^p \pi_i(\theta_i), \quad \pi_i(\theta_i) \text{ a } n(0, \tau^2) \text{ pdf.}$$

For estimating  $\theta_i$  with loss  $(\theta_i - \delta_i(\mathbf{X}))^2$ , the Bayes rule against  $\pi_i, \delta_i^\pi$ , is

$$(10.7.4) \quad \delta_i^\pi(\mathbf{X}) = \delta_i^\pi(X_i) = \frac{\tau^2}{\tau^2 + 1} X_i,$$

with Bayes risk

$$B(\pi_i, \delta_i^\pi(X_i)) = \frac{\tau^2}{\tau^2 + 1}.$$

(These types of calculations are used in Section 10.4.3 and Exercise 10.8.) Since the priors are independent, it follows (see Exercise 10.10) that

$$\delta^\pi(\mathbf{X}) = (\delta_1^\pi(X_1), \dots, \delta_p^\pi(X_p))$$

is Bayes against the prior  $\pi(\theta) = \prod_{i=1}^p \pi_i(\theta_i)$  using the loss (10.7.1). The Bayes risk is

$$\begin{aligned} B(\pi, \delta^\pi(\mathbf{X})) &= \sum_{i=1}^p \frac{\tau^2}{\tau^2 + 1} \\ &= p \frac{\tau^2}{\tau^2 + 1} \\ &\rightarrow p, \quad \text{as } \tau^2 \rightarrow \infty, \\ &= R(\theta, \mathbf{X}), \end{aligned}$$

and hence, by Exercise 10.30,  $\mathbf{X}$  is minimax.

Even though  $\mathbf{X}$  is minimax,  $\mathbf{X}$  is not unique minimax and, since the risk of  $\mathbf{X}$  is constant at the minimax value, any other minimax estimator will be better than  $\mathbf{X}$ . Unlike the one-dimensional problem where  $\mathbf{X}$  is admissible, it is not admissible in higher dimensions (three or more).

This was established by Stein (1955) who showed that, if the dimension of the problem was at least three, then there exists a better procedure. (It was shown in Stein (1955) and James and Stein (1961) that the sample mean is admissible in one and two dimensions.) More importantly, in James and Stein (1961), a better estimator was exhibited. That seemingly nonintuitive estimator (but also see Exercise 10.40) is given by  $\delta^S(\mathbf{X}) = (\delta_1^S(\mathbf{X}), \dots, \delta_p^S(\mathbf{X}))$ , where

$$(10.7.5) \quad \delta_i^S(\mathbf{X}) = \left(1 - \frac{p-2}{\sum_{j=1}^p X_j^2}\right) X_i.$$

The original proof that  $\delta^S$  dominates  $\mathbf{X}$  is quite long and cumbersome, relying on representations of noncentral chi squared distributions. A more elegant and useful proof, however, was given by Stein using his Lemma (Stein, 1973, 1981). This use of Stein's Lemma, or more accurately, employment of integration by parts, was discovered independently by Berger (1975), who also used it to establish minimaxity of a class of estimators.

Recall Stein's Lemma, given in Chapter 4. If  $X \sim n(\theta, \sigma^2)$ , then

$$E(g(X)(X - \theta)) = \sigma^2 E g'(X),$$

provided the expectations exist. Using this identity, computation of the risk of  $\delta^S$  is relatively easy. We have

$$\begin{aligned} R(\theta, \delta^S) &= E_\theta \left[ \sum_{i=1}^p (\theta_i - \delta_i^S(\mathbf{X}))^2 \right] && \text{(definition of risk)} \\ &= \sum_{i=1}^p E_\theta [\theta_i - \delta_i^S(\mathbf{X})]^2 && \text{(property of expectation)} \end{aligned}$$

$$\begin{aligned}
 &= \sum_{i=1}^p E_\theta \left[ \theta_i - \left( 1 - \frac{p-2}{\sum_{j=1}^p X_j^2} \right) X_i \right]^2 && \text{(definition of estimator)} \\
 &= \sum_{i=1}^p E_\theta \left[ (\theta_i - X_i) + \frac{p-2}{\sum_{j=1}^p X_j^2} X_i \right]^2 \\
 (10.7.6) \quad &= \sum_{i=1}^p E_\theta (\theta_i - X_i)^2 + 2 \sum_{i=1}^p E_\theta \left( (\theta_i - X_i) \frac{p-2}{\sum_{j=1}^p X_j^2} X_i \right) \\
 &\quad + \sum_{i=1}^p E_\theta \left( \frac{p-2}{\sum_{j=1}^p X_j^2} X_i \right)^2. && \text{(expand the square)}
 \end{aligned}$$

The first expectation in (10.7.6) is equal to  $p$ , since it is the risk of  $\mathbf{X}$ , and simple manipulation will show that the third expectation is equal to  $(p-2)^2 E_\theta(1/\sum_{j=1}^p X_j^2)$ . For the middle term we use Stein's Lemma:

$$\sum_{i=1}^p E_\theta \left( (\theta_i - X_i) \frac{p-2}{\sum_{j=1}^p X_j^2} X_i \right) = -(p-2) \sum_{i=1}^p E_\theta \left( \frac{\partial}{\partial X_i} \frac{X_i}{\sum_{j=1}^p X_j^2} \right).$$

Differentiating and gathering terms gives

$$\begin{aligned}
 -(p-2) \sum_{i=1}^p E_\theta \left( \frac{\partial}{\partial X_i} \frac{X_i}{\sum_{j=1}^p X_j^2} \right) &= -(p-2) \sum_{i=1}^p E_\theta \left( \frac{\sum_{j=1}^p X_j^2 - 2X_i^2}{(\sum_{j=1}^p X_j^2)^2} \right) \\
 &= -(p-2)^2 E_\theta \left( \frac{1}{\sum_{j=1}^p X_j^2} \right).
 \end{aligned}$$

Putting this all together, we have the risk of the Stein estimator to be

$$\begin{aligned}
 R(\theta, \delta^S) &= p - 2(p-2)^2 E_\theta \left( \frac{1}{\sum_{j=1}^p X_j^2} \right) + (p-2)^2 E_\theta \left( \frac{1}{\sum_{j=1}^p X_j^2} \right) \\
 &= p - (p-2)^2 E_\theta \left( \frac{1}{\sum_{j=1}^p X_j^2} \right) \\
 &< p = R(\theta, \mathbf{X}).
 \end{aligned}$$

Thus the risk of  $\delta^S$  is smaller than the risk of  $\mathbf{X}$  and  $\mathbf{X}$  is inadmissible. The above inequality is valid as long as the expectation exists, and the expectation exists as long as  $p \geq 3$ . If  $p = 1$  or  $2$ ,  $E_\theta(1/\sum_{j=1}^p X_j^2) = \infty$ .

The estimator  $\delta^S$  is one of a family of estimators defined by

$$(10.7.7) \quad \delta_i^c(\mathbf{X}) = \left( 1 - \frac{c}{\sum_{j=1}^p X_j^2} \right) X_i, \quad i = 1, \dots, p.$$

Any such estimator with  $0 < c < 2(p - 2)$  is better than  $\mathbf{X}$ , but the choice  $c = p - 2$  is optimal (see Exercise 10.41). These estimators, however, can also be uniformly improved upon in a simple way (Efron and Morris, 1973) by using a *positive-part* estimator,

$$(10.7.8) \quad \delta_i^+(\mathbf{X}) = \left( 1 - \frac{p-2}{\sum_{j=1}^p X_j^2} \right)^+ X_i, \quad i = 1, \dots, p,$$

where we define the notation  $(x)^+ = \max(0, x)$ . Hence the coordinates of the positive-part estimator cannot have a different sign from the coordinates of  $\mathbf{X}$ . Also, the positive-part estimator alleviates the strange behavior of the Stein estimator near zero. (Note that as  $\mathbf{X} \rightarrow 0$ ,  $\delta_i^S \rightarrow -\infty$  or  $+\infty$ . Although this behavior does not adversely affect the risk, it would make an experimenter uncomfortable if a small  $\mathbf{X}$  were observed.) The risk functions of  $\mathbf{X}$ ,  $\delta^S$ , and  $\delta^+$ , which depend on  $\theta$  only through  $\sum_{i=1}^p \theta_i^2$ , are shown in Figure 10.7.1. Notice that the biggest risk improvement is obtained near  $\theta = 0$ , because these estimators all shrink  $\mathbf{X}$  toward the point  $(0, 0, \dots, 0)$ . There is nothing magic about zero, however, and these estimators can shrink toward any point.

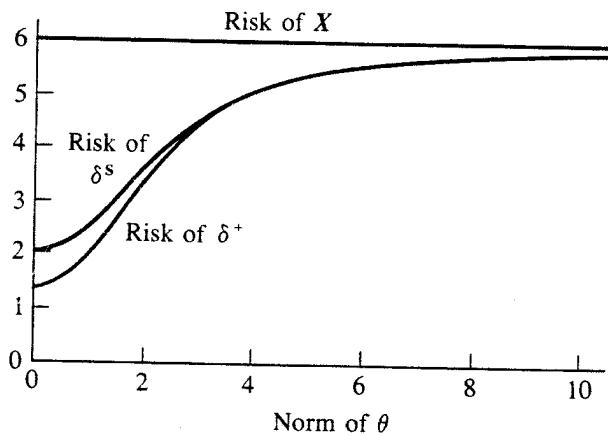


FIGURE 10.7.1 Risk functions for Stein-type estimators

Interestingly, even though  $\delta^+$  is a very good estimator (Efron and Morris, 1973), the results of Brown (1971), which generalized the work of Sacks (1963), show that  $\delta^+$  is inadmissible. Thus, there are estimators that uniformly dominate  $\delta^+$ . Even though admissible estimators for this problem have been found (Strawderman, 1971; Berger, 1976), no one has found an estimator that dominates  $\delta^+$ . Although, practically speaking,  $\delta^+$  cannot be improved upon by very much, finding an estimator that dominates it would be a theoretical achievement.

Finally, we note that the Stein Paradox carries over to set estimation in that, in three or more dimensions, the usual confidence set for a vector of normal means is inadmissible. There exist confidence sets centered at Stein-type estimators that have the same volume and higher coverage probability, or smaller volume and the same confidence coefficient. Brown (1966) and Joshi (1967) independently proved the existence of a dominating procedure for  $p \geq 3$ , and Joshi (1969) later proved the admissibility of the usual confidence set if  $p = 1$  or 2. Hwang and Casella (1982) first exhibited a dominating set and Casella and Hwang (1983, 1987) explored the set estimation problem further.

## EXERCISES

---

- 10.1** Let  $X$  have a  $n(\theta, 1)$  distribution, and consider testing  $H_0: \theta \geq \theta_0$  versus  $H_1: \theta < \theta_0$ . Use the loss function (10.2.5) and investigate the three tests that reject  $H_0$  if  $X < -z_\alpha + \theta_0$  for  $\alpha = .1, .3$ , and  $.5$ .
- For  $b = c = 1$ , graph and compare their risk functions.
  - For  $b = 3, c = 1$ , graph and compare their risk functions.
  - Graph and compare the power functions of the three tests to the risk functions in parts (a) and (b).
- 10.2** Consider testing  $H_0: p \leq \frac{1}{3}$  versus  $H_1: p > \frac{1}{3}$  where  $X \sim \text{binomial}(5, p)$  using 0–1 loss. Graph and compare the risk functions for the following two tests. Test I rejects  $H_0$  if  $X = 0$  or 1. Test II rejects  $H_0$  if  $X = 4$  or 5.
- 10.3** Consider the binomial estimation problem in Example 10.4.1 for  $n = 10$ . Graph and compare the risk functions for these two estimators,  $\delta(x) = \frac{1}{3}$  and  $\delta'(x) = x/10$ .
- 10.4** Show that the log of the likelihood function for estimating  $\sigma^2$ , based on observing  $S^2 \sim \sigma^2 \chi^2_\nu / \nu$ , can be written in the form

$$\log L(\sigma^2 | s^2) = K_1 \frac{s^2}{\sigma^2} - K_2 \log \frac{s^2}{\sigma^2} + K_3,$$

where  $K_1, K_2$ , and  $K_3$  are constants, not dependent on  $\sigma^2$ . Relate the above log likelihood to the loss function discussed in Example 10.2.3. See Anderson (1984a) for a discussion of this relationship.

- 10.5** Let  $X \sim n(\mu, \sigma^2)$ ,  $\sigma^2$  known. For each  $c \geq 0$ , define an interval estimator for  $\mu$  by  $C(x) = [x - c\sigma, x + c\sigma]$  and consider the loss in (10.2.7).
- Show that the risk function,  $R(\mu, C)$ , is given by

$$R(\mu, C) = b(2c\sigma) - P(-c \leq Z \leq c).$$

- b. Using the Fundamental Theorem of Calculus, show that

$$\frac{d}{dc} R(\mu, C) = 2b\sigma - \frac{2}{\sqrt{2\pi}} e^{-c^2/2}$$

and, hence, the derivative is an increasing function of  $c$  for  $c \geq 0$ .

- c. Show that if  $b\sigma > 1/\sqrt{2\pi}$ , the derivative is positive for all  $c \geq 0$  and, hence,  $R(\mu, C)$  is minimized at  $c = 0$ . That is, the best interval estimator is the point estimator  $C(x) = [x, x]$ .

- d. Show that if  $b\sigma \leq 1/\sqrt{2\pi}$ , the  $c$  that minimizes the risk is  $c = \sqrt{-2 \log(b\sigma\sqrt{2\pi})}$ . Hence, if  $b$  is chosen so that  $c = z_{\alpha/2}$  for some  $\alpha$ , then the interval estimator that minimizes the risk is just the usual  $1-\alpha$  confidence interval.

- 10.6** Let  $X \sim n(\mu, \sigma^2)$ , but now consider  $\sigma^2$  unknown. For each  $c \geq 0$ , define an interval estimator for  $\mu$  by  $C(x) = [x - cs, x + cs]$ , where  $s^2$  is an estimator of  $\sigma^2$  independent of  $X$ ,  $\nu S^2/\sigma^2 \sim \chi_\nu^2$  (for example, the usual sample variance). Consider a modification of the loss in (10.2.7),

$$L((\mu, \sigma), C) = \frac{b}{\sigma} \text{Len}(C) - I_C(\mu).$$

- a. Show that the risk function,  $R((\mu, \sigma), C)$ , is given by

$$R((\mu, \sigma), C) = b(2cM) - [2P(T \leq c) - 1],$$

where  $T \sim t_\nu$  and  $M = ES/\sigma$ .

- b. If  $b \leq 1/\sqrt{2\pi}$ , show that the  $c$  that minimizes the risk satisfies

$$b = \frac{1}{\sqrt{2\pi}} \left( \frac{\nu}{\nu + c^2} \right)^{(\nu+1)/2}$$

- c. Reconcile this problem with the known  $\sigma^2$  case. Show that as  $\nu \rightarrow \infty$ , the solution here converges to the solution in the known  $\sigma^2$  problem. (Be careful of the rescaling done to the loss function.)

- 10.7** The decision theoretic approach to set estimation can be quite useful (Exercise 10.34) but it can also give some unsettling results, showing the need for thoughtful implementation. Consider again the case of  $X \sim n(\mu, \sigma^2)$ ,  $\sigma^2$  unknown, and suppose that we have an interval estimator for  $\mu$  by  $C(x) = [x - cs, x + cs]$ , where  $s^2$  is an estimator of  $\sigma^2$  independent of  $X$ ,  $\nu S^2/\sigma^2 \sim \chi_\nu^2$ . This is, of course, the usual  $t$  interval, one of the great statistical procedures that has withstood the test of time. Consider the loss

$$L((\mu, \sigma), C) = b \text{Len}(C) - I_C(\mu),$$

similar to that used in Exercise 10.6, but without scaling the length. Construct another procedure  $C'$  as

$$C' = \begin{cases} [x - cs, x + cs] & \text{if } s < K \\ \emptyset & \text{if } s \geq K \end{cases},$$

where  $K$  is a positive constant. Notice that  $C'$  does *exactly the wrong thing*. When  $s^2$  is big and there is a lot of uncertainty, we would want the interval to be wide. But  $C'$  is empty! Show that we can find a value of  $K$  so that

$$R((\mu, \sigma), C') \leq R((\mu, \sigma), C), \quad \text{for every } (\mu, \sigma),$$

with strict inequality for some  $(\mu, \sigma)$ .

- 10.8** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population,  $\sigma^2$  known. Consider estimating  $\theta$  using squared error loss. Let  $\pi(\theta)$  be a  $n(\mu, \tau^2)$  prior distribution on  $\theta$  and let  $\delta^\pi$  be the Bayes estimator of  $\theta$ . Verify the following formulas for the risk function and Bayes risk.

a. For any constants  $a$  and  $b$ , the estimator  $\delta(\mathbf{x}) = a\bar{X} + b$  has risk function

$$R(\theta, \delta) = a^2 \frac{\sigma^2}{n} + (b - (1-a)\theta)^2.$$

b. Let  $\eta = \sigma^2/(n\tau^2 + \sigma^2)$ . The risk function for the Bayes estimator is

$$R(\theta, \delta^\pi) = (1-\eta)^2 \frac{\sigma^2}{n} + \eta^2(\theta - \mu)^2.$$

c. The Bayes risk for the Bayes estimator is

$$B(\pi, \delta^\pi) = \tau^2 \eta.$$

**10.9** Let  $X \sim n(\mu, 1)$ . Let  $\delta^\pi$  be the Bayes estimator of  $\mu$  for squared error loss. Compute and graph the risk functions,  $R(\mu, \delta^\pi)$ , for  $\pi(\mu) \sim n(0, 1)$  and  $\pi(\mu) \sim n(0, 10)$ . Comment on how the prior affects the risk function of the Bayes estimator.

**10.10** Let  $X_1, \dots, X_n$  be independent random variables, where  $X_i$  has cdf  $F(x|\theta_i)$ . Show that, for  $i = 1, \dots, n$ , if  $\delta_i^{\pi_i}(X_i)$  is a Bayes rule for estimating  $\theta_i$  using loss  $L(\theta_i, a_i)$  and prior  $\pi_i(\theta_i)$ , then  $\delta^\pi(\mathbf{X}) = (\delta^{\pi_1}(X_1), \dots, \delta^{\pi_n}(X_n))$  is a Bayes rule for estimating  $\theta = (\theta_1, \dots, \theta_n)$  using the loss  $\sum_{i=1}^n L(\theta_i, a_i)$  and prior  $\pi(\theta) = \prod_{i=1}^n \pi_i(\theta_i)$ .

**10.11** A loss function investigated by Zellner (1986) is the LINEX (LINEar-EXPonential) loss, a loss function that can handle asymmetries in a smooth way. The LINEX loss is given by

$$L(\theta, a) = e^{c(a-\theta)} - c(a-\theta) - 1,$$

where  $c$  is a positive constant. As the constant  $c$  varies, the loss function varies from very asymmetric to almost symmetric.

a. For  $c = .2, .5, 1$ , plot  $L(\theta, a)$  as a function of  $a - \theta$ .

b. If  $X \sim F(x|\theta)$ , show that the Bayes estimator of  $\theta$ , using a prior  $\pi$ , is given by  $\delta^\pi(X) = \frac{-1}{c} \log E(e^{-c\theta}|X)$ .

c. Let  $X_1, \dots, X_n$  be iid  $n(\theta, \sigma^2)$ , where  $\sigma^2$  is known, and suppose that  $\theta$  has the noninformative prior  $\pi(\theta) = 1$ . Show that the Bayes estimator versus LINEX loss is given by  $\delta^B(\bar{X}) = \bar{X} - (c\sigma^2/(2n))$ .

d. Calculate  $r(x, a)$  of (10.3.3) for  $\delta^B(\bar{X})$  and  $\bar{X}$  using LINEX loss.

e. Calculate  $r(x, a)$  of (10.3.3) for  $\delta^B(\bar{X})$  and  $\bar{X}$  using squared error loss.

**10.12** Consider testing  $H_0 : \mu \leq 0$  versus  $H_1 : \mu > 0$  using 0–1 loss, where  $X \sim n(\mu, 1)$ . Let  $\delta_c$  be the test that rejects  $H_0$  if  $X > c$ . The class of tests  $\{\delta_c, -\infty \leq c \leq \infty\}$ , is an essentially complete class for this problem. Let  $\delta$  be the test that rejects  $H_0$  if  $1 < X < 2$ . Find a test  $\delta_c$  that is better than  $\delta$ . (Either prove that the test is better or graph the risk functions for  $\delta$  and  $\delta_c$  and carefully explain why the proposed test should be better.)

**10.13** Again let  $X \sim n(\mu, 1)$ , but now consider testing  $H_0 : \mu = 0$  versus  $H_1 : \mu \neq 0$  using 0–1 loss. Let  $\delta_{c,d}$  be the test that accepts  $H_0$  if  $c \leq X \leq d$ . Wald (1950) showed that the class of tests  $\{\delta_{c,d} : -\infty \leq c \leq d \leq \infty\}$ , is an essentially complete class for this problem. Let  $\delta$  be the test that accepts  $H_0$  if  $1 \leq X \leq 2$  or  $-2 \leq X \leq -1$ . Find a test  $\delta_c$  that is better than  $\delta$ . (Either prove that the test is better or graph the risk functions for  $\delta$  and  $\delta_c$  and carefully explain why the proposed test should be better.)

- 10.14** Assume that the parameter space in a decision problem is finite, say  $\Theta = \{\theta_1, \dots, \theta_m\}$ . Suppose that  $\delta^\pi$  is the Bayes rule with respect to a prior distribution  $\pi$  that gives positive probability to every possible value of  $\theta$ . Show that  $\delta^\pi$  is admissible.
- 10.15** Suppose that for a certain prior distribution  $\pi$ , every Bayes rule with respect to  $\pi$  (if there is more than one) has the same risk function. Prove that these Bayes rules are admissible. In other words, unique Bayes rules are admissible.
- 10.16** Let  $X \sim n(\mu, \sigma^2)$ ,  $\sigma^2$  known. Consider estimating  $\mu$  using squared error loss. Show that the estimator that always estimates  $\mu$  to be 17, that is,  $\delta(x) = 17$  for all  $x$ , is admissible. (Note that this is similar to Example 10.4.1 but here sets of probability zero are important.)
- 10.17** Let  $\delta_1$  and  $\delta_2$  be two decision rules and suppose that the loss  $L(\theta, a)$  is convex, as defined in Theorem 10.4.3. Suppose the action space  $\mathcal{A}$  is a convex set. Define the randomized decision rule  $\delta^a$  by

$$\delta^a(\mathbf{x}) = \begin{cases} \delta_1(\mathbf{x}) & \text{with probability } a \\ \delta_2(\mathbf{x}) & \text{with probability } 1 - a \end{cases},$$

where  $a$  is a constant,  $0 < a < 1$ . Define the nonrandomized decision rule  $\delta^*$  by averaging  $\delta_1$  and  $\delta_2$  with weights determined by  $a$ , that is,  $\delta^*(\mathbf{x}) = a\delta_1(\mathbf{x}) + (1-a)\delta_2(\mathbf{x})$ . Show that  $\delta^*$  is as good as  $\delta^a$ .

- 10.18** The following calculations show that, in most cases, we can ignore randomized rules in decision theoretic point estimation, but we must be careful of them in hypothesis testing and interval estimation.
- Show that both absolute error loss and squared error loss are convex losses.
  - Show that generalized 0–1 loss, given in (10.2.3), is not a convex loss.
  - Show that the loss for interval estimators given in (10.2.7) is not a convex loss.
- 10.19** A class of decision rules  $\mathcal{C}$  is called *minimal complete* if it is a complete class and no proper subclass of  $\mathcal{C}$  is complete. Show that if a minimal complete class exists, it is exactly the class of admissible decision rules.
- 10.20** Prove the following. Let  $\mathcal{C}$  be an essentially complete class of decision rules. If  $\delta' \notin \mathcal{C}$  is admissible, then there is a  $\delta \in \mathcal{C}$  such that  $\delta$  is equivalent to  $\delta'$ .
- 10.21** Show that the admissible estimator  $\delta(x) = \frac{1}{3}$  from Example 10.4.1 is a Bayes rule for some prior  $\pi(p)$ .
- 10.22** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, 1)$  population. Consider testing  $H_0: \mu = 0$  versus  $H_1: \mu > 0$  using a loss function that satisfies  $L(\mu, a_1) \leq L(\mu, a_0)$  for all  $\mu > 0$ . Show that the class of tests with rejection regions  $R_k = \{\mathbf{x}: \bar{x} > k\}$ ,  $-\infty \leq k \leq \infty$ , is an essentially complete class.
- 10.23** Let  $X_1, \dots, X_n$  be a random sample from a population with mean  $\theta$  and variance  $\sigma^2$ ,  $0 < \sigma^2 < \infty$ . Consider estimating  $\theta$  using squared error loss.
- Show that any estimator of the form  $a\bar{X} + b$ , where  $a > 1$  and  $b$  are constants, is inadmissible.
  - Show that if  $a = 1$  and  $b \neq 0$ , then the estimator is inadmissible.
  - Show that, if  $X_i \sim n(\theta, \sigma^2)$ ,  $\sigma^2$  known, then  $a\bar{X} + b$  is admissible if  $a < 1$ .
- 10.24** Let  $X$  have a discrete uniform  $(1, \theta)$  distribution. That is,  $X$  and  $\theta$  are both integer-valued and the parameter space is  $\Theta = \{1, 2, \dots\}$ . Consider estimating  $\theta$  using squared error loss.
- Let the action space  $\mathcal{A} = \Theta$ . Show that for some priors, the estimator  $\delta(x) = E(\theta|x)$  is not the Bayes estimator of  $\theta$ .
  - Now suppose that the action space is  $\mathcal{A} = [1, \infty)$ . Assuming that the expectation exists, show that  $\delta(x) = E(\theta|x)$  is the Bayes estimator of  $\theta$ .

c. Show that the estimator  $\delta(x) = x$  is admissible, regardless of which of the above two action spaces is used. (Hint: Show that  $R(1, \delta)$  is as small as possible. Then use induction on the value of  $\theta$ .)

d. The estimator  $\delta(x) = x$  is Bayes with respect to some prior. What is the prior? Show that there are other Bayes estimators for this prior that have different risk functions than  $\delta$ . (See Exercise 10.15.)

- 10.25** A missile can travel at either high or low trajectory. The missile's effectiveness decreases linearly with the distance by which it misses its target, up to a distance of 2 miles, at which it becomes totally ineffective.

If a low trajectory is used, the missile is safe from antimissile fire, but its accuracy is subject to the proportion of cloud cover,  $\theta$ . In fact, the distance,  $d$ , by which the missile misses its target, is uniformly distributed on  $[-\theta, \theta]$ . For the target area it is reasonable to assume that  $\theta$  has a beta(2, 2) prior pdf.

If a high trajectory is used, the missile will hit the target exactly unless it is first destroyed by antimissile fire. From previous experience, the probability,  $\xi$ , of the missile being destroyed is thought to have a beta(1, 2) prior pdf. An experiment is conducted to provide further information about  $\xi$ . Two missiles are launched using a high trajectory, out of which none is shot down.

- What is the loss incurred in having the missile miss the target by a distance  $d$ , where we code 0 = perfect hit, 1 = total miss.
- What is the optimal Bayes trajectory?
- What is the optimal minimax trajectory?
- What trajectories are admissible?

- 10.26** Consider the hypothesis testing problem and loss function given in Example 10.2.4, and let  $\sigma = n = 1$ . Consider tests that reject  $H_0$  if  $X < -z_\alpha + \theta_0$ . Find the value of  $\alpha$  that yields a minimax test.

- 10.27** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population,  $\sigma^2$  unknown. Consider estimating  $\theta$  using the loss function  $L((\theta, \sigma), a) = (a - \theta)^2 / \sigma^2$ .

- Show that  $\bar{X}$  is an admissible estimator of  $\theta$ .
- Show that  $\bar{X}$  is minimax.

- 10.28** a. Show that if  $\pi$  and  $\delta^\pi$  satisfy the conditions of Theorem 10.5.1, then inequality (10.5.2) is true.  
 b. Show that a decision rule  $\delta'$  is minimax if and only if

$$\sup_{\pi \in \Pi} B(\pi, \delta') = \inf_{\delta \in \mathcal{D}} \sup_{\pi \in \Pi} B(\pi, \delta),$$

where  $\Pi$  = the class of all priors. Hence, this can be an alternate definition of minimaxity.

- c. Show that if  $\delta^{\pi_0}$  is minimax, then

$$B(\pi_0, \delta^{\pi_0}) = \sup_{\pi \in \Pi} B(\pi, \delta^\pi).$$

- 10.29** Let  $X \sim n(\theta, 1)$  and assume it is known that  $\theta \in [-m, m]$ ;  $0 < m < 1$  is a known constant. In Example 10.5.2 it was asserted that the estimator  $\delta^m(X) = m \tanh(mX)$  is minimax against squared error loss. This exercise will fill in some details.

- Show that  $\delta^m$  is Bayes against a prior that gives probability  $\frac{1}{2}$  to the points  $\pm m$ .
- Compute the Bayes risk of  $\delta^m$ ,  $B(\pi, \delta^m)$ , and show that it is equal to  $R(m, \delta^m)$ .

- c. Explain why, if  $\max_{\theta} R(\theta, \delta^m) = R(m, \delta^m)$ , then  $\delta^m$  is minimax.
- 10.30** Let  $\pi_n, n = 1, 2, \dots$ , be a sequence of prior distributions. Let  $\delta_n$  denote a Bayes rule with respect to  $\pi_n$ . If  $B(\pi_n, \delta_n)$  converges to a number  $c$  and if  $\delta$  is a decision rule with  $R(\theta, \delta) \leq c$  for all  $\theta \in \Theta$ , then  $\delta$  is minimax.
- 10.31** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\theta, \sigma^2)$  population,  $\sigma^2$  known. Consider estimating  $\theta$  using squared error loss. Use the result in Exercise 10.30 to show that  $\bar{X}$  is a minimax estimator. Note that less is required here of the difference  $B(\pi_n, \bar{X}) - B(\pi_n, \delta_n)$  than was required in Section 10.4.3 where we showed that  $\bar{X}$  is admissible.
- 10.32** Prove Corollary 10.5.1.
- 10.33** Let  $X_1, \dots, X_n$  be a random sample from a Bernoulli( $p$ ) population. Consider estimating  $p$  using squared error loss. Let  $\hat{p}_B = (\sum X_i + \sqrt{n}/4)/(n + \sqrt{n})$ .
- Show that  $\hat{p}_B$  is an equalizer rule.
  - Show that  $\hat{p}_B$  is minimax.
  - Show that  $\hat{p}_B$  is admissible.
- 10.34** Let  $X \sim f(x|\theta)$  and suppose that we want to estimate  $\theta$  with an interval estimator  $C$  using the loss in (10.2.7).
- If  $\theta$  has the prior pdf  $\pi(\theta)$ , show that the Bayes rule is given by

$$C^\pi = \{\theta : \pi(\theta|x) \geq b\}.$$

(Hint: Write  $\text{Len}(C) = \int_C 1 d\theta$  and use the Neyman–Pearson Lemma.)

b. Now let  $X \sim n(\theta, \sigma^2)$ , where  $\sigma^2$  is known. Show that, using the loss (10.2.7), the estimator  $C(x) = [x - c\sigma, x + c\sigma]$  can be approached by a sequence of Bayes rules and, using Exercise 10.30, conclude that for certain values of  $b$  and  $c$ ,  $C(x)$  is minimax. Find the values of  $b$  and  $c$ . (Hint: A normal prior will work.)

- 10.35** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population. Consider estimating  $\mu$  using squared error loss and consider the translation group from Example 10.6.1. For what values of  $a$  and  $b$  is  $T_{a,b}(x) = a\bar{x} + b$  an invariant estimator?
- 10.36** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population. Consider estimating  $\sigma^2$  using the loss

$$L((\mu, \sigma^2), a) = \left(1 - \frac{a}{\sigma^2}\right)^2.$$

Show that this estimation problem is invariant under the scale group of transformations from Example 10.6.2. Show that the sample variance  $S^2$  is an invariant estimator.

- 10.37** Let  $X_1, \dots, X_n$  be a random sample from a  $n(\mu, \sigma^2)$  population. Consider testing  $H_0: \mu \leq 0$  versus  $H_1: \mu > 0$  using the loss

$$L((\mu, \sigma^2), a_0) = \begin{cases} \mu/\sigma & \mu > 0 \\ 0 & \mu \leq 0 \end{cases} \quad \text{and} \quad L((\mu, \sigma^2), a_1) = \begin{cases} |\mu|/\sigma & \mu \leq 0 \\ 0 & \mu > 0 \end{cases}.$$

Consider the scale group from Example 10.6.3. Show that this testing problem is invariant under this group. Show that any test based on the statistic  $\bar{X}/\sqrt{S^2/n}$  is an invariant test.

- 10.38** This is an example of a hypothesis testing problem that is invariant according to the general decision theoretic definition but not according to the definition in Chapter 8. Let  $X \sim \text{binomial}(n, p)$ . The sample size  $n$  is known but  $p$  is unknown. Consider testing  $H_0: p \leq \frac{1}{2}$  versus  $H_1: p > \frac{1}{2}$  using generalized 0–1 loss with the following modification.

For  $p = \frac{1}{2}$ ,  $L(\frac{1}{2}, a_0) = L(\frac{1}{2}, a_1) = 0$ . Consider the group  $\mathcal{G} = \{g_1, g_2\}$  from Example 6.3.1 that has only two elements,

$$g_1(x) = n - x \quad \text{and} \quad g_2(x) = x.$$

a. Show that this testing problem is invariant under this group if and only if the constants from the loss function satisfy  $c_I = c_{II}$ . What are  $\bar{g}_1$ ,  $\bar{g}_2$ ,  $\tilde{g}_1$ , and  $\tilde{g}_2$ ? Why was the modification of the loss for  $p = \frac{1}{2}$  necessary?

b. Explain why the conditions of Definition 8.2.4 are not satisfied.

c. Suppose  $n$  is odd. Let  $c_I = c_{II} = 1$ . Show that a test  $\phi$  is invariant for this problem if and only if, for every  $x = 0, \dots, n$ ,  $\phi$  takes the opposite actions at  $x$  and  $n - x$ .

d. Show that no invariant test exists if  $n$  is even. (Hint: The point  $x = n/2$  creates problems.)

e. Explain why the invariant tests in part (c) do not satisfy Definition 8.2.3.

- 10.39** a. Adapt the argument of Section 10.7, and the result of Exercise 10.30, to show that if we observe

$$X_{ij} \sim n(\theta_i, \sigma^2), \quad i = 1, \dots, p, \quad j = 1, \dots, n, \quad \sigma^2 \text{ known,}$$

all independent, and we compute  $\bar{X}_i = \frac{1}{n} \sum_j X_{ij}$ , then the estimator  $\bar{\mathbf{X}} = (\bar{X}_1, \dots, \bar{X}_p)$  is minimax.

b. Independently of part (a), show that if we observe  $X_{ij} \sim n(\theta_i, \sigma^2)$ ,  $i = 1, \dots, p$ , and  $j = 1, \dots, n$ , then, by sufficiency, we can reduce the problem to that considered in Section 10.7, that of  $n = 1$ . Hence, the estimator  $\bar{\mathbf{X}} = (\bar{X}_1, \dots, \bar{X}_p)$  is minimax.

- 10.40** The form of the Stein estimator of (10.7.5) can be justified somewhat by an *empirical Bayes* argument, given in Efron and Morris (1972). Such an argument was probably known by Stein, although he makes no mention of it. The empirical Bayes explanation is quite useful, especially in data analysis (see Efron and Morris, 1973, 1975; or Casella, 1985). Let  $X_i \sim n(\theta_i, 1)$ ,  $i = 1, \dots, p$ , and  $\theta_i$  be iid  $n(0, \tau^2)$ .

a. Show that the  $X_i$ 's, marginally, are iid  $n(0, \tau^2 + 1)$ , hence,  $\sum X_i^2 / (\tau^2 + 1) \sim \chi_p^2$ .

b. Using the marginal distribution, show that  $E(1 - ((p - 2)/\sum_{j=1}^p X_j^2)) = \tau^2 / (\tau^2 + 1)$  if  $p \geq 3$ . Thus, the  $i$ th component of the Stein estimator,

$$\delta_i^S(\mathbf{X}) = (1 - ((p - 2)/\sum_{j=1}^p X_j^2))X_i,$$

is an empirical Bayes version of the  $i$ th component of the Bayes estimator  $\delta_i^\pi(\mathbf{X}) = (\tau^2 / (\tau^2 + 1))X_i$ .

- 10.41** Consider the class of Stein estimators given by (10.7.7),

$$\delta_i^c(\mathbf{X}) = \left(1 - \frac{c}{\sum_{j=1}^p X_j^2}\right) X_i, \quad i = 1, \dots, p, \quad 0 < c < 2(p - 2).$$

Let  $X_i \sim n(\theta_i, 1)$ ,  $i = 1, \dots, p$ .

a. Using sum of squared errors loss, find an expression for the risk of  $\delta^c(\mathbf{X}) = (\delta_1^c(\mathbf{X}), \dots, \delta_p^c(\mathbf{X}))$ .

b. Show that for any constant  $c$  satisfying  $0 < c < 2(p - 2)$ ,  $\delta^c(\mathbf{X})$  is better than  $\mathbf{X}$ .

c. Compute the risk of  $\delta^c(\mathbf{X})$  at  $\theta = \mathbf{0}$  and find the value of  $c$  that minimizes this risk.

d. Show that the value of  $c = p - 2$  minimizes the risk within the class of estimators  $\{\delta^c(\mathbf{X}): 0 < c < 2(p - 2)\}$ .

- 10.42** Suppose we observe  $X_{ij} \sim n(\theta_i, \sigma^2)$ ,  $\sigma^2$  known,  $i = 1, \dots, p$ ,  $j = 1, \dots, n$ , all independent, and we form the Stein estimator

$$\delta_i^S(\bar{\mathbf{X}}) = \left(1 - \frac{(p-2)(\sigma^2/n)}{\sum_{j=1}^p \bar{X}_j^2}\right) \bar{X}_i, \quad i = 1, \dots, p,$$

- where  $\bar{X}_i = \frac{1}{n} \sum_j X_{ij}$  and  $\bar{\mathbf{X}} = (\bar{X}_1, \dots, \bar{X}_p)$ . Show that  $\delta^S(\bar{\mathbf{X}})$  is minimax.
- 10.43** For  $X_i \sim n(\theta_i, 1)$ ,  $i = 1, \dots, p$ , consider the class of Stein estimators that shrink toward an arbitrary point,

$$\delta_i^S(\mathbf{X}, \theta^0) = \theta_i^0 + \left(1 - \frac{c}{\sum_{j=1}^p (X_j - \theta_j^0)^2}\right) (X_i - \theta_i^0), \quad i = 1, \dots, p,$$

where  $\theta^0 = (\theta_1^0, \dots, \theta_p^0)$  is constant and  $0 < c < 2(p-2)$ .

- a. Show that under sum of squared errors loss,  $\delta^S(\mathbf{X}, \theta^0)$  dominates  $\mathbf{X}$  for any value of  $\theta^0$ .
- b. Show that the risk of  $\delta^S(\mathbf{X}, \theta^0)$  at  $\theta = \theta^0$  is the same as the ordinary Stein estimator (10.7.5) at  $\theta = 0$ .

## Miscellanea

---

### Game Theory

A topic closely related to decision theory is *game theory*, a formal mathematical study of games in which two or more players compete. The simplest type of game is a two-person zero-sum game. Player I picks a strategy  $a \in \mathcal{A}$ . Player II picks a strategy  $\theta \in \Theta$ . Then Player I pays Player II an amount  $L(\theta, a)$ , where negative values of  $L(\theta, a)$  correspond to payments from Player II to Player I and positive values of  $L(\theta, a)$  correspond to payments from Player I to Player II.

Player I may gain some information about what strategy Player II will use by observing a random variable  $X$  whose distribution depends on Player II's strategy  $\theta$ . Player I can use this information to decide what strategy  $a$  to use. From the notation we have used, the similarity between these elements of a game and the corresponding elements in a decision problem is evident. Minimax strategies make good sense in a game since there is an intelligent opponent. If Player II always knows what strategy Player I will use, then Player II can choose  $\theta$  to maximize Player I's expected losses. Player I, knowing that Player II will do this, should use a minimax strategy. Such a strategy will minimize Player I's maximum expected loss, the loss Player I knows he will incur if Player II plays in the way we have described.

As mentioned in Section 5, in statistical problems Nature is not considered to be an adversary and the minimaxity criterion is not so compelling. A classic treatment of game theory and decision theory is given by Blackwell and Girshick (1954), including the famous theorem by von Neumann on the existence of minimax strategies. A later reference is Thomas (1984).

### The Hunt–Stein Theorem

The Hunt–Stein Theorem is one of the great items of statistical folklore, as the original paper by Hunt and Stein was never published. However, the theorem due to them is quite real and represents one of the deepest results in mathematical statistics. The most readable (for statisticians) article about this theorem is by Bondar and Milnes (1981), with one of the most general developments given by Kiefer (1957). Lehmann (1986) discusses this theorem in the testing context.

Although the Hunt–Stein Theorem relies very heavily on group theoretic concepts, the general flavor of the theorem can be appreciated. For an invariant statistical problem, the Hunt–Stein Theorem shows that, under certain conditions on the group, the performance of *any* estimator can be equalled (or bettered) by the performance of an invariant estimator. The result can apply to point estimation, set estimation, and hypothesis testing. It is often paraphrased as saying that, under certain conditions on the group, the best invariant estimator is minimax.

### *Other Bayes Analyses*

1. *Robust Bayes Analysis* The fact that Bayes rules may be quite sensitive to the (subjective) choice of a prior distribution is a cause of concern for many Bayesian statisticians. The paper of Berger (1984) introduced the idea of a *robust Bayes analysis*. This is a Bayes analysis in which estimators are sought that have good properties for a range of prior distributions. For example, we might look for an estimator  $\delta^*$  whose Bayes risk is “close” to that of the Bayes rule  $\delta^\pi$ , for all priors  $\pi$  in a class  $\prod$ . Much of the work that has been done on this topic is summarized in Berger (1985).
2. *Empirical Bayes Analysis* In a standard Bayesian analysis, there are usually parameters in the prior distribution that are to be specified by the experimenter. For example, consider the specification

$$\begin{aligned} X|\theta &\sim n(\theta, 1), \\ \theta|\tau^2 &\sim n(0, \tau^2). \end{aligned}$$

The Bayesian experimenter would specify a prior value for  $\tau^2$  and a Bayesian analysis can be done. However, as seen in Exercise 10.40, the marginal distribution of  $X$ , which is  $n(0, \tau^2 + 1)$ , contains information about  $\tau$  and can be used to estimate  $\tau$ . This idea of *estimation of prior parameters from the marginal distribution* is what distinguishes empirical Bayes analysis. Empirical Bayes methods are useful in constructing improved procedures, as illustrated in Morris (1983) and Casella and Hwang (1987). Gianola and Fernando (1986) have successfully applied these types of methods to solve practical problems.

3. *Hierarchical Bayes Analysis* Another way of dealing with the specification above, without giving a prior value to  $\tau^2$ , is with a hierarchical specification, that is, a specification of a second-stage prior on  $\tau^2$ . For example, we could use

$$\begin{aligned} X|\theta &\sim n(\theta, 1), \\ \theta|\tau^2 &\sim n(0, \tau^2), \\ \tau^2 &\sim \text{uniform}(0, \infty) \text{ (improper prior)}. \end{aligned}$$

Hierarchical modeling, both Bayes and non-Bayes, is a very effective tool and usually gives answers that are reasonably robust to the underlying model. General formulas are given in Lindley and Smith (1972) and much subsequent work has been done by Smith. More recent application of this type of methodology can be found in Dempster et al. (1984).

In some cases, answers from a hierarchical analysis are quite similar to that obtained from an empirical Bayes analysis. In particular, when the second-stage prior is relatively flat when compared with the first-stage prior and the sample pdf, the answers from the two methods are close to one another. A discussion of the hierarchical model and its relationship to empirical Bayes estimation is given by Smith (1983).

# 11 The Analysis of Variance

*"If the fresh facts which come to our knowledge  
all fit themselves into the scheme, then our hypothesis  
may gradually become a solution."*

**Sherlock Holmes**  
*The Adventure of Wisteria Lodge*

## 11.1 Introduction

The analysis of variance (commonly referred to as the ANOVA) is one of the most widely used statistical techniques. A basic idea of the ANOVA, that of partitioning variation, is a fundamental idea of experimental statistics. The ANOVA belies its name in that it is not concerned with analyzing variances but rather with analyzing *variation in means*.

In its simplest form, the ANOVA is a method of estimating the means of several populations, populations often assumed to be normally distributed. The heart of the ANOVA, however, lies in the topic of statistical design. How can we get the most information on the most populations with the fewest observations? The ANOVA design question is not our major concern, however; we will be concerned with inference, that is, with estimation and testing, in the ANOVA.

Classical ANOVA had testing as its main goal—in particular, testing what is known as “the ANOVA null hypothesis.” But more recently, especially in the light of greater computing power, experimenters have realized that testing one hypothesis (a somewhat ludicrous one at that, as we shall see) does not make for good experimental inference. Thus, although we will derive the test of the ANOVA null, it is far from the most important part of an analysis of variance. More important is estimation, both point and interval. In particular, inference based on *contrasts* (to be defined) is of major importance.

We will study the two most common types of ANOVA, the oneway ANOVA and randomized complete block ANOVA. For a thorough treatment of the different facets of ANOVA designs, there is the classic text by Cochran and Cox (1957) or the more modern, but still classical, treatments by Kirk (1982) and Montgomery (1984). The text by Box, Hunter, and Hunter (1978) provides a guide to overall strategies in experimental statistics.

## 11.2 The Oneway Analysis of Variance

In the oneway analysis of variance (also known as the oneway classification) we assume that data,  $Y_{ij}$ , are observed according to a model

$$(11.2.1) \quad Y_{ij} = \theta_i + \epsilon_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, n_i,$$

where the  $\theta_i$  are unknown parameters and the  $\epsilon_{ij}$  are error random variables.

**Example 11.2.1:** Schematically, the data,  $y_{ij}$ , from a oneway ANOVA will look like this:

Treatments					
1	2	3	...	$k$	
$y_{11}$	$y_{21}$	$y_{31}$	...	$y_{k1}$	
$y_{12}$	$y_{22}$	$y_{32}$	...	$y_{k2}$	
$\vdots$	$\vdots$	$\vdots$	...	$y_{k3}$	
		$y_{3n_3}$		$\vdots$	
	$y_{1n_1}$				
		$y_{2n_2}$		$y_{kn_k}$	

Note that we do not assume that there are equal numbers of observations in each treatment group.

As an example, consider the following experiment performed to assess the relative effects, of three toxins and a control, on the liver of a certain species of trout. The data are the amount of deterioration (in standard units) of the liver of each sacrificed fish.

Toxin 1	Toxin 2	Toxin 3	Control
28	33	18	11
23	36	21	14
14	34	20	11
27	29	22	16
31		24	
34			

||

Without loss of generality we can assume that  $E\epsilon_{ij} = 0$ , since if not, we can rescale the  $\epsilon_{ij}$  and absorb the leftover mean into  $\theta_i$ . Thus it follows that

$$EY_{ij} = \theta_i, \quad j = 1, \dots, n_i,$$

so the  $\theta_i$ s are the means of the  $Y_{ij}$ s. The  $\theta_i$ s are usually referred to as *treatment means*, since the index often corresponds to different treatments or to *levels* of a particular treatment, such as dosage levels of a particular drug.

There is an alternative model to (11.2.1), sometimes called the *overparameterized model*, which can be written as

$$(11.2.2) \quad Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, n_i,$$

where, again,  $E\epsilon_{ij} = 0$ . It follows from this model that

$$EY_{ij} = \mu + \tau_i.$$

In this formulation we think of  $\mu$  as a grand mean, that is, the common mean level of the treatments. The parameters  $\tau_i$  then denote the unique effect due to treatment  $i$ , the deviation from the mean level that is caused by the treatment. However, we cannot estimate both  $\tau_i$  and  $\mu$  separately, because there are problems with *identifiability*.

**DEFINITION 11.2.1:** A parameter  $\theta$  for a family of distributions  $\{f(x|\theta) : \theta \in \Theta\}$  is *identifiable* if distinct values of  $\theta$  correspond to distinct pdfs or pmfs. That is, if  $\theta \neq \theta'$ , then  $f(x|\theta)$  is not the same function of  $x$  as  $f(x|\theta')$ .

Identifiability is a property of the model, not of an estimator or estimation procedure. However, if the model is not identifiable, then there is difficulty in doing inference. For example, if  $f(x|\theta) = f(x|\theta')$ , then observations from both distributions look exactly the same and we would have no way of knowing whether the true value of the parameter was  $\theta$  or  $\theta'$ . In particular, both  $\theta$  and  $\theta'$  would give the likelihood function the same value.

Realize that problems with identifiability can usually be solved by redefining the model. One reason that we have not encountered identifiability problems before is that our models have not only made intuitive sense but also were identifiable (for example, modeling a normal population in terms of its mean and variance). Here, however, we have a model, (11.2.2), that makes intuitive sense but is not identifiable. In Chapter 12 we will see a parameterization of the bivariate normal distribution that models a situation well, but is not identifiable.

In the parameterization of (11.2.2), there are  $k + 1$  parameters,  $(\mu, \tau_1, \dots, \tau_k)$ , but only  $k$  means,  $EY_{ij}, i = 1, \dots, k$ . Without any further restriction on the parameters, more than one set of values for  $(\mu, \tau_1, \dots, \tau_k)$  will lead to the same distribution. It is common in this model to add the restriction that  $\sum_{i=1}^k \tau_i = 0$ , which effectively reduces the number of parameters to  $k$  and makes the model identifiable. The restriction also has the effect of giving the  $\tau_i$ s an interpretation as deviations from an overall mean level. (See Exercise 11.4 and also the discussion after (11.3.3).)

For the oneway ANOVA the model (11.2.1), the *cell means model*, which has a more straightforward interpretation, is the one that we prefer to use. In more complicated ANOVAs, however, there is sometimes an interpretive advantage in model (11.2.2).

### 11.2.1 Model and Distribution Assumptions

Under model (11.2.1), a minimum assumption that is needed before any estimation can be done is that  $E\epsilon_{ij} = 0$  and  $\text{Var } \epsilon_{ij} < \infty$ , for all  $i, j$ . Under these assumptions, we can do some estimation of the  $\theta_i$ s (as in Exercise 7.39). However, to do any confidence interval estimation or testing, we need distributional assumptions. Here are the classic ANOVA assumptions.

## Oneway ANOVA assumptions

Random variables  $Y_{ij}$  are observed according to the model

$$Y_{ij} = \theta_i + \epsilon_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, n_i,$$

where

- i.  $E\epsilon_{ij} = 0$ ,  $\text{Var } \epsilon_{ij} = \sigma_i^2 < \infty$ , for all  $i, j$ .  $\text{Cov}(\epsilon_{ij}, \epsilon_{i'j'}) = 0$  for all  $i, i', j$ , and  $j'$  unless  $i = i'$  and  $j = j'$ .
- ii. The  $\epsilon_{ij}$  are independent and normally distributed (normal errors).
- iii.  $\sigma_i^2 = \sigma^2$  for all  $i$  (equality of variance, also known as *homoscedasticity*).

Without assumption (ii) we could only do point estimation and possibly look for estimators that minimize variance within a class, but we could not do interval estimation or testing. If we assume some distribution other than normal, intervals and tests can be quite difficult (but still possible) to derive. Of course, with reasonable sample sizes and populations that are not too asymmetric, we have the CLT to rely on.

The equality of variance assumption is also quite important. Interestingly, its importance is linked to the normality assumption. In general, if it is suspected that the data badly violate the ANOVA assumptions, a first course of attack is usually to try to transform the data, nonlinearly. This is done as an attempt to more closely satisfy the ANOVA assumptions, a generally easier alternative than finding another model for the untransformed data. A number of common transformations can be found in Snedecor and Cochran (1989); also see Exercises 11.1 and 11.2. (Other research on transformations has been concerned with the Box–Cox family of power transformations. See Exercise 11.3.)

The classic paper of Box (1954) shows that the robustness of the ANOVA to the assumption of normality depends on how equal the variances are (equal being better). The problem of estimating means when variances are unequal, known as the Behrens–Fisher problem, has a rich statistical history which can be traced back to Fisher (1935, 1939). A full account of the Behrens–Fisher problem can be found in Kendall and Stuart (1979).

For the remainder of this chapter we will do what is done in most of the experimental situations and we will assume that the three classic assumptions hold. If the data are such that transformations and the CLT are needed, we assume that such measures have been taken.

### 11.2.2 The Classic ANOVA Hypothesis

The classic ANOVA test is a test of the null hypothesis

$$H_0: \quad \theta_1 = \theta_2 = \cdots = \theta_k,$$

a hypothesis that, in many cases, is silly, uninteresting, and not true. An experimenter would not usually believe that the different treatments have *exactly* the same mean. More reasonably, an experiment is done to find out which treatments are better (for example, have a larger mean) and the real interest in the ANOVA is not in testing

but in estimation. (There are some specialized situations where there is interest in the ANOVA null in its own right. For example, see Section 11.3.5.) Most situations are like the following.

**Example 11.2.2:** The ANOVA evolved as a method of analyzing agricultural experiments. For example, in a study of the effect of various fertilizers on the zinc content of spinach plants ( $y_{ij}$ ), five treatments are investigated. Each treatment consists of a mixture of fertilizer material (magnesium, potassium, and zinc) and the data look like the layout of Example 11.2.1. The five treatments, in pounds per acre, are

Treatment	Magnesium	Potassium	Zinc
1	0	0	0
2	0	200	0
3	50	200	0
4	200	200	0
5	0	200	15

The classic ANOVA null hypothesis is really of no interest since the experimenter is sure that the different fertilizer mixtures have some different effects. The interest is in quantifying these effects. ||

We will spend some time with the ANOVA null, but mostly to use it as a means to an end. Recall the connection between testing and interval estimation established in Chapter 9. By using this connection, we can derive confidence regions by deriving, then inverting, appropriate tests (an easier route here).

The alternative to the ANOVA null is simply that the means are not all equal, that is, we test

$$(11.2.3) \quad H_0: \theta_1 = \theta_2 = \cdots = \theta_k \quad \text{versus} \quad H_1: \theta_i \neq \theta_j, \text{ for some } i, j.$$

Equivalently, we can specify  $H_1$  as  $H_1: \text{not } H_0$ . Realize that if  $H_0$  is rejected, we can conclude only that there is *some* difference in the  $\theta_i$ s, but we can make no inference as to where this difference might be. (Note that if  $H_1$  is accepted, we are *not* saying that all of the  $\theta_i$ s are different, merely that at least two are.)

One problem with the ANOVA hypotheses, a problem shared by many multivariate hypotheses, is that the interpretation of the hypotheses is not easy. What would be more useful, rather than concluding just that some  $\theta_i$ s are different, is a statistical description of the  $\theta_i$ s. Such a description can be obtained by breaking down the ANOVA hypotheses into smaller, more easily describable pieces.

We have already encountered methods for breaking down complicated hypotheses into smaller, more easily understood pieces—the union–intersection and intersection–union methods of Chapter 8. For the ANOVA, the union–intersection method is best suited as the ANOVA null is the intersection of more easily understood univariate hypotheses, hypotheses expressed in terms of *contrasts*. Furthermore, in

the cases we will consider, the resulting tests based on the union–intersection method are identical to LRTs (Exercise 11.13). Hence, they enjoy all of the properties of likelihood tests.

**DEFINITION 11.2.2:** Let  $t = (t_1, \dots, t_k)$  be a set of variables, either parameters or statistics, and let  $\mathbf{a} = (a_1, \dots, a_k)$  be known constants. The function

$$(11.2.4) \quad \sum_{i=1}^k a_i t_i$$

is called a *linear combination* of the  $t_i$ s. If, furthermore,  $\sum a_i = 0$ , it is called a *contrast*.

Contrasts are important because they can be used to compare treatment means. For example, if we have means  $\theta_1, \dots, \theta_k$  and constants  $\mathbf{a} = (1, -1, 0, \dots, 0)$ , then

$$\sum_{i=1}^k a_i \theta_i = \theta_1 - \theta_2$$

is a contrast that compares  $\theta_1$  to  $\theta_2$ . (See Exercise 11.10 for more about contrasts.)

The power of the union–intersection approach is increased understanding. The individual null hypotheses, of which the ANOVA null hypothesis is the intersection, are quite easy to visualize.

**THEOREM 11.2.1:** Let  $\theta = (\theta_1, \dots, \theta_k)$  be arbitrary parameters. Then

$$\theta_1 = \theta_2 = \dots = \theta_k \Leftrightarrow \sum_{i=1}^k a_i \theta_i = 0 \quad \text{for all } \mathbf{a} \in \mathcal{A}$$

where  $\mathcal{A}$  is the set of constants satisfying  $\mathcal{A} = \{\mathbf{a} = (a_1, \dots, a_k) : \sum a_i = 0\}$ , that is, all contrasts must satisfy  $\sum a_i \theta_i = 0$ .

*Proof:* If  $\theta_1 = \dots = \theta_k = \theta$  then

$$\sum_{i=1}^k a_i \theta_i = \sum_{i=1}^k a_i \theta = \theta \sum_{i=1}^k a_i = 0, \quad (\text{since } \mathbf{a} \text{ satisfies } \sum a_i = 0)$$

proving one implication ( $\Rightarrow$ ). To prove the other implication, consider the set of  $a_i \in \mathcal{A}$  given by

$$\mathbf{a}_1 = (1, -1, 0, \dots, 0), \quad \mathbf{a}_2 = (0, 1, -1, 0, \dots, 0), \quad \dots, \quad \mathbf{a}_{k-1} = (0, \dots, 0, 1, -1).$$

(The set  $(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{k-1})$  spans the elements of  $\mathcal{A}$ . That is, any  $\mathbf{a} \in \mathcal{A}$  can be written as a linear combination of  $(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{k-1})$ .) Forming contrasts with these  $\mathbf{a}_i$ s, we get that

$$\mathbf{a}_1 \Rightarrow \theta_1 = \theta_2, \quad \mathbf{a}_2 \Rightarrow \theta_2 = \theta_3, \quad \dots, \quad \mathbf{a}_{k-1} \Rightarrow \theta_{k-1} = \theta_k,$$

which, taken together, imply that  $\theta_1 = \dots = \theta_k$ , proving the theorem.  $\square$

It immediately follows from Theorem 11.2.1 that the ANOVA null can be expressed as a hypothesis about contrasts. That is, the null hypothesis is true if and only if the hypothesis

$$H_0: \sum_{i=1}^k a_i \theta_i = 0 \quad \text{for all } (a_1, \dots, a_k) \text{ such that } \sum_{i=1}^k a_i = 0$$

is true. Moreover, if  $H_0$  is false, we now know that there must be at least one nonzero contrast. That is, the ANOVA alternative,  $H_1$ : not all  $\theta_i$ s equal, is equivalent to the alternative

$$H_1: \sum_{i=1}^k a_i \theta_i \neq 0 \quad \text{for some } (a_1, \dots, a_k) \text{ such that } \sum_{i=1}^k a_i = 0.$$

Thus, we have gained in that the use of contrasts leaves us with hypotheses that are a little easier to understand and perhaps are a little easier to interpret. The real gain, however, is that the use of contrasts now allows us to think and operate in a univariate manner.

### 11.2.3 Inferences Regarding Linear Combinations of Means

Linear combinations, in particular, contrasts, play an extremely important role in the analysis of variance. Through understanding and analyzing the contrasts, we can make meaningful inferences about the  $\theta_i$ s. In the previous section we showed that the ANOVA null is really a statement about contrasts. In fact, most interesting inferences in an ANOVA can be expressed as contrasts or sets of contrasts. We start simply with inference about a single linear combination.

Working under the oneway ANOVA assumptions, we have that

$$Y_{ij} \sim n(\theta_i, \sigma^2), \quad i = 1, \dots, k, \quad j = 1, \dots, n_i.$$

Therefore,

$$\bar{Y}_{i \cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \sim n\left(\theta_i, \frac{\sigma^2}{n_i}\right), \quad i = 1, \dots, k.$$

*A note on notation:* It is a common convention that if a subscript is replaced by a  $\cdot$  (dot), it means that subscript has been summed over. Thus,  $Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}$  and  $Y_{\cdot j} = \sum_{i=1}^k Y_{ij}$ . The addition of a “bar” indicates that a mean is taken, as in  $\bar{Y}_{i\cdot}$  above. If both subscripts are summed over and the overall mean (called the *grand mean*) is calculated, we will break this rule to keep notation a little simpler and write  $\bar{\bar{Y}} = (1/N) \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$ , where  $N = \sum_{i=1}^k n_i$ .

For any set of constants  $a = (a_1, \dots, a_k)$ ,  $\sum_{i=1}^k a_i \bar{Y}_{i\cdot}$  is also normal (Exercise 11.7) with

$$E\left(\sum_{i=1}^k a_i \bar{Y}_{i\cdot}\right) = \sum_{i=1}^k a_i \theta_i,$$

and

$$\text{Var}\left(\sum_{i=1}^k a_i \bar{Y}_{i\cdot}\right) = \sigma^2 \sum_{i=1}^k \frac{a_i^2}{n_i},$$

and furthermore

$$\frac{\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - \sum_{i=1}^k a_i \theta_i}{\sqrt{\sigma^2 \sum_{i=1}^k a_i^2 / n_i}} \sim N(0, 1).$$

Although this is nice, we are usually in the situation of wanting to make inferences about the  $\theta_i$ s without knowledge of  $\sigma$ . Therefore, we want to replace  $\sigma$  with an estimate. In each population, if we denote the sample variance by  $S_i^2$ , that is,

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2, \quad i = 1, \dots, k,$$

then  $S_i^2$  is an estimate of  $\sigma^2$ , and  $(n_i - 1)S_i^2 / \sigma^2 \sim \chi_{n_i - 1}^2$ . Furthermore, under the ANOVA assumptions, since each  $S_i^2$  estimates the same  $\sigma^2$ , we can improve the estimators by combining them. We thus use the pooled estimator of  $\sigma^2$ ,  $S_p^2$ , given by

$$(11.2.5) \quad S_p^2 = \frac{1}{N - k} \sum_{i=1}^k (n_i - 1)S_i^2 = \frac{1}{N - k} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2.$$

Note that  $N - k = \sum(n_i - 1)$ . Since the  $S_i^2$ s are independent, Lemma 5.4.1 shows that  $(N - k)S_p^2 / \sigma^2 \sim \chi_{N-k}^2$ . Also,  $S_p^2$  is independent of each  $\bar{Y}_{i\cdot}$  (Exercise 11.8) and thus

$$(11.2.6) \quad \frac{\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - \sum_{i=1}^k a_i \theta_i}{\sqrt{S_p^2 \sum_{i=1}^k a_i^2 / n_i}} \sim t_{N-k},$$

Student's  $t$  with  $N - k$  degrees of freedom.

To test

$$H_0: \sum_{i=1}^k a_i \theta_i = 0 \quad \text{versus} \quad H_1: \sum_{i=1}^k a_i \theta_i \neq 0$$

at level  $\alpha$ , we would reject  $H_0$  if

$$(11.2.7) \quad \left| \frac{\sum_{i=1}^k a_i \bar{Y}_{i\cdot}}{\sqrt{S_p^2 \sum_{i=1}^k a_i^2 / n_i}} \right| > t_{N-k,\alpha/2}.$$

(Exercise 11.9 shows some other tests involving linear combinations.) Furthermore, (11.2.6) defines a pivot that can be inverted to give an interval estimator of  $\sum a_i \theta_i$ . With probability  $1 - \alpha$ ,

$$(11.2.8) \quad \sum_{i=1}^k a_i \bar{Y}_{i\cdot} - t_{N-k,\alpha/2} \sqrt{S_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}} \leq \sum_{i=1}^k a_i \theta_i \leq \sum_{i=1}^k a_i \bar{Y}_{i\cdot} + t_{N-k,\alpha/2} \sqrt{S_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}}.$$

**Example 11.2.3:** Special values of  $a$  will give particular tests or confidence intervals. For example, to compare treatments 1 and 2, take  $a = (1, -1, 0, \dots, 0)$ . Then, using (11.2.6), to test  $H_0: \theta_1 = \theta_2$  versus  $H_1: \theta_1 \neq \theta_2$ , we would reject  $H_0$  if

$$\left| \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \right| > t_{N-k,\alpha/2}.$$

Note, the difference between this test and the two-sample  $t$  test (Exercise 8.51) is that here information from treatments  $3, \dots, k$ , as well as treatments 1 and 2, is used to estimate  $\sigma^2$ .

Alternatively, to compare treatment 1 to the average of treatments 2 and 3 (for example, treatment 1 might be a control, 2 and 3 might be experimental treatments, and we are looking for some overall effect), we would take  $a = (1, -\frac{1}{2}, -\frac{1}{2}, 0, \dots, 0)$  and reject  $H_0: \theta_1 = \frac{1}{2}(\theta_2 + \theta_3)$  if

$$\left| \frac{\bar{Y}_{1\cdot} - \frac{1}{2}\bar{Y}_{2\cdot} - \frac{1}{2}\bar{Y}_{3\cdot}}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{4n_2} + \frac{1}{4n_3} \right)}} \right| > t_{N-k,\alpha/2}.$$

Using either (11.2.6) or (11.2.8), we have a way of testing or estimating any linear combination in the ANOVA. By judiciously choosing our linear combination we can learn much about the treatment means. For example, if we look at the contrasts

$\theta_1 - \theta_2, \theta_2 - \theta_3$ , and  $\theta_1 - \theta_3$ , we can learn something about the ordering of the  $\theta_i$ s. (Of course, we have to be careful of the overall  $\alpha$  level when doing a number of tests or intervals, but we can use the Bonferroni Inequality. See Example 11.2.4.)

We also must use some care in making formal conclusions from combinations of contrasts. Consider the hypotheses

$$H_0: \theta_1 = \frac{1}{2}(\theta_2 + \theta_3) \quad \text{versus} \quad H_1: \theta_1 < \frac{1}{2}(\theta_2 + \theta_3)$$

and

$$H_0: \theta_2 = \theta_3 \quad \text{versus} \quad H_1: \theta_2 < \theta_3.$$

If we reject both null hypotheses, we can conclude that  $\theta_3$  is bigger than both  $\theta_1$  and  $\theta_2$ , although we can make no formal conclusion about the ordering of  $\theta_2$  and  $\theta_1$  from these two tests. (See Exercise 11.10.) ||

Now we will use these univariate results about linear combinations and the relationship between the ANOVA null hypothesis and contrasts given in Theorem 11.2.1 to derive a test of the ANOVA null hypothesis.

### 11.2.4 The ANOVA $F$ Test

In the previous section we saw how to deal with single linear combinations and, in particular, contrasts in the ANOVA. Also, in Section 11.2.2, we saw that the ANOVA null hypothesis is equivalent to a hypothesis about contrasts. In this section we will use this equivalence, together with the union–intersection methodology of Chapter 8, to derive a test of the ANOVA hypothesis.

From Theorem 11.2.1, the ANOVA hypothesis test can be written

$$H_0: \sum_{i=1}^k a_i \theta_i = 0 \text{ for all } \mathbf{a} \in \mathcal{A} \quad \text{versus} \quad H_1: \sum_{i=1}^k a_i \theta_i \neq 0 \text{ for some } \mathbf{a} \in \mathcal{A},$$

where  $\mathcal{A} = \{\mathbf{a} = (a_1, \dots, a_k) : \sum_{i=1}^k a_i = 0\}$ . To see this more clearly as a union–intersection test define, for each  $\mathbf{a}$ , the set

$$\Theta_{\mathbf{a}} = \{\boldsymbol{\theta} = (\theta_1, \dots, \theta_k) : \sum_{i=1}^k a_i \theta_i = 0\}.$$

Then we have

$$\theta \in \{\boldsymbol{\theta} : \theta_1 = \theta_2 = \dots = \theta_k\} \Leftrightarrow \theta \in \Theta_{\mathbf{a}} \text{ for all } \mathbf{a} \in \mathcal{A} \Leftrightarrow \theta \in \bigcap_{\mathbf{a} \in \mathcal{A}} \Theta_{\mathbf{a}},$$

showing that the ANOVA null can be written as an intersection.

Now, recalling the union-intersection methodology from Section 8.2.4, we would reject  $H_0: \theta \in \cap_{\mathbf{a} \in \mathcal{A}} \Theta_{\mathbf{a}}$  (and, hence, the ANOVA null), if we can reject

$$H_{0_{\mathbf{a}}}: \theta \in \Theta_{\mathbf{a}} \quad \text{versus} \quad H_{1_{\mathbf{a}}}: \theta \notin \Theta_{\mathbf{a}}$$

for any  $\mathbf{a}$ . We test  $H_{0_{\mathbf{a}}}$  with the  $t$  statistic of (11.2.6),

$$(11.2.9) \quad T_{\mathbf{a}} = \left| \frac{\sum_{i=1}^k a_i \bar{Y}_i - \sum_{i=1}^k a_i \theta_i}{\sqrt{S_p^2 \sum_{i=1}^k a_i^2 / n_i}} \right|.$$

We then reject  $H_{0_{\mathbf{a}}}$  if  $T_{\mathbf{a}} > k$ , for some constant  $k$ . Using the union-intersection methodology, it follows that if we could reject for any  $\mathbf{a}$ , we could reject for the  $\mathbf{a}$  that maximizes  $T_{\mathbf{a}}$ . Thus, the union-intersection test of the ANOVA null is to reject  $H_0$  if  $\sup_{\mathbf{a}} T_{\mathbf{a}} > k$ , where  $k$  is chosen so that  $P_{H_0}(\sup_{\mathbf{a}} T_{\mathbf{a}} > k) = \alpha$ .

Calculation of  $\sup_{\mathbf{a}} T_{\mathbf{a}}$  is not straightforward, although with a little care it is not difficult. The calculation is that of a constrained maximum, similar to problems previously encountered (see, for example, Exercise 7.39, where a constrained minimum is calculated). We will attack the problem in a manner similar to what we have done previously and use the Cauchy-Schwarz Inequality. (Alternatively, a method such as Lagrange multipliers could be used, but then we would have to use second-order conditions to verify that a maximum has been found.)

Most of the technical maximization arguments will be given in the following lemma and the lemma will then be applied to obtain the supremum of  $T_{\mathbf{a}}$ . The lemma is just a statement about constrained maxima of quadratic functions. The proof of the lemma may be skipped by the fainthearted.

**LEMMA 11.2.1:** Let  $(v_1, \dots, v_k)$  be constants and let  $(c_1, \dots, c_k)$  be positive constants. Then, for  $\mathcal{A} = \{\mathbf{a} = (a_1, \dots, a_k) : \sum a_i = 0\}$ ,

$$(11.2.10) \quad \max_{\mathbf{a} \in \mathcal{A}} \left\{ \frac{\left( \sum_{i=1}^k a_i v_i \right)^2}{\sum_{i=1}^k a_i^2 / c_i} \right\} = \sum_{i=1}^k c_i (v_i - \bar{v}_c)^2,$$

where  $\bar{v}_c = \sum c_i v_i / \sum c_i$ . The maximum is attained at any  $\mathbf{a}$  of the form  $a_i = K c_i (v_i - \bar{v}_c)$  where  $K$  is a nonzero constant.

**Proof:** Define  $\mathcal{B} = \{\mathbf{b} = (b_1, \dots, b_k) : \sum b_i = 0 \text{ and } \sum b_i^2 / c_i = 1\}$ . For any  $\mathbf{a} \in \mathcal{A}$ , define  $\mathbf{b} = (b_1, \dots, b_k)$  by

$$b_i = \frac{a_i}{\sqrt{\sum_{i=1}^k a_i^2 / c_i}}$$

and note that  $\mathbf{b} \in \mathcal{B}$ . For any  $\mathbf{a} \in \mathcal{A}$ ,

$$\frac{\left(\sum_{i=1}^k a_i v_i\right)^2}{\sum_{i=1}^k a_i^2/c_i} = \left(\sum_{i=1}^k b_i v_i\right)^2.$$

We will find an upper bound on  $(\sum b_i v_i)^2$ , for  $\mathbf{b} \in \mathcal{B}$ , and then we will show that the maximizing  $\mathbf{a}$ , given in the lemma, achieves the upper bound.

Since we are dealing with the sum of products, the Cauchy–Schwarz Inequality (Section 4.7) is a natural thing to try, but we have to be careful to build in the constraints involving the  $c_i$ s. We can do this in the following way. Define  $C = \sum c_i$  and write

$$\frac{1}{C^2} \left(\sum_{i=1}^k b_i v_i\right)^2 = \left\{ \sum_{i=1}^k \left(\frac{b_i}{c_i}\right) (v_i) \left(\frac{c_i}{C}\right) \right\}^2.$$

This is the square of a *covariance* for a probability measure defined by the ratios  $c_i/C$ . Formally, if we define random variables  $B$  and  $V$  by

$$P\left(B = \frac{b_i}{c_i}, V = v_i\right) = \frac{c_i}{C}, \quad i = 1, \dots, k,$$

then  $EB = \sum (b_i/c_i)(c_i/C) = \sum b_i/C = 0$ . Thus,

$$\begin{aligned} & \left\{ \sum_{i=1}^k \left(\frac{b_i}{c_i}\right) (v_i) \left(\frac{c_i}{C}\right) \right\}^2 \\ &= (EBV)^2 \\ &= (\text{Cov}(B, V))^2 \quad (\text{EB} = 0) \\ &\leq (\text{Var } B)(\text{Var } V) \quad (\text{Cauchy–Schwarz Inequality}) \\ &= \left( \sum_{i=1}^k \left(\frac{b_i}{c_i}\right)^2 \left(\frac{c_i}{C}\right) \right) \left( \sum_{i=1}^k (v_i - \bar{v}_c)^2 \left(\frac{c_i}{C}\right) \right). \quad \left( \bar{v}_c = \frac{\sum c_i v_i}{\sum c_i} \right) \end{aligned}$$

Using the fact that  $\sum b_i^2/c_i = 1$  and cancelling common terms, we obtain

$$(11.2.11) \quad \left(\sum_{i=1}^k b_i v_i\right)^2 \leq \sum_{i=1}^k c_i (v_i - \bar{v}_c)^2, \quad \text{for any } \mathbf{b} \in \mathcal{B}.$$

Finally, we see that if  $a_i = Kc_i(v_i - \bar{v}_c)$ , for any nonzero constant  $K$ , then  $\mathbf{a} \in \mathcal{A}$  and

$$b_i = \frac{Kc_i(v_i - \bar{v}_c)}{\sqrt{\sum_{i=1}^k (Kc_i(v_i - \bar{v}_c))^2/c_i}} = \frac{c_i(v_i - \bar{v}_c)}{\sqrt{\sum_{i=1}^k c_i(v_i - \bar{v}_c)^2}}.$$

Since  $\sum c_i(v_i - \bar{v}_c) = 0$ ,

$$\begin{aligned} \sum_{i=1}^k b_i v_i &= \frac{\sum_{i=1}^k c_i(v_i - \bar{v}_c)v_i}{\sqrt{\sum_{i=1}^k c_i(v_i - \bar{v}_c)^2}} \\ &= \frac{\sum_{i=1}^k c_i(v_i - \bar{v}_c)^2}{\sqrt{\sum_{i=1}^k c_i(v_i - \bar{v}_c)^2}} = \sqrt{\sum_{i=1}^k c_i(v_i - \bar{v}_c)^2} \end{aligned}$$

and the inequality in (11.2.11) is an equality. Thus, the upper bound is attained and the function is maximized at such an  $a$ .  $\square$

Returning to  $T_a$  of (11.2.9), it should be clear that maximizing  $T_a$  is equivalent to maximizing  $T_a^2$ . We have

$$\begin{aligned} T_a^2 &= \frac{\left(\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - \sum_{i=1}^k a_i \theta_i\right)^2}{S_p^2 \sum_{i=1}^k a_i^2 / n_i}, \\ &= \frac{\left(\sum_{i=1}^k a_i \bar{U}_i\right)^2}{S_p^2 \sum_{i=1}^k a_i^2 / n_i}. \quad (\bar{U}_i = \bar{Y}_{i\cdot} - \theta_i) \end{aligned}$$

Noting that  $S_p^2$  has no effect on the maximization, we can apply Lemma 11.2.1 to the above expression to get the following theorem.

**THEOREM 11.2.2:** For  $T_a$  defined in expression (11.2.9),

$$(11.2.12) \quad \sup_{a: \sum a_i = 0} T_a^2 = \frac{\sum_{i=1}^k n_i \left( (\bar{Y}_{i\cdot} - \bar{\bar{Y}}) - (\theta_i - \bar{\theta}) \right)^2}{S_p^2},$$

where  $\bar{\bar{Y}} = \sum n_i \bar{Y}_{i\cdot} / \sum n_i$  and  $\bar{\theta} = \sum n_i \theta_i / \sum n_i$ . Furthermore, under the ANOVA assumptions

$$(11.2.13) \quad \sup_{a: \sum a_i = 0} T_a^2 \sim (k-1) F_{k-1, N-k},$$

that is,  $\frac{1}{k-1} \sup_{a: \sum a_i = 0} T_a^2$  has an  $F$  distribution with  $k-1$  and  $N-k$  degrees of freedom. (Recall that  $N = \sum n_i$ .)

*Proof:* To prove (11.2.12), use Lemma 11.2.1 and identify  $v_i$  with  $\bar{U}_i$  and  $c_i$  with  $n_i$ . The result is immediate.

To prove (11.2.13), we must show that the numerator and denominator of (11.2.12) are independent chi squared random variables, each divided by its degrees

of freedom. From the ANOVA assumptions two things follow. The numerator and denominator are independent and  $S_p^2 \sim \sigma^2 \chi_{N-k}^2 / (N - k)$ . A little work must be done to show that

$$\frac{1}{\sigma^2} \sum_{i=1}^k n_i \left( (\bar{Y}_{i\cdot} - \bar{\bar{Y}}) - (\theta_i - \bar{\theta}) \right)^2 \sim \chi_{k-1}^2.$$

This can be done, however, and is left as an exercise. (See Exercise 11.6.)  $\square$

If  $H_0: \theta_1 = \theta_2 = \dots = \theta_k$  is true,  $\theta_i = \bar{\theta}$  for all  $i = 1, \dots, k$  and the  $\theta_i - \bar{\theta}$  terms drop out of (11.2.12). Thus, for an  $\alpha$  level test of the ANOVA hypotheses

$$H_0: \theta_1 = \theta_2 = \dots = \theta_k \quad \text{versus} \quad H_1: \theta_i \neq \theta_j \text{ for some } i, j,$$

we reject  $H_0$  if

$$(11.2.14) \quad \frac{\sum_{i=1}^k n_i \left( (\bar{Y}_{i\cdot} - \bar{\bar{Y}}) \right)^2}{S_p^2} > (k-1)F_{k-1, N-k, \alpha}.$$

This rejection region is usually written as

$$\text{reject } H_0 \text{ if } F = \frac{\sum_{i=1}^k n_i \left( (\bar{Y}_{i\cdot} - \bar{\bar{Y}}) \right)^2 / (k-1)}{S_p^2} > F_{k-1, N-k, \alpha}$$

and the test statistic  $F$  is called the *ANOVA F statistic*.

### 11.2.5 Simultaneous Estimation of Contrasts

We have already seen how to estimate and test a single contrast in the ANOVA; the  $t$  statistic and interval are given in (11.2.6) and (11.2.8). However, in the ANOVA we are often in the position of wanting to make more than one inference and we know that the simultaneous inference from many  $\alpha$  level tests is not necessarily at level  $\alpha$ . In the context of the ANOVA this problem has already been mentioned.

**Example 11.2.4:** Many times there is interest in pairwise differences of means. Thus, if an ANOVA has means  $\theta_1, \dots, \theta_k$ , there may be interest in interval estimates of  $\theta_1 - \theta_2, \theta_2 - \theta_3, \theta_3 - \theta_4$ , etc. With the Bonferroni Inequality, we can build a simultaneous inference statement. Define

$$C_{ij} = \left\{ \theta_i - \theta_j : \theta_i - \theta_j \in \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm t_{N-k, \alpha/2} \sqrt{S_p^2 \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} \right\}.$$

Then  $P(C_{ij}) = 1 - \alpha$  for each  $C_{ij}$ , but, for example,  $P(C_{12} \text{ and } C_{23}) < 1 - \alpha$ . However, this last inference is the kind that we want to make in the ANOVA.

Recall the Bonferroni Inequality, given in expression (1.2.10), which states that for any sets  $A_1, \dots, A_n$ ,

$$P\left(\bigcap_{i=1}^n A_i\right) \geq \sum_{i=1}^n P(A_i) - (n-1).$$

In this case we want to bound  $P(\bigcap_{i,j} C_{ij})$ , the probability that all of the pairwise intervals cover their respective differences.

If we want to make a simultaneous  $1 - \alpha$  statement about the coverage of  $m$  confidence sets, then, from the Bonferroni Inequality, we can construct each confidence set to be of level  $\gamma$ , where  $\gamma$  satisfies

$$1 - \alpha = \sum_{i=1}^m \gamma - (m-1),$$

or, equivalently,

$$\gamma = 1 - \frac{\alpha}{m}.$$

A slight generalization is also possible in that it is not necessary to require each individual inference at the same level. We can construct each confidence set to be of level  $\gamma_i$ , where  $\gamma_i$  satisfies

$$1 - \alpha = \sum_{i=1}^m \gamma_i - (m-1).$$

In an ANOVA with  $k$  treatments, simultaneous inference on all  $k(k-1)/2$  pairwise differences can be made with confidence  $1 - \alpha$  if each  $t$  interval has confidence  $1 - 2\alpha/[k(k-1)]$ . ||

An alternative and quite elegant approach to simultaneous inference is given by Scheffé (1959). Scheffé's procedure, sometimes called the *S method*, allows for simultaneous confidence intervals (or tests) on *all* contrasts. (Exercise 11.15 shows that Scheffé's method can also be used to set up simultaneous intervals for any linear combination, not just for contrasts.) The procedure allows us to set a confidence coefficient that will be valid for *all contrast intervals simultaneously*, not just a specified group. The Scheffé procedure would be preferred if a large number of contrasts are to be examined. If the number of contrasts is small, the Bonferroni bound will almost certainly be smaller. (See the *Miscellanea* section for a discussion of other types of multiple comparison procedures.)

The proof that the Scheffé procedure has simultaneous  $1 - \alpha$  coverage on all contrasts follows easily from the union-intersection nature of the ANOVA test.

**THEOREM 11.2.3:** Under the ANOVA assumptions, if  $M = \sqrt{(k-1)F_{k-1, N-k, \alpha}}$ , then the probability is  $1 - \alpha$  that

$$\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - M \sqrt{S_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}} \leq \sum_{i=1}^k a_i \theta_i \leq \sum_{i=1}^k a_i \bar{Y}_{i\cdot} + M \sqrt{S_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}},$$

simultaneously for all  $\mathbf{a} \in \mathcal{A} = \{\mathbf{a} = (a_1, \dots, a_k) : \sum a_i = 0\}$ .

*Proof.* The simultaneous probability statement requires  $M$  to satisfy

$$P \left( \left| \sum_{i=1}^k a_i \bar{Y}_{i\cdot} - \sum_{i=1}^k a_i \theta_i \right| \leq M \sqrt{S_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}} \text{ for all } \mathbf{a} \in \mathcal{A} \right) = 1 - \alpha,$$

or, equivalently,

$$P(T_a^2 \leq M^2, \text{ for all } \mathbf{a} \in \mathcal{A}) = 1 - \alpha,$$

where  $T_a$  is defined in (11.2.9). However, since

$$P(T_a^2 \leq M^2, \text{ for all } \mathbf{a} \in \mathcal{A}) = P \left( \sup_{\mathbf{a}: \sum a_i = 0} T_a^2 \leq M^2 \right),$$

Theorem 11.2.2 shows that choosing  $M^2 = (k-1)F_{k-1, N-k, \alpha}$  satisfies the probability requirement.  $\square$

One of the real strengths of the Scheffé procedure is that it allows legitimate “data snooping.” That is, in classical statistics it is taboo to test hypotheses that have been suggested by the data, since this can bias the results and, hence, invalidate the inference. (We normally would not test  $H_0: \theta_1 = \theta_2$  just because we noticed that  $\bar{Y}_{1\cdot}$  was different from  $\bar{Y}_{2\cdot}$ . See Exercise 11.19.) However, with Scheffé’s procedure such a strategy is legitimate. The intervals or tests are valid for *all* contrasts. Whether they have been suggested by the data makes no difference. They already have been taken care of by the Scheffé procedure.

Of course, we must pay for all of the inferential power offered by the Scheffé procedure. The payment is in the form of the lengths of the intervals. In order to guarantee the simultaneous confidence level, the intervals may be quite long. For

example, it can be shown (Exercise 11.16) that in comparing the  $t$  and  $F$  distributions, for any  $\nu$ ,  $\alpha$ , and  $k$ , the cutoff points satisfy

$$t_{\nu, \alpha/2} \leq \sqrt{(k-1)F_{k-1, \nu, \alpha}}$$

and so the Scheffé intervals are always wider, sometimes much wider, than the single-contrast intervals (another argument in favor of the doctrine that nothing substitutes for careful planning and preparation in experimentation). The interval length phenomenon carries over to testing. It also follows from the above inequality that Scheffé tests are less powerful than  $t$  tests.

### 11.2.6 Partitioning Sums of Squares

The ANOVA provides a useful way of thinking about the way in which different treatments affect a measured variable—the idea of allocating variation to different sources. The basic idea of allocating variation can be summarized in the following identity.

**THEOREM 11.2.4:** For any numbers  $y_{ij}$ ,  $i = 1, \dots, k$ , and  $j = 1, \dots, n_i$ ,

$$(11.2.15) \quad \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^k n_i (\bar{y}_{i\cdot} - \bar{\bar{y}})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2,$$

where  $\bar{y}_{i\cdot} = \frac{1}{n_i} \sum_j y_{ij}$  and  $\bar{\bar{y}} = \sum_i n_i \bar{y}_{i\cdot} / \sum_i n_i$ .

*Proof:* The proof is quite simple and relies only on the fact that, when dealing with means, the cross-term often disappears. Write

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} ((y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \bar{\bar{y}}))^2,$$

expand the right-hand side and regroup terms. (See Exercise 11.23.) □

The sums in (11.2.15) are called *sums of squares* and are thought of as measuring variation in the data ascribable to different sources. (They are sometimes called *corrected sums of squares*, where the word *corrected* refers to the fact that a mean has been subtracted.) In particular, the terms in the oneway ANOVA model,

$$Y_{ij} = \theta_i + \epsilon_{ij},$$

are in one-to-one correspondence with the terms in (11.2.15). Equation (11.2.15) shows how to allocate variation to the treatments (variation *between* treatments) and to random error (variation *within* treatments). The left-hand side of (11.2.15) measures variation without regard to categorization by treatments, while the two terms on the

right-hand side measure variation due only to treatments and variation due only to random error, respectively. The fact that these sources of variation satisfy the above identity shows that the variation in the data, measured by sums of squares, is additive in the same way as the ANOVA model.

One reason why it is easier to deal with sums of squares is that, under normality, corrected sums of squares are chi squared random variables and we have already seen that independent chi squareds can be added to get new chi squareds.

Under the ANOVA assumptions, in particular if  $Y_{ij} \sim N(\theta_i, \sigma^2)$ , it is easy to show (Exercise 11.21) that

$$(11.2.16) \quad \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 \sim \chi_{N-k}^2,$$

since for each  $i = 1, \dots, k$ ,  $\frac{1}{\sigma^2} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 \sim \chi_{n_i-1}^2$ , all independent, and  $\sum_{i=1}^k \chi_{n_i-1}^2 \sim \chi_{N-k}^2$ . Furthermore, if  $\theta_i = \theta_j$  for every  $i, j$ , then

$$(11.2.17) \quad \frac{1}{\sigma^2} \sum_{i=1}^k n_i (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 \sim \chi_{k-1}^2 \quad \text{and} \quad \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{\bar{Y}})^2 \sim \chi_{N-1}^2.$$

Thus, under  $H_0 : \theta_1 = \dots = \theta_k$ , the sum of squares partitioning of (11.2.15) is a partitioning of chi squared random variables. When scaled, the left-hand side is distributed as a  $\chi_{N-1}^2$ , and the right-hand side is the sum of two independent random variables distributed, respectively, as  $\chi_{k-1}^2$  and  $\chi_{N-k}^2$ . Note that the  $\chi^2$  partitioning is true only if the terms on the right-hand side of (11.2.15) are independent, which follows in this case from the normality in the ANOVA assumptions. The partitioning of  $\chi^2$ 's does hold in a slightly more general context and a characterization of this is sometimes referred to as Cochran's Theorem. (See Searle (1971) and also the Miscellanea section.)

In general, it is possible to partition a sum of squares into sums of squares of uncorrelated contrasts, each with one degree of freedom. If the sum of squares has  $\nu$  degrees of freedom and is  $\chi_\nu^2$ , it is possible to partition it into  $\nu$  independent terms, each of which is  $\chi_1^2$ .

For a treatment contrast  $\sum a_i \bar{Y}_{i\cdot}$ , we define a *contrast sum of squares* as  $(\sum a_i \bar{Y}_{i\cdot})^2$ . In a oneway ANOVA it is always possible to find sets of constants  $\mathbf{a}^{(l)} = (a_1^{(l)}, \dots, a_k^{(l)})$ ,  $l = 1, \dots, k-1$ , to satisfy

$$(11.2.18) \quad \sum_{i=1}^k (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 = \left( \sum_{i=1}^k a_i^{(1)} \bar{Y}_{i\cdot} \right)^2 + \left( \sum_{i=1}^k a_i^{(2)} \bar{Y}_{i\cdot} \right)^2 + \dots + \left( \sum_{i=1}^k a_i^{(k-1)} \bar{Y}_{i\cdot} \right)^2$$

and

$$\sum_{i=1}^k \frac{a_i^{(l)} a_i^{(l')}}{n_i} = 0 \quad \text{for all } l \neq l'.$$

Thus, the individual contrast sums of squares are all uncorrelated, hence independent under normality (Lemma 5.4.2). When suitably normalized, the left-hand side of (11.2.18) is distributed as a  $\chi^2_{k-1}$  and the right-hand side is  $k-1$   $\chi^2_1$ s. (Such contrasts are called *orthogonal contrasts*. See Exercises 11.10 and 11.11 and the proof of Theorem 11.3.1.)

It is common to summarize the results of an ANOVA  $F$  test in a standard form, called an ANOVA table. The table also gives a number of useful, intermediate statistics. The headings should be self-explanatory.

ANOVA table for oneway classification

Source of variation	Degrees of freedom	Sum of squares	Mean square	$F$ statistic
Between treatment groups	$k - 1$	$SSB = \sum n_i (\bar{y}_i - \bar{\bar{y}})^2$	$MSB = SSB/(k - 1)$	$F = \frac{MSB}{MSW}$
Within treatment groups	$N - k$	$SSW = \sum \sum (y_{ij} - \bar{y}_{i\cdot})^2$	$MSW = SSW/(N - k)$	
Total	$N - 1$	$SST = \sum \sum (y_{ij} - \bar{\bar{y}})^2$		

**Example 11.2.1 (Continued):** The ANOVA table for the fish toxin data is

Source of variation	Degrees of freedom	Sum of squares	Mean square	$F$ statistic
Treatments	3	995.90	331.97	26.09
Within	15	190.83	12.72	
Total	18	1,186.73		

The  $F$  statistic of 26.09 is highly significant, showing that there is strong evidence the toxins produce different effects. ||

It follows from equation (11.2.15) that the sum of squares column "adds"—that is,  $SSB + SSW = SST$ . Similarly, the degrees of freedom column adds. The mean square column, however, does not as these are means rather than sums.

The ANOVA table contains no new statistics; it merely gives an orderly form for calculation and presentation. The  $F$  statistic is exactly the same as derived before

and, moreover, MSW is the usual, pooled, unbiased estimator of  $\sigma^2, S_p^2$  of (11.2.5) (Exercise 11.24).

### 11.3 Randomized Complete Block Designs

The previous section was concerned with a *oneway* classification of the data, that is, there was only one categorization (treatment) in the experiment. In general, the ANOVA allows for many types of categorization, not only to examine many different treatments simultaneously but also to take advantage of experimental situations. We now look at the simplest of the latter case, perhaps one of the most commonly used ANOVAs, the Randomized Complete Block (RCB) ANOVA.

There are a number of new terms in the name RCB and we will take them one at a time. First, a *block* (or *blocking factor*) is another type of categorization. The main difference between a block and a treatment is that a block is in an experiment for the express purpose of removing variation, not because there is any interest in finding block differences. The practice of blocking originated in agriculture, where experimenters took advantage of similar growing conditions to control experimental variances.

**Example 11.3.1:** A field experiment is conducted to study the adaptability of three varieties of strawberries to Venezuelan soil. The data are yields in kilograms from four blocks of land over a two-week period.

		Variety of strawberry		
		A	B	C
Block	1	6.3	10.1	8.4
	2	6.9	10.8	9.4
	3	5.3	9.8	9.0
	4	6.2	10.5	9.2

||

The blocks are called *complete* blocks if every treatment appears in every block, so that the data are in a rectangular array. We will assume that there is one observation for each treatment-block combination and that data are observed according to the additive model

$$(11.3.1) \quad Y_{ij} = \mu + \tau_i + b_j + \epsilon_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, r,$$

where  $\mu$  is an overall mean,  $\tau_i$  are treatment effects,  $b_j$  are block effects, and  $\epsilon_{ij}$  are error random variables. Notice that we are now using an overparameterized model, as discussed at (11.2.2) for the oneway ANOVA. As was previously mentioned, when there is more than one factor, the overparameterized model seems easier to understand.

Remember that with the overparameterized model the treatment and block effects represent deviations from an overall mean level.

**Example 11.3.2:** Schematically, the data,  $y_{ij}$ , from a RCB ANOVA will look like:

		Treatments				
		1	2	3	...	$k$
Blocks	1	$y_{11}$	$y_{21}$	$y_{31}$	...	$y_{k1}$
	2	$y_{12}$	$y_{22}$	$y_{32}$	...	$y_{k2}$
	:	:	:	:	:	:
	$r$	$y_{1r}$	$y_{2r}$	$y_{3r}$	...	$y_{kr}$

Note that there is only one observation for each treatment-block combination so, unlike the oneway ANOVA, there is really no replication. No two observations were taken under the same experimental conditions. It is possible to have replications in a RCB, that is, we can have two observations taken under the same experimental conditions. The inferences from such designs can address many questions that we choose not to go into here. But the essential features of RCB designs are found in this no-replication model. ||

We have now defined two of the three terms in a RCB ANOVA, *block* and *complete*. We turn to the term *randomized*. The meaning of the term *randomized* refers to the way that the observations are taken in each block. In each block, the treatments are run in a random manner, using a randomization restricted to take place within blocks. By way of contrast, the oneway ANOVA of Section 11.2 is a *completely randomized design*, since the observations are taken in a manner that is random throughout the data, with no blocks to restrict randomization. (See the discussion following equation (11.3.19).)

There is a further distinction between treatments and blocks. In the previous section, when dealing with the oneway ANOVA, all of the treatments (or levels of a treatment) of interest were included in the experiment. With blocks, however, all of the levels of interest are *not* in the experiment. For example, in Example 11.3.1, although we are interested in inferring about treatment differences regardless of the type of soil (block), we certainly cannot have all types of soil in the experiment. Therefore, we regard the blocks in the experiment as representative of a population of blocks (for example, all soil types). If the blocks are a random sample from a population of blocks, they are a particular case of a *random factor*.

**DEFINITION 11.3.1:** A *factor* is a variable defining a categorization, in particular in an ANOVA. A factor is a *fixed factor* if all the values of interest are included in the experiment. A factor is a *random factor* if all the values of interest are *not* included in the experiment and those that are can be considered to be randomly chosen from all the values of interest.

In an ANOVA, the treatments are always a fixed factor, as all of the levels of interest are in the experiment. By definition, blocks are always a random factor since not all levels of interest can be in the experiment. However, blocks are a special type of random factor. There is really no interest in blocks; they are there only because we know that the treatments will behave differently on different blocks. Fortunately, as we will see, all of this does not affect the treatment inferences, as long as we restrict inferences to treatment contrasts. This is actually quite startling. Whether we assume blocks to be fixed or random, confidence intervals and tests on treatments are unchanged. (We actually prove that the means and variances of treatment contrasts are independent of block effects, so confidence intervals and tests on treatments are unchanged for error distributions that are determined by their mean and variance, the normal being a particular case of this.)

Whether blocks are random does, however, have an effect on the block inferences. (See Scheffé (1959) and Searle (1971, 1987) for different sides of the block inference story, and Hocking (1973, 1985) or Samuels, Casella, and McCabe (1988) for some resolutions.)

### 11.3.1 Model and Distribution Assumptions

The RCB model given in (11.3.1) is, formally, a conditional (or hierarchical) model. If we assume that the blocks are random, then the actual block means in the experiment,  $\mathbf{b} = (b_1, \dots, b_r)$ , are a realization of a random variable,  $\mathbf{B} = (B_1, \dots, B_r)$ , but a realization that we do not observe. Conditionally, on  $\mathbf{B} = \mathbf{b}$ , we have the model

$$(11.3.2) \quad Y_{ij} | \mathbf{b} = \mu + \tau_i + b_j + \epsilon_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, r,$$

and unconditionally we have

$$(11.3.3) \quad Y_{ij} = \mu + \tau_i + B_j + \epsilon_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, r.$$

Notice the similarity to a Bayesian formulation. Model (11.3.2) is similar in spirit to a simple Bayes model and specifying a distribution on  $\mathbf{B}$  would be like specifying a prior distribution. When doing calculations we will be very clear as to whether we are conditioning on  $\mathbf{B} = \mathbf{b}$ . Note also that the model (11.3.3) is not identifiable (Definition 11.2.1). Fortunately, we are concerned only with treatment contrasts, so identifiability is not a problem.

We now state assumptions for the RCB model. We will not state them in utmost generality, but rather in a typical amount of generality. For example, as with the oneway ANOVA, point estimation can be done without a normality assumption, but we will not go into such details here.

#### **RCB ANOVA assumptions**

Random variables  $Y_{ij}$  are observed according to the model

$$Y_{ij} | \mathbf{b} = \mu + \tau_i + b_j + \epsilon_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, r,$$

where

- i. The random variables  $\epsilon_{ij} \sim \text{iid } n(0, \sigma^2)$  for  $i = 1, \dots, k$ , and  $j = 1, \dots, r$  (normal errors with equal variances).
- ii. The random variables  $B_1, \dots, B_r$ , whose realized (but unobserved) values are  $b_1, \dots, b_r$ , are iid  $n(0, \sigma_B^2)$  and are independent of  $\epsilon_{ij}$  for all  $i, j$ .

Of course, the distributional assumption on the blocks is of importance only when we are doing calculations unconditionally. Most of the time we will be operating conditionally on  $\mathbf{b}$  and this assumption will not be needed. (Note that assumption (i) here is a summarization of the three assumptions of the oneway ANOVA.)

The mean and variance of  $Y_{ij}$  are

$$E(Y_{ij}|\mathbf{b}) = \mu + \tau_i + b_j, \quad \text{Var}(Y_{ij}|\mathbf{b}) = \sigma^2 \quad (\text{conditionally})$$

and

$$EY_{ij} = \mu + \tau_i, \quad \text{Var } Y_{ij} = \sigma_B^2 + \sigma^2. \quad (\text{unconditionally})$$

(See Exercise 11.25 for details.)

Furthermore, conditional on  $\mathbf{b}$ , the  $Y_{ij}$ s are uncorrelated, since

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{i'j'}|\mathbf{b}) &= \text{Cov}(\mu + \tau_i + b_j + \epsilon_{ij}, \mu + \tau_{i'} + b_{j'} + \epsilon_{i'j'}|\mathbf{b}) \quad (\text{from (11.3.2)}) \\ &= \text{Cov}(\epsilon_{ij}, \epsilon_{i'j'}|\mathbf{b}) \quad (\text{property of covariance}) \\ &= 0. \quad (\epsilon_{ij}\text{s are uncorrelated}) \end{aligned}$$

(Note that we have used the independence of the  $\epsilon_{ij}$ s and the  $B_j$ s to get that the  $\epsilon_{ij}$ s are uncorrelated conditionally. If we assume only that  $B$  and  $\epsilon_{ij}$  were uncorrelated, then the fact that the  $\epsilon_{ij}$ s are uncorrelated conditionally does not necessarily follow.)

Interestingly, although the  $Y_{ij}$ s are uncorrelated conditionally, there is correlation in the blocks unconditionally. For  $Y_{ij}$  and  $Y_{i'j}$  in block  $j$ , with  $i \neq i'$ ,

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{i'j}) &= \text{Cov}(\mu + \tau_i + B_j + \epsilon_{ij}, \mu + \tau_{i'} + B_j + \epsilon_{i'j}) \quad (\text{from (11.3.3)}) \\ &= \text{Cov}(B_j + \epsilon_{ij}, B_j + \epsilon_{i'j}) \quad (\text{property of covariance}) \\ &= EB_j^2 \quad (\epsilon\text{s and } B\text{s are uncorrelated}) \\ &= \sigma_B^2, \quad (\text{definition, since } EB_j = 0) \end{aligned}$$

showing that not only does the model imply that there is correlation in the blocks, but there is positive correlation. This is a consequence of the additive model (11.3.3) and the assumption that the  $\epsilon$ s and  $B$ s are independent. In addition, the correlation between  $Y_{ij}$  and  $Y_{i'j}$  is

$$\frac{\text{Cov}(Y_{ij}, Y_{i'j})}{\sqrt{(\text{Var } Y_{ij})(\text{Var } Y_{i'j})}} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma^2},$$

a quantity called the *intraclass correlation*.

### 11.3.2 Treatment Contrasts

The most important inference from a RCB ANOVA concerns the treatments and, in particular, the estimation of contrasts between the treatments. In a manner similar to the oneway ANOVA, we will derive the distribution of treatment contrasts. Again, we will do all calculations conditional on  $\mathbf{b}$ .

Working under the RCB ANOVA assumptions we have that, conditionally,

$$(11.3.4) \quad Y_{ij} | \mathbf{b} \sim n(\mu + \tau_i + b_j, \sigma^2), \quad i = 1, \dots, k, \quad j = 1, \dots, r.$$

The parameter of interest is the treatment contrast  $\sum_{i=1}^k a_i \tau_i$ , whose estimator  $\sum_{i=1}^k a_i \bar{Y}_i$  is conditionally unbiased since

$$\begin{aligned} E\left(\sum_{i=1}^k a_i \bar{Y}_i | \mathbf{b}\right) &= E\left(\sum_{i=1}^k a_i \frac{1}{r} \sum_{j=1}^r Y_{ij} | \mathbf{b}\right) \\ &= \sum_{i=1}^k a_i \frac{1}{r} \sum_{j=1}^r (\mu + \tau_i + b_j) \quad (\text{using (11.3.2)}) \\ (11.3.5) \quad &= \sum_{i=1}^k a_i (\mu + \tau_i + \bar{b}) \quad (\bar{b} = \sum b_j / r) \\ &= \sum_{i=1}^k a_i \tau_i. \quad (\sum a_i = 0) \end{aligned}$$

Thus conditionally and, hence, unconditionally  $\sum a_i \bar{Y}_i$  is an unbiased estimator of the contrast  $\sum a_i \tau_i$ .

Similarly, since the  $Y_{ij}$ s are conditionally uncorrelated, it is easy to show

$$(11.3.6) \quad \text{Var}\left(\sum_{i=1}^k a_i \bar{Y}_i | \mathbf{b}\right) = \frac{\sigma^2}{r} \sum_{i=1}^k a_i^2.$$

Since the conditional variance and the conditional mean are both free of  $\mathbf{b}$ , it follows (Exercise 11.26) that the conditional variance is also the unconditional variance.

Since, conditional on  $\mathbf{b}$  the  $Y_{ij}$ s are normal, from (11.3.5) and (11.3.6) we have

$$\sum_{i=1}^k a_i \bar{Y}_i \sim n\left(\sum_{i=1}^k a_i \tau_i, \frac{\sigma^2}{r} \sum_{i=1}^k a_i^2\right).$$

and, therefore,

$$(11.3.7) \quad \frac{\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - \sum_{i=1}^k a_i \tau_i}{\sqrt{\frac{\sigma^2}{r} \sum_{i=1}^k a_i^2}} \sim n(0, 1).$$

This shows that, under normality, not only the conditional mean and variance, but the entire conditional distribution of  $\sum_{i=1}^k a_i \bar{Y}_{i\cdot}$  is free of  $b$ . Thus the conditional distribution is the unconditional distribution and inferences from this conditional distribution are not dependent on  $b$ .

Note that the assumption of normality implies that the mean and variance determine the distribution; hence if they are free of  $b$  then the entire distribution is free of  $b$ . If the  $Y_{ij}$ 's are not normal, it does not necessarily follow that the mean and variance determine the distribution, so it will not necessarily follow that the conditional distribution equals the unconditional distribution. Thus, one very important implication of the normality assumption is that it allows us to do all calculations conditional on blocks and have these calculations be valid unconditionally.

As in Section 11.2, (11.3.7) is not of practical value unless  $\sigma^2$  is known. Estimation of  $\sigma^2$  in the RCB poses a slightly different problem than in the oneway ANOVA. Here, we have no replication, so we must depend on the model for our estimate of  $\sigma^2$ . To do this, the assumption of the additive model will be used extensively.

Under the RCB assumptions,  $\sigma^2$  is the variance of  $\epsilon_{ij}$ , which can be written as

$$\epsilon_{ij} = Y_{ij} - (\mu + \tau_i + b_j),$$

and we can approximate  $\epsilon_{ij}$  by substituting estimates for  $\mu$ ,  $\tau_i$ , and  $b_j$ ,

$$(11.3.8) \quad \hat{\epsilon}_{ij} = Y_{ij} - (\hat{\mu} + \hat{\tau}_i + \hat{b}_j),$$

forming the *residuals* from the model.

The residual from a model is what is left after the effects have been estimated. By their very definition, the residuals are model-dependent and this represents a very big difference from the oneway ANOVA. There, the residuals,  $(Y_{ij} - \bar{Y}_{i\cdot})$ , came from within a treatment and could be used to estimate  $\sigma^2$  even if the model (11.2.1) was incorrect. However, we have no such luxury here; our estimate of  $\sigma^2$  is model-dependent.

To estimate  $\mu$ ,  $\tau_i$ , and  $b_j$  we can use the respective means. Remember that, in an overparameterized model such as (11.3.2) the parameters  $\tau_i$  and  $b_j$  represent deviations from a mean level, so we have

$$\begin{aligned} \hat{\mu} &= \bar{\bar{Y}}, \\ \hat{\tau}_i &= \bar{Y}_{i\cdot} - \bar{\bar{Y}}, \\ \hat{b}_j &= \bar{Y}_{\cdot j} - \bar{\bar{Y}}. \end{aligned}$$

Using (11.3.8), this gives

$$(11.3.9) \quad \hat{\epsilon}_{ij} = Y_{ij} - \left( \bar{\bar{Y}} + (\bar{Y}_{i\cdot} - \bar{\bar{Y}}) + (\bar{Y}_{\cdot j} - \bar{\bar{Y}}) \right)$$

$$= Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{\bar{Y}}.$$

It is straightforward to calculate (Exercise 11.27) that

$$\mathbb{E}(\hat{\epsilon}_{ij} | \mathbf{b}) = 0,$$

$$\text{Var}(\hat{\epsilon}_{ij} | \mathbf{b}) = \mathbb{E}(\hat{\epsilon}_{ij}^2 | \mathbf{b}) = \left( \frac{r-1}{r} \right) \left( \frac{k-1}{k} \right) \sigma^2.$$

Therefore, from the properties of expected values,

$$\mathbb{E} \left( \frac{1}{(r-1)(k-1)} \sum_{i=1}^k \sum_{j=1}^r \hat{\epsilon}_{ij}^2 | \mathbf{b} \right) = \sigma^2,$$

and the statistic

$$S_R^2 = \frac{1}{(r-1)(k-1)} \sum_{i=1}^k \sum_{j=1}^r \hat{\epsilon}_{ij}^2$$

is an unbiased estimator of  $\sigma^2$ . We use the subscript R in  $S_R^2$  to remind us that this estimate is based on model-dependent residuals.

Now, substituting for  $\sigma^2$  in (11.3.7), we would hope that

$$(11.3.10) \quad \frac{\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - \sum_{i=1}^k a_i \tau_i}{\sqrt{\frac{S_R^2}{r} \sum_{i=1}^k a_i^2}} \sim t_{(r-1)(k-1)}.$$

We know that the degrees of freedom of  $S_R^2$  are  $(r-1)(k-1)$  because of the constraints on the  $\hat{\epsilon}_{ij}$ s (see Exercise 11.29), so if (11.3.10) is distributed as a  $t$ , it must have  $(r-1)(k-1)$  degrees of freedom.

From the definition of the  $t$  distribution, to show that (11.3.10) is distributed as Student's  $t$  we must show two things:

- i.  $S_R^2$  is independent of  $\sum_{i=1}^k a_i \bar{Y}_{i\cdot}$ .
- ii.  $(r-1)(k-1)S_R^2/\sigma^2 \sim \chi^2_{(r-1)(k-1)}$ .

To show that  $S_R^2$  is independent of  $\sum_{i=1}^k a_i \bar{Y}_{i\cdot}$ , under normality it is sufficient (by Lemma 5.4.2) to show that

$$(11.3.11) \quad \text{Cov}(\hat{\epsilon}_{i'j'}, \bar{Y}_{i\cdot} | \mathbf{b}) = 0 \quad \text{for every } i, i', j'.$$

To establish (11.3.11), write

$$\begin{aligned}
 & \text{Cov}(\hat{\epsilon}_{i'j'}, \bar{Y}_{i\cdot} | \mathbf{b}) \\
 &= E(\hat{\epsilon}_{i'j'} \bar{Y}_{i\cdot} | \mathbf{b}) && (\text{since } E(\hat{\epsilon}_{i'j'} | \mathbf{b}) = 0) \\
 &= E((Y_{i'j'} - \bar{Y}_{i'\cdot} - \bar{Y}_{\cdot j'} + \bar{\bar{Y}})\bar{Y}_{i\cdot} | \mathbf{b}) \\
 &= E((Y_{i'j'} - \bar{Y}_{i'\cdot})\bar{Y}_{i\cdot} | \mathbf{b}) - E((\bar{Y}_{\cdot j'} - \bar{\bar{Y}})\bar{Y}_{i\cdot} | \mathbf{b}) \\
 &= ((\mu + \tau_i + \bar{b})(b_{j'} - \bar{b})) - ((\mu + \tau_i + \bar{b})(b_{j'} - \bar{b})) \quad (\text{Exercise 11.30}) \\
 &= 0.
 \end{aligned}$$

Thus,  $S_R^2$  is independent of  $\sum_{i=1}^k a_i \bar{Y}_{i\cdot}$ . It only remains to show that  $S_R^2$  is distributed as a multiple of a chi squared random variable. Here we run into a problem. Since the  $\hat{\epsilon}_{ij}$ s are correlated, our typical chi squared induction proof (as in Lemma 5.4.1) does not work in a straightforward way and we have to rely on more advanced means. A straightforward proof will work, however, if either  $k = 2$  or  $r = 2$ . (See Exercise 11.31.)

Cochran's Theorem, which is discussed in the *Miscellanea* section, can be applied to  $S_R^2$  and tells us that  $(r-1)(k-1)S_R^2/\sigma^2 \sim \chi_{(r-1)(k-1)}^2$ . We can also prove this fact directly, using a partitioning of the sum of squares as was discussed in Section 11.2.6. The proof of the following theorem, like that of Lemma 11.2.1, is not for the fainthearted.

**THEOREM 11.3.1:** Under the assumptions of the RCB ANOVA,

$$(r-1)(k-1)S_R^2/\sigma^2 \sim \chi_{(r-1)(k-1)}^2.$$

*Proof:* We will show that

$$(r-1)(k-1) \frac{S_R^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^r \hat{\epsilon}_{ij}^2$$

can be written as the sum of  $(r-1)(k-1)$  terms, each of which is distributed as a  $\chi_1^2$ , and all of them independent.

We know from Exercise 11.29 that  $\sum_{i=1}^k \sum_{j=1}^r \hat{\epsilon}_{ij}^2$  has  $(r-1)(k-1)$  degrees of freedom, so we expect, as in the partitioning in Section 11.2.6, that the sum can be written as the sum of squares of  $(r-1)(k-1)$  independent contrasts.

A contrast among the  $\hat{\epsilon}_{ij}$ s is a linear combination  $\sum_{i=1}^k \sum_{j=1}^r a_{ij} \hat{\epsilon}_{ij}$ , where the  $a_{ij}$ s satisfy  $\sum_{i=1}^k \sum_{j=1}^r a_{ij} = 0$ ,  $\sum_{j=1}^r a_{ij} = 0$  for each  $i$  and  $\sum_{i=1}^k a_{ij} = 0$  for each  $j$ . For two contrasts  $\sum_{i=1}^k \sum_{j=1}^r a_{ij} \hat{\epsilon}_{ij}$  and  $\sum_{i=1}^k \sum_{j=1}^r c_{ij} \hat{\epsilon}_{ij}$  to be uncorrelated (hence independent under normality), we must also have  $\sum_{i=1}^k \sum_{j=1}^r a_{ij} c_{ij} = 0$  (see Exercise 11.32). We will construct  $(r-1)(k-1)$  sets of constants whose corresponding contrasts are all independent, are each distributed as a  $\chi_1^2$ , and sum to  $\frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^r \hat{\epsilon}_{ij}^2$ .

We will construct our contrasts using constants that we get from a "product-type" construction. We start with the  $k-1$  sets of contrasts defined by the rows

	1	2	3	4	...	$k - 1$	$k$
1	$k - 1$	-1	-1	-1	...	-1	-1
2	0	$k - 2$	-1	-1	...	-1	-1
3	0	0	$k - 3$	-1	...	-1	-1
:	:	:	:	:	⋮	-1	-1
$k - 1$	0	0	0	0	...	1	-1

Each row sums to zero so each row can be used to define a contrast between treatments.  
Similarly, we also define the  $r - 1$  sets of contrasts by the columns

	1	2	3	...	$r - 1$
1	$r - 1$	0	0	...	0
2	-1	$r - 2$	0	...	0
3	-1	-1	$r - 3$	...	0
4	-1	-1	-1	...	0
:	⋮	⋮	⋮	⋮	⋮
$r - 1$	-1	-1	-1	...	1
$r$	-1	-1	-1	...	-1

Here, every column sums to zero and defines a contrast. We now construct our set of uncorrelated contrasts by taking products. For example, from row  $l$  and column  $m$ , we form a rectangular array of constants. Write row  $l$  across the top and column  $m$  down the left side. Now an element in the array is the product of the value on the top line and the value in the left column. From row  $l$  and column  $m$  we obtain

	0	0	0	...	0	$k - l$	-1	...	-1
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
$r - m$	0	0	0	...	0	$(r - m)(k - l)$	$-(r - m)$	...	$-(r - m)$
-1	.	.	.	...	0	$-(k - l)$	1	...	1
.	.	.	.	...	0	.	1	...	1
-1	0	0	0	...	0	$-(k - l)$	1	...	1

If we denote this set of constants by  $a_{ij}^{lm}$  (normalized so that  $\sum_{i,j} (a_{ij}^{lm})^2 = 1$ ) and define  $T_{lm} = \sum_{i=1}^k \sum_{j=1}^r a_{ij}^{lm} \hat{\epsilon}_{ij}$ , it can be shown (using Exercise 11.32) that

- i. For each  $l, m, T_{lm} \sim N(0, \sigma^2)$ .
- ii. For each  $l, m, l', m', l \neq l', m \neq m'$ ,  $\text{Cov}(T_{lm}, T_{l'm'}) = 0$ .

Furthermore, a tedious amount of algebra will verify that

$$\frac{1}{\sigma^2} \sum_{l=1}^{k-1} \sum_{m=1}^{r-1} (T_{lm})^2 = \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^r \hat{\epsilon}_{ij}^2.$$

Note that, by construction, there are only  $(r-1)(k-1)$  quantities  $T_{lm}$ .

It follows from (i) and (ii) that the above sum, and hence the residual sum of squares, is distributed  $\chi^2_{(r-1)(k-1)}$ .  $\square$

We now have that  $S_R^2$  is independent of  $\sum a_i \bar{Y}_i$  and  $(r-1)(k-1)S_R^2/\sigma^2 \sim \chi^2_{(r-1)(k-1)}$ . Putting these results together, we see that (11.3.10) defines a  $t$  random variable. To test

$$H_0: \sum_{i=1}^k a_i \tau_i = 0 \quad \text{versus} \quad H_1: \sum_{i=1}^k a_i \tau_i \neq 0$$

at level  $\alpha$ , we

$$(11.3.12) \quad \text{reject } H_0 \text{ if } \left| \frac{\sum_{i=1}^k a_i \bar{Y}_i}{\sqrt{\left(\frac{S_R^2}{r}\right) \sum_{i=1}^k a_i^2}} \right| > t_{(r-1)(k-1), \alpha/2}.$$

More importantly, we get an interval estimator of  $\sum a_i \tau_i$ . With probability  $1 - \alpha$ ,

$$(11.3.13) \quad \begin{aligned} \sum_{i=1}^k a_i \bar{Y}_i - t_{(r-1)(k-1), \alpha/2} \sqrt{\frac{S_R^2}{r} \sum_{i=1}^k a_i^2} &\leq \sum_{i=1}^k a_i \tau_i \\ &\leq \sum_{i=1}^k a_i \bar{Y}_i + t_{(r-1)(k-1), \alpha/2} \sqrt{\frac{S_R^2}{r} \sum_{i=1}^k a_i^2}. \end{aligned}$$

### 11.3.3 Simultaneous Estimation and Testing

Having now successfully dealt with treatment contrasts in the RCB ANOVA, we can use the methodology of Sections 11.2.4 and 11.2.5, which will directly carry over to

this case, to look at simultaneous inference. In particular, Lemma 11.2.1 can be used to get overall tests and simultaneous intervals.

As in Section 11.2.5, we can use the union–intersection method to get a test of

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_k \quad \text{versus} \quad H_1: \tau_i \neq \tau_j \text{ for some } i, j,$$

based on the supremum of the statistic

$$(11.3.14) \quad T_a = \left| \frac{\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - \sum_{i=1}^k a_i \tau_i}{\sqrt{\frac{S_R^2}{r} \sum_{i=1}^k a_i^2}} \right|,$$

where the supremum is over all  $\mathbf{a} = (a_1, \dots, a_k)$  such that  $\sum a_i = 0$ . The supremum can be calculated using Lemma 11.2.1.

**THEOREM 11.3.2:** For  $T_a$  defined in expression (11.3.14)

$$(11.3.15) \quad \sup_{\mathbf{a}: \sum a_i = 0} T_a^2 = \frac{\sum_{i=1}^k r \left( (\bar{Y}_{i\cdot} - \bar{\bar{Y}}) - (\tau_i - \bar{\tau}) \right)^2}{S_R^2},$$

where  $\bar{\bar{Y}} = \sum \bar{Y}_{i\cdot} / k$  and  $\bar{\tau} = \sum \tau_i / k$ . Furthermore, under the RCB ANOVA assumptions

$$(11.3.16) \quad \sup_{\mathbf{a}: \sum a_i = 0} T_a^2 \sim (k-1) F_{k-1, (r-1)(k-1)}.$$

*Proof:* Equation (11.3.15) is established by a direct application of Lemma 11.2.1, and details are omitted. To prove (11.3.16), we must show that the numerator and denominator are independent chi squared random variables, each divided by its degrees of freedom. Using the fact that

$$\text{Cov}(\bar{Y}_{i\cdot} - \bar{\bar{Y}}, Y_{i'j'} - \bar{Y}_{\cdot j'} - \bar{Y}_{i\cdot} + \bar{\bar{Y}}) = 0,$$

Lemma 5.4.2 can be used to show that the numerator and denominator are independent. Finally, we can show that

$$\sum_{i=1}^k r \left( (\bar{Y}_{i\cdot} - \bar{\bar{Y}}) - (\tau_i - \bar{\tau}) \right)^2 \sim \chi_{k-1}^2,$$

and the result follows. (See Exercise 11.34 for details.) □

In summary, for an  $\alpha$  level test of the RCB ANOVA hypotheses

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_k \quad \text{versus} \quad H_1: \tau_i \neq \tau_j \text{ for some } i, j,$$

we reject  $H_0$  if

$$\frac{\sum_{i=1}^k r \left( (\bar{Y}_{i\cdot} - \bar{\bar{Y}}) \right)^2}{S_R^2} > (k-1)F_{k-1,(r-1)(k-1),\alpha},$$

or, equivalently, if

$$F = \frac{\sum_{i=1}^k r \left( (\bar{Y}_{i\cdot} - \bar{\bar{Y}}) \right)^2 / (k-1)}{S_R^2} > F_{k-1,(r-1)(k-1),\alpha}.$$

Since we again built the  $F$  test from contrasts, simultaneous interval estimation using the Scheffé procedure directly follows.

**THEOREM 11.3.3:** Under the RCB ANOVA assumptions, if  $M = \sqrt{(k-1)F_{k-1,(r-1)(k-1),\alpha}}$ , then the probability is  $1-\alpha$  that

$$\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - M \sqrt{\frac{S_R^2}{r} \sum_{i=1}^k a_i^2} \leq \sum_{i=1}^k a_i \tau_i \leq \sum_{i=1}^k a_i \bar{Y}_{i\cdot} + M \sqrt{\frac{S_R^2}{r} \sum_{i=1}^k a_i^2},$$

simultaneously for all  $a = (a_1, \dots, a_k)$  such that  $\sum a_i = 0$ .  $\square$

In fact, any simultaneous inference procedure that would work in the oneway ANOVA can be adapted to work for treatment contrasts in the RCB ANOVA.

### 11.3.4 Partitioning Sums of Squares

The RCB ANOVA can also be summarized in an ANOVA table using the following identity, which provides a partitioning of the total sum of squares analogous to that in Theorem 11.2.4.

**THEOREM 11.3.4:** For any numbers  $y_{ij}$ ,  $i = 1, \dots, k$  and  $j = 1, \dots, r$ ,

$$(11.3.17) \quad \begin{aligned} \sum_{i=1}^k \sum_{j=1}^r (y_{ij} - \bar{y})^2 &= \sum_{i=1}^k r(\bar{y}_{i\cdot} - \bar{\bar{y}})^2 + \sum_{j=1}^r k(\bar{y}_{\cdot j} - \bar{\bar{y}})^2 \\ &\quad + \sum_{i=1}^k \sum_{j=1}^r (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{\bar{y}})^2 \end{aligned}$$

where  $\bar{y}_{i\cdot} = \frac{1}{r} \sum_j y_{ij}$ ,  $\bar{y}_{\cdot j} = \frac{1}{k} \sum_i y_{ij}$ , and  $\bar{\bar{y}} = \sum \sum y_{ij}/(rk)$ .

*Proof:* The proof is similar to that of Theorem 11.2.4 and represents a further partitioning of the data. This case is slightly easier since here we do not allow unequal  $n_i$ . Apply Theorem 11.2.4 to the left-hand side of (11.3.17) to get

$$\sum_{i=1}^k \sum_{j=1}^r (y_{ij} - \bar{y})^2 = \sum_{i=1}^k r(\bar{y}_{i\cdot} - \bar{y})^2 + \sum_{i=1}^k \sum_{j=1}^r (y_{ij} - \bar{y}_{i\cdot})^2.$$

Now we must show that

$$(11.3.18) \quad \sum_{i=1}^k \sum_{j=1}^r (y_{ij} - \bar{y}_{i\cdot})^2 = \sum_{j=1}^r k(\bar{y}_{\cdot j} - \bar{y})^2 + \sum_{i=1}^k \sum_{j=1}^r (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y})^2.$$

As before, by adding  $\pm(\bar{y}_{\cdot j} - \bar{y})$  inside the square on the left-hand side of (11.3.18) and expanding, we find that the cross-term is zero and (11.3.17) is established (Exercise 11.35).  $\square$

As with the oneway ANOVA, the partitioning of the RCB sums of squares follows the model

$$Y_{ij} = \mu + \tau_i + b_j + \epsilon_{ij},$$

where the right-hand side of equation (11.3.17) measures variation due to treatments, variation due to blocks, and variation due to residual error, respectively. Again, as in the oneway ANOVA, the above identity shows that, when measured by sums of squares, the variation in the data is additive in the same way as the RCB ANOVA model.

The RCB ANOVA  $F$  test can be summarized in the following way.

RCB ANOVA table

Source of variation	Degrees of freedom	Sum of squares	Mean square	$F$ statistic
Blocks	$r - 1$	$SSBI = \sum k(\bar{y}_{\cdot j} - \bar{y})^2$		
Between treatment groups	$k - 1$	$SSB = \sum r(\bar{y}_{i\cdot} - \bar{y})^2$	$MSB = SSB/(k - 1)$	$F = \frac{MSB}{MSR}$
Residual	$(r - 1)(k - 1)$	$SSR = \sum \sum (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y})^2$	$MSR = SSR/[(r - 1)(k - 1)]$	
Total	$rk - 1$	$SST = \sum \sum (y_{ij} - \bar{y})^2$		

Again, the Sum of squares column “adds,” that is,  $SSB_l + SSB + SSR = SST$ . The degrees of freedom column also adds.

**Example 11.3.1 (Continued):** The ANOVA table for the strawberry variety data is

Source of variation	Degrees of freedom	Sum of squares	Mean square	F statistic
Blocks	3	1.72	.57	
Varieties	2	35.58	17.79	146.22
Residual	6	.73	.12	
Total	11	38.03		

Note that 146.22 is not exactly equal to  $17.79/.12$ . But 146.22 is correct if the calculations are done to higher accuracy. Also note that we have referred to the error term as “residual,” not as “within error” as in the oneway ANOVA. This is to make the distinction clear, that our error estimate here is model-dependent and does not come from a “within” source. It is possible to have a within error term in a RCB ANOVA, which would happen if replications were taken within the treatment-block combinations. In such a case there would be a “within” row added to the above table. Searle (1971) treats this case in some detail.

Equation (11.3.18) can be interpreted as

$$(11.3.19) \quad SSW(\text{oneway}) = SSB_l(\text{RCB}) + SSR(\text{RCB}),$$

showing a possible advantage to blocking. That is, the sum of squares used for estimating error is reduced, possibly leading to a smaller error estimate and hence more significant results. The reduction is not certain, however, because it is the mean squares that are used and the degrees of freedom have been reduced from  $k(r - 1)$  in a oneway to  $(r - 1)(k - 1)$  in a RCB.

Realize that equation (11.3.19) just relates numbers and cannot be used to decide how to analyze a particular ANOVA. The oneway ANOVA and the RCB ANOVA are quite different in a most fundamental way. Recall that the oneway ANOVA, as described in Section 11.2, is a special case of a completely randomized design; the data are collected in a random order throughout the experiment. By its very nature, a block design restricts randomization to within the blocks and, hence, cannot be a completely randomized design. Any attempt to analyze it as such can only create bias in the analysis.

### 11.3.5 Implications of Random Blocking

Thus far in this section, we have operated conditionally on the blocks used in the experiment. However, we have seen that if inferences are restricted to those concerned

with treatment contrasts, then the inferences made conditional on blocks are the same as those made unconditional on blocks. This follows from the fact that, if we make our treatment inferences using either (11.3.10) or (11.3.16), we are basing our inference on statistics whose conditional distribution, and hence unconditional distribution, is independent of the value of  $b$ . Furthermore, as long as  $Y_{ij}|b$  is normal, the actual form of the distribution of the random blocks is of no consequence (since the first two moments of the normal determine the entire distribution); our inferences on treatment contrasts remain the same.

Realize that the fact that the blocks are *complete* plays an important role in freeing treatment contrasts from block effects. If blocks are *incomplete*, that is, if not every block contains every treatment, then the treatment contrasts will not, in general, be independent of block effects. In complicated situations, an incomplete block design may be preferred and the resulting ANOVA is more complicated than those considered here. Furthermore, even if an incomplete design is not preferred, it may be dictated by data-gathering problems.

A reasonable question to wonder about is what inference can be made about blocks, that is, can block effects be tested or estimated? This is an area where statisticians do not generally agree—it is almost a matter of taste. The formal mathematical statistics can be done in different correct ways and different correct answers can be obtained. In particular, the complete answer to this question is tied to the parameterization used for the model, but the details are too involved to give here. Searle (1971) shows algebraic relationships between different parameterizations, and Hocking (1973, 1985) goes further in explaining the different parameterizations. A more statistical answer to this question is given by Samuels, Casella, and McCabe (1988).

Looking at equation (11.3.19), it is clear that if the block sum of squares is very small, then blocking has increased the error estimate (since the degrees of freedom used for calculating the error mean square will have decreased). In other words, blocking pays off only if the blocks are significant. Also, by their very nature, blocks are not controllable, that is, blocks are just there—we cannot manipulate them (as, for example, in Example 11.3.1). So, if we can infer about blocks and find differences, such knowledge, in general, does us no good since we cannot implement it.

This brings us to the only inference about blocks that makes sense in the model (11.3.2). The only control we have over blocks is whether to use them, that is, at the beginning of an experiment we may have to decide whether to conduct a oneway or RCB ANOVA. Thus, for future reference, we would be interested in determining if the blocks are different. This can be easily done under some additional assumptions.

To conclude that the blocks are the same is to conclude  $b_1 = b_2 = \dots = b_r$ . Thus, we would be interested in testing

$$H_0: b_1 = b_2 = \dots = b_r \quad \text{versus} \quad H_1: b_i \neq b_j, \text{ for some } i, j.$$

If we assume that the blocks are not random but fixed, then, under  $H_0$ , the blocks are equal, so derivation of a test on blocks would be exactly the same as the derivation of a test on treatments. Thus, we can test the hypothesis that *the blocks used in the experiment have no effect* with the statistic

$$\frac{\sum_{i=1}^r k \left( (\bar{Y}_{\cdot j} - \bar{\bar{Y}}) \right)^2}{S_R^2},$$

which, under  $H_0$ , is distributed as  $(r-1)F_{r-1,(r-1)(k-1)}$ . (This is a rare case in which an ANOVA null is plausible and of interest.)

## EXERCISES

---

- 11.1** An ANOVA variance-stabilizing transformation stabilizes variances in the following approximate way. Let  $Y$  have mean  $\theta$  and variance  $v(\theta)$ .
- Use arguments as in Section 7.4.2 to show that a one-term Taylor series approximation of the variance of  $g(Y)$  is given by  $\text{Var}(g(Y)) = [\frac{d}{d\theta}g(\theta)]^2 v(\theta)$ .
  - Show that the approximate variance of  $g^*(Y)$  is independent of  $\theta$ , where  $g^*(y) = \int [1/\sqrt{v(y)}] dy$ .
- 11.2** Verify that the following transformations are approximately variance-stabilizing in the sense of Exercise 11.1.
- $Y \sim \text{Poisson}$ ,  $g^*(y) = \sqrt{y}$
  - $Y \sim \text{binomial}(n, p)$ ,  $g^*(y) = \sin^{-1}(\sqrt{y/n})$
  - $Y$  has variance  $v(\theta) = K\theta^2$  for some constant  $K$ ,  $g^*(y) = \log(y)$ .
- 11.3** The Box–Cox family of power transformations (Box and Cox, 1964) is defined by

$$g_\lambda^*(y) = \begin{cases} (y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0 \end{cases}$$

where  $\lambda$  is a free parameter.

- a. Show that, for each  $y$ ,  $g_\lambda^*(y)$  is continuous in  $\lambda$ . In particular, show that

$$\lim_{\lambda \rightarrow 0} (y^\lambda - 1)/\lambda = \log y.$$

- b. Find the function  $v(\theta)$ , the approximate variance of  $Y$ , that  $g_\lambda^*(y)$  stabilizes. (Note that  $v(\theta)$  will most likely also depend on  $\lambda$ .)

Analysis of transformed data in general, and the Box–Cox power transformation in particular, has been the topic of some controversy in the statistical literature. See Bickel and Doksum (1981), Box and Cox (1982), and Hinkley and Rungger (1984).

- 11.4** Suppose that random variables  $Y_{ij}$  are observed according to the overparameterized oneway ANOVA model in (11.2.2). Show that, without some restriction on the parameters, this model is not identifiable by exhibiting two distinct collections of parameters that lead to exactly the same distribution of the  $Y_{ij}$ s.
- 11.5** Under the oneway ANOVA assumptions, show that the set of statistics  $(\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \dots, \bar{Y}_{k\cdot}, S_p^2)$  is sufficient for  $(\theta_1, \theta_2, \dots, \theta_k, \sigma^2)$ .
- 11.6** Complete the proof of Theorem 11.2.2 by showing that

$$\frac{1}{\sigma^2} \sum_{i=1}^k n_i \left( (\bar{Y}_{i\cdot} - \bar{\bar{Y}}) - (\theta_i - \bar{\theta}) \right)^2 \sim \chi_{k-1}^2.$$

(Hint: Define  $\bar{U}_i = \bar{Y}_{i\cdot} - \theta_i$ ,  $i = 1, \dots, k$ . Show that  $\bar{U}_i$  are independent  $\text{n}(0, \sigma^2/n_i)$ . Then, adapt the induction argument of Lemma 5.4.1 to show that  $\sum n_i (\bar{U}_i - \bar{\bar{U}})^2 / \sigma^2 \sim \chi_{k-1}^2$ , where  $\bar{\bar{U}} = \sum n_i \bar{U}_i / \sum n_i$ .)

- 11.7** Show that under the oneway ANOVA assumptions, for any set of constants  $\mathbf{a} = (a_1, \dots, a_k)$ , the quantity  $\sum a_i \bar{Y}_i$  is normally distributed with mean  $\sum a_i \theta_i$  and variance  $\sigma^2 \sum a_i^2/n_i$ . (See Corollary 4.6.2.)
- 11.8** Show that under the oneway ANOVA assumptions,  $S_p^2 = \frac{1}{N-k} \sum_{i=1}^k (n_i - 1) S_i^2$  is independent of each  $\bar{Y}_i$ ,  $i = 1, \dots, k$ . (See Lemma 5.4.2.)
- 11.9** Using an argument similar to that which led to the  $t$  test in (11.2.7), show how to construct a  $t$  test for
- $H_0: \sum a_i \theta_i = \delta$  versus  $H_1: \sum a_i \theta_i \neq \delta$ ,
  - $H_0: \sum a_i \theta_i \leq \delta$  versus  $H_1: \sum a_i \theta_i > \delta$ ,
- where  $\delta$  is a specified constant.
- 11.10** Suppose we have a oneway ANOVA with five treatments. Denote the treatment means by  $\theta_1, \dots, \theta_5$ , where  $\theta_1$  is a control and  $\theta_2, \dots, \theta_5$  are alternate new treatments and assume that an equal number of observations per treatment is taken. Consider the four contrasts  $\sum a_i \theta_i$  defined by

$$\mathbf{a}_1 = \left(1, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}\right),$$

$$\mathbf{a}_2 = \left(0, 1, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}\right),$$

$$\mathbf{a}_3 = \left(0, 0, 1, -\frac{1}{2}, -\frac{1}{2}\right),$$

$$\mathbf{a}_4 = (0, 0, 0, 1, -1).$$

- Argue that the results of the four  $t$  tests using these contrasts can lead to conclusions about the ordering of  $\theta_1, \dots, \theta_5$ . What conclusions might be made?
- Show that any two contrasts  $\sum a_i \bar{Y}_i$  formed from the four  $\mathbf{a}_i$ s in part (a) are uncorrelated. (Recall that these are called orthogonal contrasts.)
- For the fertilizer experiment of Example 11.2.2, the following contrasts were planned:

	Treatment				
	1	2	3	4	5
$\mathbf{a}_1 =$	-1	1	0	0	0
$\mathbf{a}_2 =$	0	-1	$\frac{1}{2}$	$\frac{1}{2}$	0
$\mathbf{a}_3 =$	0	0	1	-1	0
$\mathbf{a}_4 =$	0	-1	0	0	1

Show that these contrasts are not orthogonal. Interpret these contrasts in the context of the fertilizer experiment, and argue that they are a sensible set of contrasts.

- 11.11** Show that under the oneway ANOVA assumptions, for any sets of constants  $\mathbf{a} = (a_1, \dots, a_k)$  and  $\mathbf{b} = (b_1, \dots, b_k)$ ,

$$\text{Cov}(\sum a_i \bar{Y}_i, \sum b_i \bar{Y}_i) = \sigma^2 \sum \frac{a_i b_i}{n_i}.$$

Hence, in the oneway ANOVA, contrasts are uncorrelated (orthogonal) if  $\sum a_i b_i / n_i = 0$ .

- 11.12** Suppose that we have a oneway ANOVA with equal numbers of observations on each treatment, that is,  $n_i = n, i = 1, \dots, k$ . In this case the  $F$  test can be considered an average  $t$  test.

a. Show that a  $t$  test of  $H_0: \theta_i = \theta_{i'}$  versus  $H_1: \theta_i \neq \theta_{i'}$  can be based on the statistic

$$t_{ii'}^2 = \frac{(\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot})^2}{S_p^2(2/n)}.$$

b. Show that

$$\frac{1}{k(k-1)} \sum_{i,i'} t_{ii'}^2 = F,$$

where  $F$  is the usual ANOVA  $F$  statistic. (*Hint:* See Exercise 5.9a.) (Communicated by George McCabe, who learned it from John Tukey.)

- 11.13** Under the oneway ANOVA assumptions,

a. Show that the likelihood ratio test of  $H_0: \theta_1 = \theta_2 = \dots = \theta_k$  is given by the  $F$  test of (11.2.14).

b. Show that this test is unbiased.

- 11.14** a. Show that the  $F$  statistic from a oneway ANOVA is invariant under a location-scale transformation of the data, that is, if  $Y_i \rightarrow aY_i + b$ , where  $a > 0$  and  $b$  are constants, then the  $F$  statistic is unchanged.

b. Show that the  $F$  statistic from a RCB ANOVA is also invariant under a location-scale transformation of the data.

- 11.15** The Scheffé simultaneous interval procedure actually works for all linear combinations, not just contrasts. Show that under the oneway ANOVA assumptions, if  $M = \sqrt{kF_{k,N-k,\alpha}}$  (note the change in the numerator degrees of freedom), then the probability is  $1 - \alpha$  that

$$\sum_{i=1}^k a_i \bar{Y}_{i\cdot} - M \sqrt{S_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}} \leq \sum_{i=1}^k a_i \theta_i \leq \sum_{i=1}^k a_i \bar{Y}_{i\cdot} + M \sqrt{S_p^2 \sum_{i=1}^k \frac{a_i^2}{n_i}},$$

simultaneously for all  $\mathbf{a} = (a_1, \dots, a_k)$ . (It is probably easiest to proceed by first establishing, in the spirit of Lemma 11.2.1, that if  $v_1, \dots, v_k$  are constants and  $c_1, \dots, c_k$  are positive constants, then

$$\max_{\mathbf{a}} \left\{ \frac{\left( \sum_{i=1}^k a_i v_i \right)^2}{\sum_{i=1}^k a_i^2 / c_i} \right\} = \sum_{i=1}^k c_i v_i^2.$$

The proof of Theorem 11.2.3 can then be adapted to establish the result.

- 11.16** a. Show that for the  $t$  and  $F$  distributions, for any  $\nu, \alpha$ , and  $k$ ,

$$t_{\nu, \alpha/2} \leq \sqrt{(k-1)F_{k-1, \nu, \alpha}}.$$

(Recall the relation between the  $t$  and the  $F$ . This inequality is a consequence of the fact that the distributions  $kF_{k,\nu}$  are stochastically increasing in  $k$  for fixed  $\nu$ , but is actually a weaker statement. See Exercise 5.32.)

- b. Explain how the above inequality shows that the simultaneous Scheffé intervals are always wider than the single-contrast intervals.
- c. Show that it also follows from the above inequality that Scheffé tests are less powerful than  $t$  tests.

- 11.17** In Theorem 11.2.1 we saw that the ANOVA null is equivalent to all contrasts being zero. We can also write the ANOVA null as the intersection over another set of hypotheses.
- a. Show that the hypotheses

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_k \quad \text{versus} \quad H_1: \theta_i \neq \theta_j \text{ for some } i, j$$

and the hypotheses

$$H_0: \theta_i - \theta_j = 0 \text{ for all } i, j \quad \text{versus} \quad H_1: \theta_i - \theta_j \neq 0 \text{ for some } i, j$$

are equivalent.

- b. Express  $H_0$  and  $H_1$  of the ANOVA test as unions and intersections of the sets

$$\Theta_{ij} = \{\theta = (\theta_1, \dots, \theta_k) : \theta_i - \theta_j = 0\}.$$

Describe how these expressions can be used to construct another (different) union-intersection test of the ANOVA null hypothesis. (See the Miscellanea section about the  $Q$  distribution in multiple comparisons.)

- 11.18** A multiple comparison procedure called the *Protected LSD* (Protected Least Significant Difference) is performed as follows. If the ANOVA  $F$  test rejects  $H_0$  at level  $\alpha$ , then for each pair of means  $\theta_i$  and  $\theta_{i'}$ , declare the means different if

$$\frac{|\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}|}{\sqrt{S_p^2 \left( \frac{1}{n_i} + \frac{1}{n_{i'}} \right)}} > t_{\alpha/2, N-k}.$$

Note that each  $t$  test is done at the same  $\alpha$  level as the ANOVA  $F$  test. Here we are using an *experimentwise*  $\alpha$  level, where

$$\text{experimentwise } \alpha = P \left( \begin{array}{c|c} \text{at least one false} & \text{all the means} \\ \text{assertion of difference} & \text{are equal} \end{array} \right).$$

- a. Prove that no matter how many means are in the experiment, simultaneous inference from the Protected LSD is made at level  $\alpha$ .
- b. The *ordinary* (or *unprotected*) LSD simply does the individual  $t$  tests, at level  $\alpha$ , no matter what the outcome of the ANOVA  $F$  test. Show that the ordinary LSD can have an experimentwise error rate greater than  $\alpha$ . (The unprotected LSD does maintain a *comparisonwise* error rate of  $\alpha$ .)
- c. Perform the LSD procedure on the fish toxin data of Example 11.2.1. What are the conclusions?

**11.19** To see that “data snooping,” that is, testing hypotheses that are suggested by the data, is generally not a good practice,

a. Show that, for any random variable  $Y$  and constants  $a, b$  with  $a > b$  and  $P(Y > b) < 1$ ,  $P(Y > a|Y > b) > P(Y > a)$ .

b. Apply the inequality in part (a) to the size of a data-suggested hypothesis test by letting  $Y$  be a test statistic and  $a$  be a cutoff point.

**11.20** Let  $X_i \sim \text{gamma}(\lambda_i, 1)$  independently for  $i = 1, \dots, n$ . Define  $Y_i = \frac{X_{i+1}}{\sum_{j=1}^i X_j}, i = 1, \dots, n-1$ , and  $Y_n = \sum_{i=1}^n X_i$ .

a. Find the joint and marginal distributions of  $Y_i, i = 1, \dots, n$ .

b. Connect your results to any distributions that are commonly employed in the ANOVA.

**11.21** Prove the oneway ANOVA expressions of (11.2.16) and (11.2.17):

a.  $\frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 \sim \chi^2_{N-k}$ ,

b.  $\frac{1}{\sigma^2} \sum_{i=1}^k n_i (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 \sim \chi^2_{k-1}$  and  $\frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{\bar{Y}})^2 \sim \chi^2_{N-1}$ .

(Searle and Pukelsheim (1985) have applied Kruskal’s chi squared induction proof to the oneway ANOVA.)

**11.22** Show that if the oneway ANOVA null hypothesis is true, then

a.  $\sum n_i (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 / (k-1)$  gives an unbiased estimate of  $\sigma^2$ .

b. Show how to use the method of Example 5.4.1 to derive the ANOVA  $F$  test.

**11.23** a. Illustrate the partitioning of the sums of squares in the ANOVA by calculating the complete ANOVA table for the following data. To determine diet quality, male weanling rats were fed diets with various protein levels. Each of fifteen rats was randomly assigned to one of three diets, and their weight gain in grams was recorded.

Diet protein level		
Low	Medium	High
3.89	8.54	20.39
3.87	9.32	24.22
3.26	8.76	30.91
2.70	9.30	22.78
3.82	10.45	26.33

b. Analytically verify the partitioning of the ANOVA sums of squares by completing the proof of Theorem 11.2.4.

c. Illustrate the relationship between the  $t$  and  $F$  statistics, given in Exercise 11.12b, using the data of part (a).

**11.24** Calculate the expected values of MSB and MSW given in the oneway ANOVA table. (Such expectations are formally known as *expected mean squares* and can be used to help identify  $F$  tests in complicated ANOVAs. An algorithm exists for calculating expected mean squares. See, for example, Kirk (1982) for details about the algorithm.)

**11.25** Show details for the following calculations.

a. The conditional mean and variance of  $Y_{ij}$ , under the RCB model.

b. The unconditional mean and variance of  $Y_{ij}$ , under the RCB model. ( $\text{Var } Y_{ij}$  can be calculated using the formula  $\text{Var } Y_{ij} = \text{Var}(\text{E}(Y_{ij}|b)) + \text{E}(\text{Var}(Y_{ij}|b))$ .)

**11.26** a. Verify equation (11.3.6), that is, show that in the RCB ANOVA

$$\text{Var}\left(\sum_{i=1}^k a_i \bar{Y}_{i\cdot} | \mathbf{b}\right) = \frac{\sigma^2}{r} \sum_{i=1}^k a_i^2.$$

- b. Show that  $\text{Var}(\sum a_i \bar{Y}_{i\cdot})$ , the unconditional variance of  $\sum a_i \bar{Y}_{i\cdot}$ , is equal to the conditional variance.  
c. Calculate  $\text{Var}(\sum a_i \bar{Y}_{i\cdot})$ , the unconditional variance of  $\sum a_i \bar{Y}_{i\cdot}$ , directly from the unconditional distribution of the  $Y_{ij}$ s and show that

$$\text{Var}(\sum a_i \bar{Y}_{i\cdot}) = \frac{1}{r}(\sigma^2 + \sigma_B^2)(1 - \rho) \sum a_i^2,$$

where  $\rho$  = the intraclass correlation. Using the definition of the intraclass correlation, show directly that

$$\text{Var}(\sum a_i \bar{Y}_{i\cdot}) = \text{Var}(\sum a_i \bar{Y}_{i\cdot} | \mathbf{b}).$$

- 11.27** For the RCB residual given in (11.3.9),  $\hat{\epsilon}_{ij} = Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{\bar{Y}}$ , show that  
a.  $E(\hat{\epsilon}_{ij} | \mathbf{b}) = 0$   
b.  $\text{Var}(\hat{\epsilon}_{ij} | \mathbf{b}) = E(\hat{\epsilon}_{ij}^2 | \mathbf{b}) = \left(\frac{r-1}{r}\right) \left(\frac{k-1}{k}\right) \sigma^2$ . (The expectation in part (b) requires detailed calculation. After expanding  $\hat{\epsilon}_{ij}^2$ , show that

$$E((Y_{ij} - \bar{Y}_{i\cdot})^2 | \mathbf{b}) = \left(\frac{r-1}{r}\right) \sigma^2 + (b_j - \bar{b})^2,$$

$$E((Y_{ij} - \bar{Y}_{i\cdot})(\bar{Y}_{\cdot j} - \bar{\bar{Y}}) | \mathbf{b}) = \frac{1}{k} \left(\frac{r-1}{r}\right) \sigma^2 + (b_j - \bar{b})^2,$$

and

$$E((\bar{Y}_{\cdot j} - \bar{\bar{Y}})^2 | \mathbf{b}) = \frac{1}{k} \left(\frac{r-1}{r}\right) \sigma^2 + (b_j - \bar{b})^2,$$

which, when put together, give part (b).)

- 11.28** The fact that the RCB residuals are constrained to add to zero implies that they are correlated.  
a. Show that

$$\text{Cov}(\hat{\epsilon}_{ij}, \hat{\epsilon}_{i'j'}) = \sigma^2 \frac{(r-1)(k-1)}{rk}, \quad i = i', \quad j = j',$$

$$\text{Cov}(\hat{\epsilon}_{ij}, \hat{\epsilon}_{i'j'}) = -\sigma^2 \frac{k-1}{rk}, \quad i = i', \quad j \neq j',$$

$$\text{Cov}(\hat{\epsilon}_{ij}, \hat{\epsilon}_{i'j'}) = -\sigma^2 \frac{r-1}{rk}, \quad i \neq i', \quad j = j',$$

$$\text{Cov}(\hat{\epsilon}_{ij}, \hat{\epsilon}_{i'j'}) = \sigma^2 \frac{1}{rk}, \quad i \neq i', \quad j \neq j'.$$

- b. As a partial check on your calculations in part (a), calculate  $\sum_{j>j'} \text{Cov}(\hat{\epsilon}_{ij}, \hat{\epsilon}_{i'j'})$  from  $\sum_{j=1}^r \text{Var} \hat{\epsilon}_{ij}$ , using the fact that  $\sum_{j=1}^r \hat{\epsilon}_{ij} = 0$  for each  $i$ .

- c. As another partial check on your calculations in part (a), calculate  $\sum_{i>i'} \text{Cov}(\hat{\epsilon}_{ij}, \hat{\epsilon}_{i'j'})$  from  $\sum_{i=1}^k \text{Var} \hat{\epsilon}_{ij}$ , using the fact that  $\sum_{i=1}^k \hat{\epsilon}_{ij} = 0$  for each  $j$ .

**11.29** In this exercise we will show that the RCB estimate of  $\sigma^2$ ,  $S_R^2$ , has  $(r-1)(k-1)$  degrees of freedom.

- Show that  $\sum_{i=1}^k \hat{\epsilon}_{ij} = 0$ , for each  $j = 1, \dots, r$ , and  $\sum_{j=1}^r \hat{\epsilon}_{ij} = 0$ , for each  $i = 1, \dots, k$ .
- Show that part (a) implies that  $S_R^2$  can be calculated without  $\hat{\epsilon}_{ir}, i = 1, \dots, k$ , and  $\hat{\epsilon}_{kj}, j = 1, \dots, r$ .
- Argue that  $r+k-1$  terms in  $S_R^2$  can be discarded and thus  $S_R^2$  has  $(r-1)(k-1)$  degrees of freedom.

**11.30** Finish the proof showing that  $S_R^2$  is independent of  $\sum_{i=1}^k a_i \bar{Y}_{i\cdot}$ , under normality, by establishing (11.3.11). That is, show

$$\text{Cov}(\hat{\epsilon}_{i'j'}, \bar{Y}_{i\cdot} | \mathbf{b}) = 0 \quad \text{for all } i, i', j'.$$

Expand the covariance into the following four terms and establish

$$\begin{aligned} E(Y_{i'j'} \bar{Y}_{i\cdot} | \mathbf{b}) &= (\mu + \tau_{i'} + b_{j'})(\mu + \tau_i + \bar{b}) + \delta_{ii'} \sigma^2 / r, \\ E(\bar{Y}_{i'j'} \bar{Y}_{i\cdot} | \mathbf{b}) &= (\mu + \tau_{i'} + \bar{b})(\mu + \tau_i + \bar{b}) + \delta_{ii'} \sigma^2 / r, \\ E(\bar{Y}_{i\cdot} \bar{Y}_{i\cdot} | \mathbf{b}) &= (\mu + \bar{\tau} + b_{j'})(\mu + \tau_i + \bar{b}) + \sigma^2 / (rk), \\ E(\bar{Y}_{i\cdot} \bar{Y}_{i\cdot} | \mathbf{b}) &= (\mu + \bar{\tau} + \bar{b})(\mu + \tau_i + \bar{b}) + \sigma^2 / (rk). \end{aligned}$$

**11.31** The parts of this exercise are to be done directly, without referring to Theorem 11.3.1.

- Show that if  $k = 2$  in the RCB ANOVA (there are two treatments), then

$$(r-1) \frac{S_R^2}{\sigma^2} = \sum_{i=1}^2 \sum_{j=1}^r \frac{\hat{\epsilon}_{ij}^2}{\sigma^2}$$

is a chi squared random variable with  $r-1$  degrees of freedom.

- Similarly, if  $r = 2$ , (there are two blocks), then

$$(k-1) \frac{S_R^2}{\sigma^2} = \sum_{i=1}^k \sum_{j=1}^2 \frac{\hat{\epsilon}_{ij}^2}{\sigma^2}$$

is a chi squared random variable with  $k-1$  degrees of freedom.

**11.32** Show that if  $\sum_{i=1}^k \sum_{j=1}^r a_{ij} \hat{\epsilon}_{ij}$  and  $\sum_{i=1}^k \sum_{j=1}^r c_{ij} \hat{\epsilon}_{ij}$  are two linear combinations of RCB residuals, then

$$\text{a. } \text{Cov} \left( \sum_{i=1}^k \sum_{j=1}^r a_{ij} \hat{\epsilon}_{ij}, \sum_{i=1}^k \sum_{j=1}^r c_{ij} \hat{\epsilon}_{ij} \right)$$

$$= \sum_{i=1}^k \sum_{j=1}^r a_{ij} \left\{ c_{ij} - \frac{1}{r} \sum_{j'=1}^r c_{ij'} - \frac{1}{k} \sum_{i'=1}^k c_{i'j} + \frac{1}{rk} \sum_{i'=1}^k \sum_{j'=1}^r c_{i'j'} \right\}.$$

- If the  $a_{ij}$ s and  $c_{ij}$ s define contrasts (as defined in Theorem 11.3.1) then the above contrasts are uncorrelated if  $\sum_{i=1}^k \sum_{j=1}^r a_{ij} c_{ij} = 0$ .

- Show that if the  $a_{ij}$ s define a contrast, then

$$E\left(\sum_{i=1}^k \sum_{j=1}^r a_{ij} \hat{\epsilon}_{ij}\right) = 0 \quad \text{and} \quad \text{Var}\left(\sum_{i=1}^k \sum_{j=1}^r a_{ij} \hat{\epsilon}_{ij}\right) = \tau^2 \sum_{i=1}^k \sum_{j=1}^r a_{ij}^2,$$

where  $\text{Var } \hat{\epsilon}_{ij} = \tau^2 = \sigma^2(r-1)(k-1)/(rk)$ . Hence, under the RCB ANOVA assumptions, if  $\sum_{i=1}^k \sum_{j=1}^r a_{ij}^2 = 1$ ,

$$\frac{1}{\tau^2} \left( \sum_{i=1}^k \sum_{j=1}^r a_{ij} \hat{\epsilon}_{ij} \right)^2 \sim \chi_1^2.$$

- 11.33** A variation of a RCB, one with only a mean level and no treatments, is given by the model

$$Y_{ij} | \mathbf{c} = \mu + c_i + \epsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, k,$$

where  $\mathbf{c} = (c_1, \dots, c_k)$  is a realization of the iid random variables  $C_i \sim N(0, \sigma_C^2)$ ,  $\epsilon_{ij} \sim N(0, \sigma^2)$ , and all the  $C_i$  and  $\epsilon_{ij}$  are independent.

- Calculate the unconditional expected values of  $\sum (\bar{Y}_{i \cdot} - \bar{\bar{Y}})^2$  and  $\sum \sum (Y_{ij} - \bar{Y}_{i \cdot})^2$ .
- Find an unbiased estimator of  $\sigma_C^2$ .

- 11.34** There are two details left to complete in the proof of Theorem 11.3.2.

- Show that  $\text{Cov}(\bar{Y}_{i \cdot} - \bar{\bar{Y}}, Y_{i' j'} - \bar{Y}_{i' \cdot} - \bar{Y}_{j'} + \bar{\bar{Y}}) = 0$ , which together with the normality assumption, implies that the numerator and denominator of (11.3.15) are independent.
- Then, using a technique as in Exercise 11.6, show that

$$\frac{1}{\sigma^2} \sum_{i=1}^k r \left( (\bar{Y}_{i \cdot} - \bar{\bar{Y}}) - (\tau_i - \bar{\tau}) \right)^2 \sim \chi_{k-1}^2,$$

a fact needed to deduce that (11.3.16) holds.

- 11.35** a. Illustrate the partitioning of the sums of squares in the RCB ANOVA by calculating the complete ANOVA table for the following data. The effectiveness of three anticoagulant drugs in dissolving blood clots was studied. Each of five subjects (blocks) received all three drugs (in random order with adequate washout time in between), and the length of time (in seconds) required for a cut of specified size to stop bleeding was recorded. The data are

		Anticoagulant drug		
		A	B	C
Person (Block)	1	127.5	129.0	135.5
	2	130.6	129.1	138.0
	3	118.3	111.7	110.1
	4	155.5	144.3	162.3
	5	180.7	174.4	181.8

- b. Analytically verify the partitioning of the RCB ANOVA sums of squares by completing the proof Theorem 11.3.4 by verifying (11.3.18).

## Miscellanea

---

### Cochran's Theorem

Sums of squares of normal random variables, when properly scaled and centered, are distributed as chi squared random variables. This type of result is first due to Cochran (1934). Cochran's Theorem gives necessary and sufficient conditions on the scaling required for squared and summed iid normal random variables to be distributed as a chi squared random variable. The conditions are not difficult, but they are best stated in terms of properties of matrices, and will not be treated here.

It is an immediate consequence of Cochran's Theorem that in the oneway ANOVA the  $\chi^2$  random variables partition as discussed in Section 11.2.6. Furthermore, another consequence is that in a RCB ANOVA,  $(r - 1)(k - 1)S_R^2/\sigma^2$  has a chi squared distribution. Here it is a matter of checking that the matrix that can be used to define  $S_R^2$  satisfies the conditions of the theorem.

Cochran's Theorem has been generalized to the extent that necessary and sufficient conditions are known for the distribution of squared normals (not necessarily iid) to be chi squared. See Theorem 5 and Corollary 5.1 in Chapter 2 of Searle (1971), who treats this topic in detail.

### Multiple Comparisons

We have seen two ways of doing simultaneous inference in this chapter, the Scheffé procedure and use of the Bonferroni Inequality. There is a plethora of other simultaneous inference procedures. Most are concerned with inference on pairwise comparisons, that is, differences between means. These procedures can be applied to estimate treatment means in both the oneway ANOVA and the RCB ANOVA.

A method due to Tukey (see Miller, 1981), sometimes known as the  $Q$  method, applies a Scheffé-type maximization argument, but only over pairwise differences, not all contrasts. The  $Q$  distribution is the distribution of

$$Q = \max_{i,j} \left| \frac{(\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}) - (\theta_i - \theta_j)}{\sqrt{S_p^2 \left( \frac{1}{n} + \frac{1}{n} \right)}} \right|,$$

where  $n_i = n$  for all  $i$ . (Recently, Hayter (1984) has shown that if  $n_i \neq n_j$  and the  $n$  above is replaced by the harmonic mean  $n_h$ , where  $1/n_h = \frac{1}{2}((1/n_i) + (1/n_j))$ , the resulting procedure is conservative.) The  $Q$  method is an improvement over Scheffé's  $S$  method in that if there is interest only in pairwise differences, the  $Q$  method is more powerful (shorter intervals). This is easy to see because, by definition, the  $Q$  maximization will produce a smaller maximum than the  $S$  method.

Other types of multiple comparison procedures, which deal with pairwise differences, are more powerful than the  $S$  method. Some procedures are the LSD (Least Significant Difference) Procedure, Protected LSD, Duncan's Procedure, Student–Neumann–Keuls' Procedure. These last two are *multiple range* procedures. The cutoff point to which comparisons are made changes between comparisons.

One difficulty in fully understanding multiple comparison procedures is that the definition of Type I Error is not inviolate. Some of these procedures have changed the definition of Type I Error for multiple comparisons, so exactly what is meant by “ $\alpha$  level” is not always clear. Some of the types of error rates considered are called *experimentwise error rate*, *comparisonwise error rate*, and *familywise error rate*. Miller (1981) is a good reference for this topic. A humorous, but illuminating treatment of this subject is given in Carmer and Walker (1982).

### **Selection and Ranking**

In many ANOVAs, the main interest is in determining which treatment is “best,” that is, which has the largest mean. Some information about this can be gained from multiple comparison procedures, but these provide a rather roundabout answer if the only question is, “Which treatment is best?” Statistical procedures that more directly answer this question are called *selection and ranking procedures*.

Bechhofer (1954) studied the “natural” rule, the rule that simply asserts that the treatment with the largest observed sample mean is the one with the largest population mean. He studied how large a sample size is required in order to have a prescribed probability,  $P^*$ , that the treatment with the largest population mean will produce the largest sample mean. (Because of variation, it is possible to observe a larger sample mean from a population with a smaller true mean.) In order to achieve this, attention must be restricted to the subset of the parameter space where the largest population mean exceeds the second largest by at least  $\delta$ , a specified positive amount.

Gupta (1965) studied a subset selection procedure, an inference procedure that answers the question, “Which treatment is best?,” in a different way. A subset selection procedure selects a subset of the treatments and asserts that the best treatment is one of those in the selected subset. The Gupta subset selection procedure selects all treatments with observed sample means that satisfy

$$\bar{y}_{i \cdot} \geq \max_{1 \leq j \leq k} \bar{y}_{j \cdot} - d.$$

The constant  $d$  is chosen so that there is a prescribed probability,  $P^*$ , that the selected subset includes the best treatment. No restriction is placed on the parameter space when deriving a subset selection procedure, but the selected subset may be quite large and the assertion rather imprecise.

Selection and ranking procedures have been developed for many types of populations and sampling plans—binomial, gamma, etc. A comprehensive bibliography can be found in Gupta and Panchapakesan (1979).

### **Stein Estimation in ANOVA**

The setting of the ANOVA is perfect for application of Stein estimation (Section 10.7) and can lead to improved procedures. For example, in the oneway ANOVA, we have

$$\bar{Y}_{i \cdot} \sim n \left( \theta_i, \frac{\sigma^2}{n_i} \right), \quad i = 1, \dots, k, \quad \text{independent},$$

where the  $\bar{Y}_{i \cdot}$ s are the cell means.

The classic way to estimate  $\theta_i$  is to use  $\bar{Y}_{i \cdot}$ . Suppose, however, we assume there is a loss function (for example, summed squared error, as in (10.7.1)) that joins the estimation together (as will most often be the case in ANOVA). Then a Stein-type estimator can offer improvement.

Using the methods of Section 10.7, an alternate estimator for  $(\theta_1, \dots, \theta_k)$  is

$$\delta_i(\bar{Y}_{1\cdot}, \dots, \bar{Y}_{k\cdot}) = \left(1 - \frac{(k-2)\sigma^2}{\sum n_j \bar{Y}_{j\cdot}^2}\right)^+ \bar{Y}_{i\cdot}, \quad i = 1, \dots, k,$$

which provides a risk improvement over the classic estimator. This Stein-type estimator can further be improved by choosing a meaningful place toward which to shrink (the above estimator shrinks toward zero). One such estimator, due to Lindley (1962), shrinks toward the grand mean of the observations. It is given by

$$\delta_i^L(\bar{Y}_{1\cdot}, \dots, \bar{Y}_{k\cdot}) = \bar{\bar{Y}} + \left(1 - \frac{(k-3)\sigma^2}{\sum n_j (\bar{Y}_{j\cdot} - \bar{\bar{Y}})^2}\right)^+ (\bar{Y}_{i\cdot} - \bar{\bar{Y}}), \quad i = 1, \dots, k.$$

Other choices of a shrinkage target might be even more appropriate. Discussion of this, including methods for improving on confidence statements, such as the Scheffé  $S$  method, is given in Casella and Hwang (1987). Morris (1983) also discusses applications of these types of estimators.

### *Other Types of Analyses of Variance*

The two types of ANOVA that we have considered, oneway ANOVAs and RCB ANOVAs, are the simplest types. For example, an extension of a complete block design is an *incomplete* block design. Sometimes there are physical constraints that prohibit putting all treatments in each block and an incomplete block design is needed. Deciding how to arrange the treatments in such a design is both difficult and critical. Of course, as the design gets more complicated, so does the analysis.

Study of the subject of statistical design, which is concerned with getting the most information from the fewest observations, leads to more complicated and more efficient ANOVAs in many situations. ANOVAs based on designs such as *fractional factorials*, *Latin squares*, and *balanced incomplete blocks* can be efficient methods of gathering much information about a phenomenon. Good overall references for this subject are Cochran and Cox (1957), Kirk (1982), and Montgomery (1984).

# 12 Linear Regression

*"Still, it is an error to argue in front of your data. You find yourself insensibly twisting them round to fit your theories."*

**Sherlock Holmes**

*The Adventure of Wisteria Lodge*

## 12.1 Introduction

The technique of regression, in particular linear regression, probably wins the prize as the most popular statistical tool. There are all forms of regression: linear, nonlinear, simple, multiple, parametric, nonparametric, etc. In this chapter we will look at the simplest case, linear regression with one predictor variable. (This is usually called *simple* linear regression, as opposed to *multiple* linear regression, which deals with many predictor variables.) Good overall references are Neter, Wasserman, and Kutner (1985) and Draper and Smith (1981). A more theoretical treatment is given in Kendall and Stuart (1979).

The major purpose of regression is to explore the dependence of one variable on another. In particular, in simple linear regression we have a relationship of the form

$$(12.1.1) \quad Y_i = \alpha + \beta x_i + \epsilon_i,$$

where  $Y_i$  is a random variable and  $x_i$  is another observable variable. The quantities  $\alpha$  and  $\beta$ , the *intercept* and *slope* of the regression, are assumed to be fixed and unknown parameters and  $\epsilon_i$  is, necessarily, a random variable. It is also common to suppose that  $E\epsilon_i = 0$  (otherwise we could just rescale the excess into  $\alpha$ ), so that, from (12.1.1), we have

$$(12.1.2) \quad EY_i = \alpha + \beta x_i.$$

In general, the function that gives  $EY$  as a function of  $x$  is called the *population regression function*. Equation (12.1.2) defines the population regression function for simple linear regression.

One main purpose of regression is to predict  $Y_i$  from knowledge of  $x_i$ , using a relationship like (12.1.2). In common usage this is often interpreted as saying that  $Y_i$  depends on  $x_i$ . It is common to refer to  $Y_i$  as the *dependent* variable and to refer to  $x_i$  as the *independent* variable. This terminology is confusing, however, since this use of the word “independent” is different from our previous usage. (The  $x_i$ s are not necessarily random variables, so they cannot be statistically “independent” according to our usual meaning.) We will not use this confusing terminology but will use alternate, more descriptive, terminology, referring to  $Y_i$  as the *response* variable and to  $x_i$  as the *predictor* variable.

Actually, to keep straight the fact that our inferences about the relationship between  $Y_i$  and  $x_i$  assume knowledge of  $x_i$ , we could write (12.1.2) as

$$(12.1.3) \quad E(Y_i|x_i) = \alpha + \beta x_i.$$

We will tend to use (12.1.3), to reinforce the conditional aspect of any inferences.

Recall that in Chapter 4 we encountered the word *regression* in connection with conditional expectations (see Exercise 4.13). There, the regression of  $Y$  on  $X$  was defined as  $E(Y|x)$ , the conditional expectation of  $Y$  given  $X = x$ . More generally, the word *regression* is used in statistics to signify a relationship between variables. When we refer to *regression that is linear*, we can mean that the conditional expectation of  $Y$  given  $X = x$  is a linear function of  $x$ . Note that, in equation (12.1.3), it does not matter whether  $x_i$  is fixed and known or it is a realization of the observable random variable  $X_i$ . In either case, equation (12.1.3) has the same interpretation. This will not be the case in Section 12.2.4, however, when we will be concerned with inference using the joint distribution of  $X_i$  and  $Y_i$ .

The term *linear regression* refers to a specification that is *linear in the parameters*. Thus, the specifications  $E(Y_i|x_i) = \alpha + \beta x_i^2$  and  $E(\log Y_i|x_i) = \alpha + \beta(1/x_i)$  both specify linear regressions. The first specifies a linear relationship between  $Y_i$  and  $x_i^2$ , and the second between  $\log Y_i$  and  $1/x_i$ . In contrast, the specification  $E(Y_i|x_i) = \alpha + \beta^2 x_i$  does not specify a linear regression.

The term *regression* has an interesting history, dating back to the work of Sir Francis Galton in the 1800s. (See Freedman, Pisani, and Purves (1978) for more details or Stigler (1986) for an in-depth historical treatment.) Galton investigated the relationship between heights of fathers and heights of sons. He found, not surprisingly, that tall fathers tend to have tall sons and short fathers tend to have short sons. However, he also found that very tall fathers tend to have shorter sons and very short fathers tend to have taller sons. (Think about it—it makes sense.) Galton called this phenomenon *regression toward the mean* (employing the usual meaning of *regression*, “to go back”) and from this usage we get the present use of the word *regression*.

**Example 12.1.1:** A more modern use of regression is to predict crop yield of grapes. In July, the grape vines produce clusters of berries, and a count of these clusters can be used to predict the final crop yield at harvest time. Typical data are like the following, which give the cluster counts and yield (tons/acre) for a number of years.

Year	Yield ( $Y$ )	Cluster count ( $x$ )
1971	5.6	116.37
1973	3.2	82.77
1974	4.5	110.68
1975	4.2	97.50
1976	5.2	115.88
1977	2.7	80.19
1978	4.8	125.24
1979	4.9	116.15
1980	4.7	117.36
1981	4.1	93.31
1982	4.4	107.46
1983	5.4	122.30

The data from 1972 are missing because the crop was destroyed by a hurricane.  
A plot of these data would show that there is a strong linear relationship. ||

When we write an equation like (12.1.3) we are implicitly making the assumption that the regression of  $Y$  on  $X$  is linear. That is, the conditional expectation of  $Y$ , given that  $X = x$ , is a linear function of  $x$ . This assumption may not be justified, because there may be no underlying theory to support a linear relationship. However, since a linear relationship is so convenient to work with, we might want to assume that the regression of  $Y$  on  $X$  can be adequately approximated by a linear function. Thus, we really do not expect (12.1.3) to hold, but instead we hope that

$$(12.1.4) \quad E(Y_i|x_i) \approx \alpha + \beta x_i$$

is a reasonable approximation. If we start from the (rather strong) assumption that the pair  $(X_i, Y_i)$  has a bivariate normal distribution, it immediately follows that the regression of  $Y$  on  $X$  is linear. In this case, the conditional expectation  $E(Y|x)$  is linear in the parameters (see Definition 4.5.3 and the subsequent discussion).

There is one final distinction to be made. When doing a regression analysis, that is, when investigating the relationship between a predictor and a response variable, there are two steps to the analysis. The first step is a totally data-oriented one, in which we attempt only to summarize the observed data. (This step is always done, since we almost always calculate sample means and variances or some other summary statistic. However, this part of the analysis now tends to get more complicated.) It is important to keep in mind that this “data fitting” step is not a matter of statistical inference. Since we are interested only in the data at hand, we do not have to make any assumptions about parameters.

The second step in the regression analysis is the statistical one, in which we attempt to infer conclusions about the relationship in the population, that is, about the population regression function. To do this, we need to make assumptions about the population. In particular, if we want to make inferences about the slope and intercept of a population linear relationship, we need to assume that there are parameters that correspond to these quantities.

## 12.2 Simple Linear Regression

In a simple linear regression problem, we observe data consisting of  $n$  pairs of observations,  $(x_1, y_1), \dots, (x_n, y_n)$ . In this section, we will consider a number of different models for these data. The different models will entail different assumptions about whether  $x$  or  $y$  or both are observed values of random variables  $X$  or  $Y$ .

In each model we will be interested in investigating a linear relationship between  $x$  and  $y$ . The  $n$  data points will not fall exactly on a straight line, but we will be interested in *summarizing* the sample information by *fitting a line* to the observed data points. We will find that many different approaches lead us to the same line.

Based on the data  $(x_1, y_1), \dots, (x_n, y_n)$ , define the following quantities. The *sample means* are

$$(12.2.1) \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

The *sums of squares* are

$$(12.2.2a) \quad S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{and} \quad S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

and the *sum of cross-products* is

$$(12.2.2b) \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Then the most common estimates of  $\alpha$  and  $\beta$  in (12.1.4), which we will subsequently justify under various models, are denoted by  $a$  and  $b$ , respectively, and are given by

$$(12.2.3) \quad b = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad a = \bar{y} - b\bar{x}.$$

### 12.2.1 Least Squares: A Mathematical Solution

Our first derivation of estimates for  $\alpha$  and  $\beta$  makes no statistical assumptions about the observations  $(x_i, y_i)$ . Simply consider  $(x_1, y_1), \dots, (x_n, y_n)$  as  $n$  pairs of numbers plotted in a scatterplot as in Figure 12.2.1 on page 558. (The 24 data points pictured in Figure 12.2.1 are listed in Table 12.2.1.) Think of drawing through this cloud of points a straight line that comes “as close as possible” to all the points.

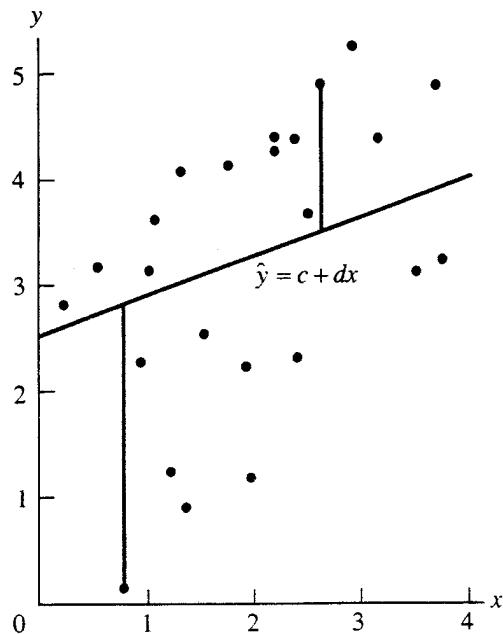


FIGURE 12.2.1 Data from Table 12.2.1: Vertical distances measured in RSS

$x$	$y$	$x$	$y$	$x$	$y$	$x$	$y$
3.74	3.22	0.20	2.81	1.22	1.23	1.76	4.12
3.66	4.87	2.50	3.71	1.00	3.13	0.51	3.16
0.78	0.12	3.50	3.11	1.29	4.05	2.17	4.40
2.40	2.31	1.35	0.90	0.95	2.28	1.99	1.18
2.18	4.25	2.36	4.39	1.05	3.60	1.53	2.54
1.93	2.24	3.13	4.36	2.92	5.39	2.60	4.89
$\bar{x} = 1.95$		$\bar{y} = 3.18$		$S_{xx} = 22.82$		$S_{yy} = 43.62$	
						$S_{xy} = 15.48$	

TABLE 12.2.1 Data pictured in Figure 12.2.1

For any line  $y = c + dx$ , the *residual sum of squares* (RSS) is defined to be

$$\text{RSS} = \sum_{i=1}^n (y_i - (c + dx_i))^2.$$

The RSS measures the *vertical* distance from each data point to the line  $c + dx$  and then sums the squares of these distances. (Two such distances are shown in Figure 12.2.1.) The *least squares estimates* of  $\alpha$  and  $\beta$  are defined to be those values  $a$  and  $b$  such that the line  $a + bx$  minimizes RSS. That is, the least squares estimates,  $a$  and  $b$ , satisfy

$$\min_{c,d} \sum_{i=1}^n (y_i - (c + dx_i))^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2.$$

This function of two variables,  $c$  and  $d$ , can be minimized in the following way. For any fixed value of  $d$ , the value of  $c$  that gives the minimum value can be found

by writing

$$\sum_{i=1}^n (y_i - (c + dx_i))^2 = \sum_{i=1}^n ((y_i - dx_i) - c)^2.$$

From Theorem 5.2.1, the minimizing value of  $c$  is

$$(12.2.4) \quad c = \frac{1}{n} \sum_{i=1}^n (y_i - dx_i) = \bar{y} - d\bar{x}.$$

Thus, for a given value of  $d$ , the minimum value of RSS is

$$\sum_{i=1}^n ((y_i - dx_i) - (\bar{y} - d\bar{x}))^2 = \sum_{i=1}^n ((y_i - \bar{y}) - d(x_i - \bar{x}))^2 = S_{yy} - 2dS_{xy} + d^2 S_{xx}.$$

The value of  $d$  that gives the overall minimum value of RSS is obtained by setting the derivative of this quadratic function of  $d$  equal to zero. The minimizing value is

$$(12.2.5) \quad d = \frac{S_{xy}}{S_{xx}}.$$

This value is, indeed, a minimum since the coefficient of  $d^2$  is positive. Thus, by (12.2.4) and (12.2.5),  $a$  and  $b$  from (12.2.3) are the values of  $c$  and  $d$  that minimize the residual sum of squares.

The RSS is only one of many reasonable ways of measuring the distance from the line  $c + dx$  to the data points. For example, rather than using vertical distances we could use horizontal distances. This is equivalent to graphing the  $y$  variable on the horizontal axis and the  $x$  variable on the vertical axis and using vertical distances as we did above. Using the above results (interchanging the roles of  $x$  and  $y$ ) we find the least squares line is  $\hat{x} = a' + b'y$  where

$$b' = \frac{S_{xy}}{S_{yy}} \quad \text{and} \quad a' = \bar{x} - b'\bar{y}.$$

Reexpressing the line so that  $y$  is a function of  $x$ , we obtain  $\hat{y} = -(a'/b') + (1/b')x$ .

Usually the line obtained by considering horizontal distances is different from the line obtained by considering vertical distances. Using the values in Table 12.2.1, the *regression of  $y$  on  $x$*  (vertical distances) is  $\hat{y} = 1.86 + .68x$ . The *regression of  $x$  on  $y$*  (horizontal distances) is  $\hat{x} = -2.31 + 2.82y$ . In Figure 12.3.2 (page 587), these two lines are shown (along with a third line discussed in Section 12.3). If these two lines were the same, then the slopes would be the same and  $b/(1/b')$  would equal 1. But, in fact,  $b/(1/b') \leq 1$  with equality only in special cases. Note that

$$\frac{b}{1/b'} = bb' = \frac{(S_{xy})^2}{S_{xx}S_{yy}}.$$

Using the version of Hölder's Inequality in (4.7.9) with  $p = q = 2$ ,  $a_i = x_i - \bar{x}$ , and  $b_i = y_i - \bar{y}$ , we see that  $(S_{xy})^2 \leq S_{xx}S_{yy}$  and, hence, the ratio is less than one.

If  $x$  is the predictor variable,  $y$  is the response variable, and we think of predicting  $y$  from  $x$ , then the vertical distance measured in RSS is reasonable. It measures the distance from  $y_i$  to the predicted value of  $y_i$ ,  $\hat{y}_i = c + dx_i$ . But if we do not make this distinction between  $x$  and  $y$ , then it is unsettling that another reasonable criterion, horizontal distance, gives a different line.

The least squares method should be considered only as a method of "fitting a line" to a set of data, not as a method of statistical inference. We have no basis for constructing confidence intervals or testing hypotheses because, in this section, we have not used any statistical model for the data. When thinking of  $a$  and  $b$  in the context of this section, it might be better to call them least squares *solutions* rather than least squares *estimates* because they are the solutions of the mathematical problem of minimizing the RSS rather than estimates derived from a statistical model. But, as we shall see, these least squares solutions have optimality properties in certain statistical models.

### 12.2.2 Best Linear Unbiased Estimators: A Statistical Solution

In this section we show that the estimates  $a$  and  $b$  from (12.2.3) are optimal in the class of linear, unbiased estimates under a fairly general statistical model. The model is described as follows. Assume that the values  $x_1, \dots, x_n$  are known, fixed values. (Think of them as values the experimenter has chosen and set in a laboratory experiment.) The values  $y_1, \dots, y_n$  are observed values of uncorrelated random variables  $Y_1, \dots, Y_n$ . The linear relationship assumed between the  $xs$  and the  $ys$  is

$$(12.2.6) \quad EY_i = \alpha + \beta x_i, \quad i = 1, \dots, n,$$

where we also assume that

$$(12.2.7) \quad \text{Var } Y_i = \sigma^2.$$

There is no subscript in  $\sigma^2$ , because we are assuming that all the  $Y_i$ 's have the same (unknown) variance. These assumptions about the first two moments of the  $Y_i$ 's are the only assumptions we need to make to proceed with the derivation in this subsection. For example, we do not need to specify a probability distribution for the  $Y_1, \dots, Y_n$ .

The model in (12.2.6) and (12.2.7) can also be expressed in this way. We assume that

$$(12.2.8) \quad Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n,$$

where  $\epsilon_1, \dots, \epsilon_n$  are uncorrelated random variables with

$$(12.2.9) \quad E\epsilon_i = 0 \quad \text{and} \quad \text{Var } \epsilon_i = \sigma^2.$$

The  $\epsilon_1, \dots, \epsilon_n$  are called the *random errors*. Since  $Y_i$  depends only on  $\epsilon_i$  and the  $\epsilon_i$ s are uncorrelated, the  $Y_i$ s are uncorrelated. Also, from (12.2.8) and (12.2.9), the expressions for  $EY_i$  and  $\text{Var } Y_i$  in (12.2.6) and (12.2.7) are easily verified.

To derive estimators for the parameters  $\alpha$  and  $\beta$ , we restrict attention to the class of *linear estimators*. An estimator is a linear estimator if it is of the form

$$(12.2.10) \quad \sum_{i=1}^n d_i Y_i,$$

where  $d_1, \dots, d_n$  are known, fixed constants. (Exercise 7.39 concerns linear estimators of a population mean.) Among the class of linear estimators, we further restrict attention to unbiased estimators. This restricts the values of  $d_1, \dots, d_n$  that can be used.

An unbiased estimator of the slope  $\beta$  must satisfy

$$E \sum_{i=1}^n d_i Y_i = \beta,$$

regardless of the true value of the parameters  $\alpha$  and  $\beta$ . This implies that

$$\begin{aligned} \beta &= E \sum_{i=1}^n d_i Y_i \\ &= \sum_{i=1}^n d_i EY_i \\ &= \sum_{i=1}^n d_i (\alpha + \beta x_i) \\ &= \alpha \left( \sum_{i=1}^n d_i \right) + \beta \left( \sum_{i=1}^n d_i x_i \right). \end{aligned}$$

This equality is true for *all*  $\alpha$  and  $\beta$  if and only if

$$(12.2.11) \quad \sum_{i=1}^n d_i = 0 \quad \text{and} \quad \sum_{i=1}^n d_i x_i = 1.$$

Thus,  $d_1, \dots, d_n$  must satisfy (12.2.11) in order for the estimator to be an unbiased estimator of  $\beta$ .

In Chapter 7 we called an unbiased estimator “best” if it had the smallest variance among all unbiased estimators. Similarly, an estimator is the *best linear, unbiased estimator (BLUE)* if it is the linear, unbiased estimator with the smallest variance. We will now show that the choice of  $d_i = (x_i - \bar{x})/S_{xx}$  that defines the estimator

$b = S_{xY}/S_{xx}$  is the best choice in that it results in the linear, unbiased estimator of  $\beta$  with the smallest variance. (The  $d_i$ s must be known, fixed constants but the  $x_i$ s are known, fixed constants so this choice of  $d_i$ s is legitimate.)

*A note on notation:* The notation  $S_{xY}$  stresses the fact that  $S_{xY}$  is a random variable that is a function of the random variables  $Y_1, \dots, Y_n$ .  $S_{xY}$  also depends on the nonrandom quantities  $x_1, \dots, x_n$ .

Because  $Y_1, \dots, Y_n$  are uncorrelated with equal variance  $\sigma^2$ , the variance of *any* linear estimator is given by

$$\text{Var} \sum_{i=1}^n d_i Y_i = \sum_{i=1}^n d_i^2 \text{Var} Y_i = \sum_{i=1}^n d_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^n d_i^2.$$

The BLUE of  $\beta$  is, therefore, defined by constants  $d_1, \dots, d_n$  that satisfy (12.2.11) and have the minimum value of  $\sum_{i=1}^n d_i^2$ . (The presence of  $\sigma^2$  has no effect on the minimization over linear estimators since it appears as a multiple of the variance of every linear estimator.)

The minimizing values of the constants  $d_1, \dots, d_n$  can now be found by using Lemma 11.2.1. To apply the lemma to our minimization problem, make the following correspondences, where the left-hand sides are notation from Lemma 11.2.1 and the right-hand sides are our current notation. Let

$$k = n, \quad v_i = x_i, \quad c_i = 1, \quad \text{and} \quad a_i = d_i,$$

which implies  $\bar{v}_c = \bar{x}$ . If  $d_i$  is of the form

$$(12.2.12) \quad d_i = K c_i (v_i - \bar{v}_c) = K(x_i - \bar{x}), \quad i = 1, \dots, n,$$

then, by Lemma 11.2.1,  $d_1, \dots, d_n$  maximizes

$$(12.2.13) \quad \frac{(\sum_{i=1}^n d_i x_i)^2}{\sum_{i=1}^n d_i^2}$$

among all  $d_1, \dots, d_n$  that satisfy  $\sum d_i = 0$ . Furthermore, since

$$\{(d_1, \dots, d_n) : \sum d_i = 0, \sum d_i x_i = 1\} \subset \{(d_1, \dots, d_n) : \sum d_i = 0\},$$

if  $d_i$ s of the form (12.2.12) also satisfy (12.2.11), they certainly maximize (12.2.13) among all  $d_1, \dots, d_n$  that satisfy (12.2.11). (Since the set over which the maximum is taken is smaller, the maximum cannot be larger.) Now, using (12.2.12), we have

$$\sum_{i=1}^n d_i x_i = \sum_{i=1}^n K(x_i - \bar{x}) x_i = K S_{xx}.$$

The second constraint in (12.2.11) is satisfied if  $K = \frac{1}{S_{xx}}$ . Therefore, with  $d_1, \dots, d_n$  defined by

$$(12.2.14) \quad d_i = \frac{(x_i - \bar{x})}{S_{xx}}, \quad i = 1, \dots, n,$$

both constraints of (12.2.11) are satisfied and this set of  $d_i$ 's produces the maximum. Finally, note that for all  $d_1, \dots, d_n$  that satisfy (12.2.11),

$$\frac{(\sum_{i=1}^n d_i x_i)^2}{\sum_{i=1}^n d_i^2} = \frac{1}{\sum_{i=1}^n d_i^2}.$$

Thus, for  $d_1, \dots, d_n$  that satisfy (12.2.11), maximization of (12.2.13) is equivalent to minimization of  $\sum d_i^2$ . Hence, we can conclude that the  $d_i$ 's defined in (12.2.14) give the minimum value of  $\sum d_i^2$  among all  $d_i$ 's that satisfy (12.2.11) and the linear, unbiased estimator defined by these  $d_i$ 's, namely,

$$b = \sum_{i=1}^n \frac{(x_i - \bar{x})}{S_{xx}} y_i = \frac{S_{xy}}{S_{xx}},$$

is the BLUE of  $\beta$ .

A geometric description of this construction of the BLUE of  $\beta$  is given in Figure 12.2.2, where we take  $n = 3$ . The figure shows three-dimensional space with coordinates  $d_1, d_2$ , and  $d_3$ . The two planes represent the vectors  $(d_1, d_2, d_3)$  that satisfy the two linear constraints in (12.2.11) and the line where the two planes intersect are the vectors  $(d_1, d_2, d_3)$  that satisfy both equalities. For any point on the line,  $\sum_{i=1}^n d_i^2$  is the square of the distance from the point to the origin  $\mathbf{0}$ . The vector  $(d_1, d_2, d_3)$  that defines the BLUE is the point on the line that is closest to  $\mathbf{0}$ . The sphere in the figure is the smallest sphere that intersects the line and the point of intersection is the point  $(d_1, d_2, d_3)$  that defines the BLUE of  $\beta$ . This, we have shown, is the point with  $d_i = (x_i - \bar{x})/S_{xx}$ .

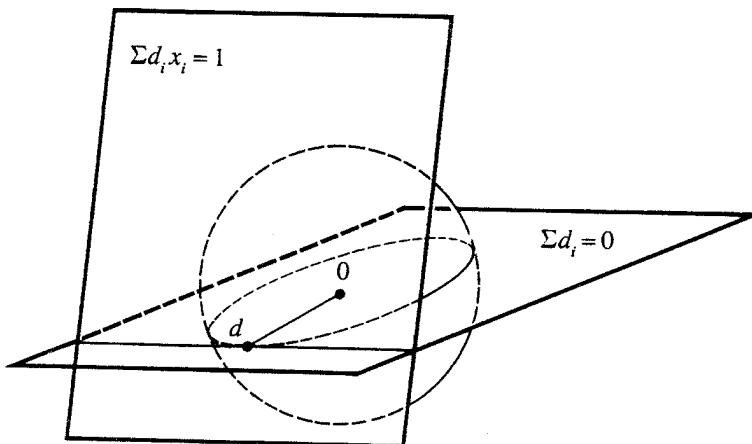


FIGURE 12.2.2 Geometric description of the BLUE

The variance of  $b$  is

$$(12.2.15) \quad \text{Var } b = \sigma^2 \sum_{i=1}^n d_i^2 = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Since  $x_1, \dots, x_n$  are values chosen by the experimenter, they can be chosen to make  $S_{xx}$  large and the variance of the estimator small. That is, the experimenter can *design the experiment* to make the estimator more precise. Suppose that all the  $x_1, \dots, x_n$  must be chosen in an interval  $[e, f]$ . Then, if  $n$  is even, the choice of  $x_1, \dots, x_n$  that makes  $S_{xx}$  as large as possible is to take half of the  $x_i$ s equal to  $e$  and half equal to  $f$  (Exercise 12.2). This would be the best design in that it would give the most precise estimate of the slope  $\beta$  if the experimenter were certain that the model described by (12.2.6) and (12.2.7) was correct. In practice, however, this design is seldom used because an experimenter is hardly ever certain of the model. This *two-point design* gives information about the value of  $E(Y|x)$  at only two values,  $x = e$  and  $x = f$ . If the population regression function  $E(Y|x)$ , which gives the mean of  $Y$  as a function of  $x$ , is nonlinear, it could never be detected from data obtained using the “optimal” two-point design.

We have shown that  $b$  is the BLUE of  $\beta$ . A similar analysis will show that  $a$  is the BLUE of the intercept  $\alpha$ . The constants  $d_1, \dots, d_n$  that define a linear estimator of  $\alpha$  must satisfy

$$(12.2.16) \quad \sum_{i=1}^n d_i = 1 \quad \text{and} \quad \sum_{i=1}^n d_i x_i = 0.$$

The details of this derivation are left as Exercise 12.4. The fact that least squares estimators are BLUES holds in other linear models, also. This general result is called the *Gauss–Markov Theorem* (see Hocking (1985) or the more general treatment in Harville (1981)).

### 12.2.3 Models and Distribution Assumptions

In this section, we will introduce two more models for paired data  $(x_1, y_1), \dots, (x_n, y_n)$  that are called simple linear regression models.

To obtain the least squares estimates in Section 12.2.1, we used no statistical model. We simply solved a mathematical minimization problem. Thus, we could not derive any statistical properties about the estimators obtained by this method because there were no probability models to work with. There are not really any parameters for which we could construct hypothesis tests or confidence intervals.

In Section 12.2.2 we made some statistical assumptions about the data. Specifically, we made assumptions about the first two moments, the mean, variance, and covariance, of the data. These are all statistical assumptions, related to probability models for the data, and we derived statistical properties for the estimators. The properties of unbiasedness and minimum variance, that we proved for the estimators  $a$  and  $b$  of the parameters  $\alpha$  and  $\beta$ , are statistical properties.

To obtain these properties we did not have to specify a complete probability model for the data, only assumptions about the first two moments. We were able to obtain a general optimality property under these minimal assumptions, but the optimality was only in a restricted class of estimators—linear, unbiased estimators. We were not able to derive exact tests and confidence intervals under this model because the model does not specify enough about the probability distribution of the data. We now present two statistical models that completely specify the probabilistic structure of the data.

### Conditional normal model

The *conditional normal model* is the most common simple linear regression model and the most straightforward to analyze. The observed data are the  $n$  pairs,  $(x_1, y_1), \dots, (x_n, y_n)$ . The values of the predictor variable,  $x_1, \dots, x_n$ , are considered to be known, fixed constants. As in Section 12.2.2, think of them as being chosen and set by the experimenter. The values of the response variable,  $y_1, \dots, y_n$ , are observed values of random variables,  $Y_1, \dots, Y_n$ . The random variables  $Y_1, \dots, Y_n$  are assumed to be independent. Furthermore, the distribution of the  $Y_i$ s is normal, specifically,

$$(12.2.17) \quad Y_i \sim n(\alpha + \beta x_i, \sigma^2), \quad i = 1, \dots, n.$$

Thus the population regression function is a linear function of  $x$ , that is,  $E(Y|x) = \alpha + \beta x$ , and all the  $Y_i$ s have the same variance,  $\sigma^2$ . The conditional normal model can be expressed similar to (12.2.8) and (12.2.9). Namely,

$$(12.2.18) \quad Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n,$$

where  $\epsilon_1, \dots, \epsilon_n$  are iid  $n(0, \sigma^2)$  random variables.

The conditional normal model is a special case of the model considered in Section 12.2.2. The population regression function,  $E(Y|x) = \alpha + \beta x$ , and the variance,  $\text{Var } Y = \sigma^2$ , are as in that model. The uncorrelatedness of  $Y_1, \dots, Y_n$  (or, equivalently,  $\epsilon_1, \dots, \epsilon_n$ ) has been strengthened to independence. And, of course, rather than just the first two moments of the distribution of  $Y_1, \dots, Y_n$ , the exact form of the probability distribution is now specified.

The joint pdf of  $Y_1, \dots, Y_n$  is the product of the marginal pdfs, because of the independence. It is given by

$$(12.2.19) \quad \begin{aligned} f(\mathbf{y}|\alpha, \beta, \sigma^2) &= f(y_1, \dots, y_n|\alpha, \beta, \sigma^2) \\ &= \prod_{i=1}^n f(y_i|\alpha, \beta, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp [-(y_i - (\alpha + \beta x_i))^2/(2\sigma^2)] \end{aligned}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp \left[ - \left( \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \right) / (2\sigma^2) \right].$$

It is this joint probability distribution that will be used to develop the statistical procedures in Sections 12.2.4 and 12.2.5. For example, the expression in (12.2.19) will be used to find MLEs of  $\alpha$ ,  $\beta$ , and  $\sigma^2$ .

### Bivariate normal model

In all the previous models we have discussed, the values of the predictor variable,  $x_1, \dots, x_n$ , have been fixed, known constants. But sometimes these values are actually observed values of random variables,  $X_1, \dots, X_n$ . In Galton's example in Section 12.1,  $x_1, \dots, x_n$  were observed heights of fathers. But the experimenter certainly did not choose these heights before collecting the data. Thus it is necessary to consider models in which the predictor variable, as well as the response variable, is random. One such model that is fairly simple is the bivariate normal model. A more complex model is discussed in Section 12.3.

In the bivariate normal model the data  $(x_1, y_1), \dots, (x_n, y_n)$  are observed values of the bivariate random vectors  $(X_1, Y_1), \dots, (X_n, Y_n)$ . The random vectors are independent and the joint distribution of  $(X_i, Y_i)$  is assumed to be bivariate normal. Specifically, it is assumed that

$$(X_i, Y_i) \sim \text{bivariate normal}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho).$$

The joint pdf and various properties of a bivariate normal distribution are given in Definition 4.5.3 and the subsequent discussion. The joint pdf of all the data  $(X_1, Y_1), \dots, (X_n, Y_n)$  is the product of these bivariate pdfs.

In a simple linear regression analysis, we are still thinking of  $x$  as the predictor variable and  $y$  as the response variable. That is, we are most interested in predicting the value of  $y$  having observed the value of  $x$ . This naturally leads to basing inference on the conditional distribution of  $Y$  given  $X = x$ . For a bivariate normal model, the conditional distribution of  $Y$  given  $X = x$  is normal. The population regression function is now a true conditional expectation, as the notation suggests, and is

$$(12.2.20) \quad E(Y|x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X) = \left[ \mu_Y - \rho \frac{\sigma_Y}{\sigma_X} \mu_X \right] + \left[ \rho \frac{\sigma_Y}{\sigma_X} \right] x.$$

The bivariate normal model *implies* that the population regression is a linear function of  $x$ . We need not assume this as in the previous models. Here  $E(Y|x) = \alpha + \beta x$  where  $\beta = \rho \frac{\sigma_Y}{\sigma_X}$  and  $\alpha = \mu_Y - \rho \frac{\sigma_Y}{\sigma_X} \mu_X$ . Also, as in the conditional normal model, the conditional variance of the response variable  $Y$  does not depend on  $x$ ,

$$(12.2.21) \quad \text{Var}(Y|x) = \sigma_Y^2 (1 - \rho^2).$$

For the bivariate normal model, the linear regression analysis is almost always carried out using the conditional distribution of  $(Y_1, \dots, Y_n)$  given  $X_1 = x_1, \dots, X_n = x_n$ , rather than the unconditional distribution of  $(X_1, Y_1), \dots, (X_n, Y_n)$ . But then we

are in the same situation as the conditional normal model described above. The fact that  $x_1, \dots, x_n$  are observed values of random variables is immaterial if we condition on these values and, in general, in simple linear regression we do not use the fact of bivariate normality except to define the conditional distribution. (Indeed, for the most part, the marginal distribution of  $X$  is of no consequence whatsoever. In linear regression it is the conditional distribution that matters.) Inference based on point estimators, intervals, or tests is the same for the two models. See Brown (1990) for an alternative view.

### 12.2.4 Estimation and Testing with Normal Errors

In this and the next subsections we develop inference procedures under the conditional normal model, the regression model defined by (12.2.17) or (12.2.18).

First, we find the maximum likelihood estimates of the three parameters,  $\alpha, \beta$ , and  $\sigma^2$ . Using the joint pdf in (12.2.19), we see that the log likelihood function is

$$\log L(\alpha, \beta, \sigma^2 | \mathbf{x}, \mathbf{y}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2}{2\sigma^2}.$$

For any fixed value of  $\sigma^2$ ,  $\log L$  is maximized as a function of  $\alpha$  and  $\beta$  by those values,  $\hat{\alpha}$  and  $\hat{\beta}$ , that minimize

$$\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2.$$

But this function is just the RSS from Section 12.2.1! There we found that the minimizing values are

$$\hat{\beta} = b = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\alpha} = a = \bar{y} - b\bar{x} = \bar{y} - \hat{\beta}\bar{x}.$$

Thus, the least squares estimators of  $\alpha$  and  $\beta$  are also the MLEs of  $\alpha$  and  $\beta$ . The values  $\hat{\alpha}$  and  $\hat{\beta}$  are the maximizing values, for any fixed value of  $\sigma^2$ . Now, substituting in the log likelihood, to find the MLE of  $\sigma^2$  we need to maximize

$$-\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{2\sigma^2}.$$

This maximization is similar to finding the MLE of  $\sigma^2$  in ordinary normal sampling (Example 7.2.8) and we leave the details to Exercise 12.5. The MLE of  $\sigma^2$ , under the conditional normal model, is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2,$$

the RSS, evaluated at the least squares line, divided by the sample size. Henceforth, when we refer to RSS we mean the RSS evaluated at the least squares line.

In Section 12.2.2, we showed that  $\hat{\alpha}$  and  $\hat{\beta}$  were linear, unbiased estimators of  $\alpha$  and  $\beta$ . However,  $\hat{\sigma}^2$  is not an unbiased estimator of  $\sigma^2$ . For the calculation of  $E\hat{\sigma}^2$  and in many subsequent calculations, the following lemma will be useful.

**LEMMA 12.2.1:** Let  $Y_1, \dots, Y_n$  be uncorrelated random variables with  $\text{Var } Y_i = \sigma^2$  for all  $i = 1, \dots, n$ . Let  $c_1, \dots, c_n$  and  $d_1, \dots, d_n$  be two sets of constants. Then

$$\text{Cov}\left(\sum_{i=1}^n c_i Y_i, \sum_{i=1}^n d_i Y_i\right) = \left(\sum_{i=1}^n c_i d_i\right) \sigma^2.$$

*Proof:* This type of result has been encountered before. It is similar to Lemma 5.4.2 and Exercise 11.11. However, here we do not need either normality or independence of  $Y_1, \dots, Y_n$ .  $\square$

To find the bias in  $\hat{\sigma}^2$ , we proceed as in Section 11.3.2. From (12.2.18) we have

$$\epsilon_i = Y_i - \alpha - \beta x_i.$$

We define the *residuals from the regression* to be

$$(12.2.22) \quad \hat{\epsilon}_i = Y_i - \hat{\alpha} - \hat{\beta} x_i,$$

and thus

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^2 = \frac{1}{n} \text{RSS}.$$

It can be calculated (Exercise 12.6) that

$$E\hat{\epsilon}_i = 0$$

and a lengthy calculation (also in Exercise 12.6) gives

$$(12.2.23) \quad \text{Var } \hat{\epsilon}_i = E\hat{\epsilon}_i^2 = \left( \frac{n-2}{n} + \frac{1}{S_{xx}} \left( \frac{1}{n} \sum_{j=1}^n x_j^2 + x_i^2 - 2(x_i - \bar{x})^2 - 2x_i \bar{x} \right) \right) \sigma^2.$$

Thus,

$$E\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n E\hat{\epsilon}_i^2$$

$$\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^n \left[ \frac{n-2}{n} + \frac{1}{S_{xx}} \left( \frac{1}{n} \sum_{j=1}^n x_j^2 + x_i^2 - 2(x_i - \bar{x})^2 - 2x_i \bar{x} \right) \right] \sigma^2 \\
&= \left[ \frac{n-2}{n} + \frac{1}{nS_{xx}} \left\{ \sum_{j=1}^n x_j^2 + \sum_{i=1}^n x_i^2 - 2S_{xx} - 2 \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right\} \right] \sigma^2 \\
&\quad (\sum x_i \bar{x} = \frac{1}{n}(\sum x_i)^2) \\
&= \left( \frac{n-2}{n} + 0 \right) \sigma^2 \\
&\quad (\sum x_i^2 - \frac{1}{n}(\sum x_i)^2 = S_{xx}) \\
&= \frac{n-2}{n} \sigma^2.
\end{aligned}$$

The MLE  $\hat{\sigma}^2$  is a biased estimator of  $\sigma^2$ . The more commonly used estimator of  $\sigma^2$ , which is unbiased, is

$$(12.2.24) \quad S^2 = \frac{n}{n-2} \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2.$$

To develop estimation and testing procedures, based on these estimators, we need to know their sampling distributions. These are summarized in the following theorem.

**THEOREM 12.2.1:** Under the conditional normal regression model (12.2.17), the sampling distributions of the estimators  $\hat{\alpha}$ ,  $\hat{\beta}$ , and  $S^2$  are

$$\begin{aligned}
\hat{\alpha} &\sim N\left(\alpha, \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^n x_i^2\right), \\
\hat{\beta} &\sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right),
\end{aligned}$$

with

$$\text{Cov}(\hat{\alpha}, \hat{\beta}) = \frac{-\sigma^2 \bar{x}}{S_{xx}}.$$

Furthermore,  $(\hat{\alpha}, \hat{\beta})$  and  $S^2$  are independent and

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi_{n-2}^2.$$

*Proof:* We first show that  $\hat{\alpha}$  and  $\hat{\beta}$  have the indicated normal distributions. The estimators  $\hat{\alpha}$  and  $\hat{\beta}$  are both linear functions of the independent normal random variables

$Y_1, \dots, Y_n$ . Thus, by Corollary 4.6.2, they both have normal distributions. Specifically, in Section 12.2.2, we showed that  $\hat{\beta} = \sum_{i=1}^n d_i Y_i$ , where the  $d_i$  are given in (12.2.14), and we also showed that

$$\mathbb{E}\hat{\beta} = \beta \quad \text{and} \quad \text{Var } \hat{\beta} = \frac{\sigma^2}{S_{xx}}.$$

The estimator  $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$  can be expressed as  $\hat{\alpha} = \sum_{i=1}^n c_i Y_i$  where

$$c_i = \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}},$$

and thus it is straightforward to verify that

$$\mathbb{E}\hat{\alpha} = \sum_{i=1}^n c_i \mathbb{E}Y_i = \sum_{i=1}^n \left( \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right) (\alpha + \beta x_i) = \alpha,$$

and

$$\text{Var } \hat{\alpha} = \sigma^2 \sum_{i=1}^n c_i^2 = \sigma^2 \left[ \frac{1}{n S_{xx}} \sum_{i=1}^n x_i^2 \right],$$

showing that  $\hat{\alpha}$  and  $\hat{\beta}$  have the specified distributions. Also,  $\text{Cov}(\hat{\alpha}, \hat{\beta})$  is easily calculated using Lemma 12.2.1. Details are left to Exercise 12.7.

We next show that  $\hat{\alpha}$  and  $\hat{\beta}$  are independent of  $S^2$ , a fact that will follow from Lemma 12.2.1 and Lemma 5.4.2. From the definition of  $\hat{\epsilon}_i$  in (12.2.22), we can write

$$(12.2.25) \quad \hat{\epsilon}_i = \sum_{j=1}^n [\delta_{ij} - (c_j + d_j x_i)] Y_i,$$

where

$$\begin{aligned} \delta_{ij} &= \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}, \\ c_j &= \frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{S_{xx}}, \end{aligned}$$

and

$$d_j = \frac{(x_j - \bar{x})}{S_{xx}}.$$

Since  $\hat{\alpha} = \sum c_i Y_i$  and  $\hat{\beta} = \sum d_i Y_i$ , application of Lemma 12.2.1 together with some algebra will show that

$$\text{Cov}(\hat{\epsilon}_i, \hat{\alpha}) = \text{Cov}(\hat{\epsilon}_i, \hat{\beta}) = 0, \quad i = 1, \dots, n.$$

Details are left to Exercise 12.8. Thus, it follows from Lemma 5.4.2 that, under normal sampling,  $S^2 = \sum \hat{\epsilon}_i^2 / (n - 2)$  is independent of  $\hat{\alpha}$  and  $\hat{\beta}$ .

To prove that  $(n - 2)S^2 / \sigma^2 \sim \chi_{n-2}^2$ , we write  $(n - 2)S^2$  as the sum of  $n - 2$  independent random variables, each of which has a  $\chi_1^2$  distribution. That is, we find constants  $a_{ij}, i = 1, \dots, n$  and  $j = 1, \dots, n - 2$ , that satisfy

$$(12.2.26) \quad \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{j=1}^{n-2} \left( \sum_{i=1}^n a_{ij} Y_i \right)^2$$

where

$$\sum_{i=1}^n a_{ij} = 0, \quad j = 1, \dots, n - 2 \quad \text{and} \quad \sum_{i=1}^n a_{ij} a_{ij'} = 0, \quad j \neq j'.$$

The details are somewhat similar to those in Theorem 11.3.1 but, unfortunately, more involved because of the general nature of the  $x_i$ s. We omit details.  $\square$

The RSS from the *linear* regression contains information about the worth of a polynomial fit of a higher order, over and above a linear fit. Since, in this model, we assume that the population regression is linear, the variation in this higher-order fit is just random variation. Robson (1959) gives a general recursion formula for finding coefficients for such higher-order polynomial fits, a formula that can be adapted to explicitly find the  $a_{ij}$ s of (12.2.26). Alternatively, Cochran's Theorem (Miscellanea section of Chapter 11) can be used to establish that  $\sum \hat{\epsilon}_i^2 / \sigma^2 \sim \chi_{n-2}^2$ .

Inferences regarding the two parameters  $\alpha$  and  $\beta$  are usually based on the following two Student's  $t$  distributions. Their derivations follow immediately from the normal and  $\chi^2$  distributions and the independence in Theorem 12.2.1. We have

$$(12.2.27) \quad \frac{\hat{\alpha} - \alpha}{S \sqrt{(\sum_{i=1}^n x_i^2) / (n S_{xx})}} \sim t_{n-2}$$

and

$$(12.2.28) \quad \frac{\hat{\beta} - \beta}{S / \sqrt{S_{xx}}} \sim t_{n-2}.$$

The joint distribution of these two  $t$  statistics is called a *bivariate Student's t distribution*. This distribution is derived in a manner analogous to the univariate case. We use the fact that the joint distribution of  $\hat{\alpha}$  and  $\hat{\beta}$  is bivariate normal and the same variance estimate  $S$  is used in both univariate  $t$  statistics. This joint distribution would be used if we wanted to do simultaneous inference regarding  $\alpha$  and  $\beta$ . However, we shall deal only with the inferences regarding one parameter at a time.

Usually there is more interest in  $\beta$  than in  $\alpha$ . The parameter  $\alpha$  is the expected value of  $Y$  at  $x = 0$ ,  $E(Y|x = 0)$ . Depending on the problem, this may or may not

be an interesting quantity. In particular, the value  $x = 0$  may not be a reasonable value for the predictor variable. However,  $\beta$  is the rate of change of  $E(Y|x)$  as a function of  $x$ . That is,  $\beta$  is the amount that  $E(Y|x)$  changes if  $x$  is changed by one unit. Thus, this parameter relates to the entire range of  $x$  values and contains the information about whatever linear relationship exists between  $Y$  and  $x$ . (See Exercise 12.10.) Furthermore, the value  $\beta = 0$  is of particular interest.

If  $\beta = 0$ , then  $E(Y|x) = \alpha + \beta x = \alpha$  and  $Y \sim N(\alpha, \sigma^2)$ , which does not depend on  $x$ . In a well-thought-out experiment leading to a regression analysis we do not expect this to be the case, but we would be interested in knowing this if it were true.

The test that  $\beta = 0$  is quite similar to the ANOVA test that all treatments are equal. In the ANOVA, the null hypothesis states that the treatments are unrelated to the response *in any way*, while in linear regression the null hypothesis  $\beta = 0$  states that the treatments ( $x$ ) are unrelated to the response in a linear way.

To test

$$(12.2.29) \quad H_0: \beta = 0 \quad \text{versus} \quad H_1: \beta \neq 0,$$

using (12.2.28), we reject  $H_0$  at level  $\alpha$  if

$$\left| \frac{\hat{\beta} - 0}{S/\sqrt{S_{xx}}} \right| > t_{n-2,\alpha/2},$$

or, equivalently, if

$$(12.2.30) \quad \frac{\hat{\beta}^2}{S^2/S_{xx}} > F_{1,n-2,\alpha}.$$

Recalling the formula for  $\hat{\beta}$  and that  $RSS = \sum \hat{\epsilon}_i^2$ , we have

$$\frac{\hat{\beta}^2}{S^2/S_{xx}} = \frac{S_{xy}^2/S_{xx}}{RSS/(n-2)} = \frac{\text{Regression Sum of Squares}}{\text{Residual Sum of Squares/df}}.$$

This last formula is summarized in the *regression ANOVA table*, which is like the ANOVA tables encountered in Chapter 11. For simple linear regression, the table, resulting in the test given in (12.2.30), is given in Table 12.2.2. Note that the table involves only a hypothesis about  $\beta$ . The parameter  $\alpha$  and the estimate  $\hat{\alpha}$  play the same role here as the grand mean did in Chapter 11. They merely serve to locate the overall level of the data and are “corrected” for in the sums of squares.

**TABLE 12.2.2** ANOVA table for simple linear regression

Source of variation	Degrees of freedom	Sum of squares	Mean square	F statistic
Regression (slope)	1	Reg. SS = $S_{xy}^2 / S_{xx}$	MS(Reg) = Reg. SS	$F = \frac{MS(Reg)}{MS(Resid)}$
Residual	$n - 2$	RSS = $\sum \epsilon_i^2$	MS(Resid) = RSS/(n - 2)	
Total	$n - 1$	SST = $\sum (y_i - \bar{y})^2$		

**Example 12.1.1 (Continued):** The regression ANOVA for the grape crop yield data is

ANOVA table for grape data

Source of variation	Degrees of freedom	Sum of squares	Mean square	F statistic
Regression	1	6.66	6.66	50.23
Residual	10	1.33	.133	
Total	11	7.99		

showing a highly significant slope of the regression line. ||

We draw one final parallel with the analysis of variance. It may not be obvious from Table 12.2.2, but the partitioning of the sum of squares of the ANOVA has an analogue in regression. We have

$$\text{Total Sum of Squares} = \text{Regression Sum of Squares} + \text{Residual Sum of Squares} \quad (12.2.31)$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

where  $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ . Notice the similarity of these sums of squares to those in ANOVA. The total sum of squares is, of course, the same. The RSS measures deviation of the fitted line from the observed values, and the regression sum of squares, analogous to the ANOVA treatment sum of squares, measures the deviation of predicted values ("treatment means") from the grand mean. Also, as in the ANOVA, the sum of squares identity is valid because of the disappearance of the cross-term (Exercise 12.11). The total and residual sums of squares in (12.2.31) are clearly the same as in Table 12.2.2. But the regression sum of squares looks different. However, they are equal (Exercise 12.11). That is,

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \frac{S_{xy}^2}{S_{xx}}.$$

The expression  $S_{xy}^2/S_{xx}$  is easier to use for computing and provides the link with the  $t$  test. But,  $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  is the more easily interpreted expression.

A statistic that is used to quantify how well the fitted line describes the data is the *coefficient of determination*. It is defined as the ratio of the regression sum of squares to the total sum of squares. It is usually referred to as  $r^2$ , and can be written in the various forms

$$r^2 = \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{S_{xy}^2}{S_{xx} S_{yy}}.$$

The coefficient of determination measures the proportion of the total variation in  $y_1, \dots, y_n$  (measured by  $S_{yy}$ ) that is explained by the fitted line (measured by the regression sum of squares). From (12.2.31),  $0 \leq r^2 \leq 1$ . If  $y_1, \dots, y_n$  all fall exactly on the fitted line, then  $y_i = \hat{y}_i$  for all  $i$  and  $r^2 = 1$ . If  $y_1, \dots, y_n$  are not close to the fitted line, then the residual sum of squares will be large and  $r^2$  will be near 0. The coefficient of determination can also be (perhaps more straightforwardly) derived as the square of the sample correlation coefficient of the  $n$  pairs  $(y_1, x_1), \dots, (y_n, x_n)$  or of the  $n$  pairs  $(y_1, \hat{y}_1), \dots, (y_n, \hat{y}_n)$ .

Expression (12.2.28) can be used to construct a  $100(1 - \alpha)\%$  confidence interval for  $\beta$  given by

$$(12.2.32) \quad \hat{\beta} - t_{n-2,\alpha/2} \frac{S}{\sqrt{S_{xx}}} < \beta < \hat{\beta} + t_{n-2,\alpha/2} \frac{S}{\sqrt{S_{xx}}}.$$

Also, a level  $\alpha$  test of  $H_0: \beta = \beta_0$  versus  $H_1: \beta \neq \beta_0$  rejects  $H_0$  if

$$(12.2.33) \quad \left| \frac{\hat{\beta} - \beta_0}{S/\sqrt{S_{xx}}} \right| > t_{n-2,\alpha/2}.$$

As mentioned above, it is common to test  $H_0: \beta = 0$  versus  $H_1: \beta \neq 0$  to determine if there is some linear relationship between the predictor and response variables. However, the above test is more general, since any value of  $\beta_0$  can be specified. The regression ANOVA, which is locked into a “recipe,” can test only  $H_0: \beta = 0$ .

### 12.2.5 Estimation and Prediction at a Specified $x = x_0$

Associated with a specified value of the predictor variable, say  $x = x_0$ , there is a population of  $Y$  values. In fact, according to the conditional normal model, a random observation from this population is  $Y \sim n(\alpha + \beta x_0, \sigma^2)$ . After observing the regression data  $(x_1, y_1), \dots, (x_n, y_n)$  and estimating the parameters  $\alpha, \beta$ , and  $\sigma^2$ , perhaps the experimenter is going to set  $x = x_0$  and obtain a new observation, call it  $Y_0$ . There might be interest in estimating the mean of the population from which this

observation will be drawn, or even predicting what this observation will be. We will now discuss these types of inferences.

We assume that  $(x_1, Y_1), \dots, (x_n, Y_n)$  satisfy the conditional normal regression model and based on these  $n$  observations we have the estimates  $\hat{\alpha}$ ,  $\hat{\beta}$ , and  $S^2$ . Let  $x_0$  be a specified value of the predictor variable. First, consider estimating the mean of the  $Y$  population associated with  $x_0$ , that is,  $E(Y|x_0) = \alpha + \beta x_0$ . The obvious choice for our point estimator is  $\hat{\alpha} + \hat{\beta}x_0$ . This is an unbiased estimator since  $E(\hat{\alpha} + \hat{\beta}x_0) = E\hat{\alpha} + (E\hat{\beta})x_0 = \alpha + \beta x_0$ . Using the moments given in Theorem 12.2.1, we can also calculate

$$\begin{aligned}\text{Var}(\hat{\alpha} + \hat{\beta}x_0) &= \text{Var } \hat{\alpha} + (\text{Var } \hat{\beta})x_0^2 + 2x_0 \text{Cov}(\hat{\alpha}, \hat{\beta}) \\ &= \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^n x_i^2 + \frac{\sigma^2 x_0^2}{S_{xx}} - \frac{2\sigma^2 x_0 \bar{x}}{S_{xx}} \\ &= \frac{\sigma^2}{S_{xx}} \left( \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 + \bar{x}^2 - 2x_0 \bar{x} + x_0^2 \right) \quad (\pm \bar{x}) \\ &= \frac{\sigma^2}{S_{xx}} \left( \frac{1}{n} \left[ \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right] + (x_0 - \bar{x})^2 \right) \quad (\text{recombine terms}) \\ &= \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right). \quad (\sum x_i^2 - \frac{1}{n}(\sum x_i)^2 = S_{xx})\end{aligned}$$

Finally, since  $\hat{\alpha}$  and  $\hat{\beta}$  are both linear functions of  $Y_1, \dots, Y_n$ , so is  $\hat{\alpha} + \hat{\beta}x_0$ . Thus  $\hat{\alpha} + \hat{\beta}x_0$  has a normal distribution, specifically,

$$(12.2.34) \quad \hat{\alpha} + \hat{\beta}x_0 \sim N \left( \alpha + \beta x_0, \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \right).$$

By Theorem 12.2.1,  $(\hat{\alpha}, \hat{\beta})$  and  $S^2$  are independent. Thus  $S^2$  is also independent of  $\hat{\alpha} + \hat{\beta}x_0$  (Theorem 4.6.5) and

$$\frac{\hat{\alpha} + \hat{\beta}x_0 - (\alpha + \beta x_0)}{S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

This pivot can be inverted to give the  $100(1 - \alpha)\%$  confidence interval for  $\alpha + \beta x_0$ ,

$$\begin{aligned}(12.2.35) \quad \hat{\alpha} + \hat{\beta}x_0 - t_{n-2, \alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} &< \alpha + \beta x_0 \\ &< \hat{\alpha} + \hat{\beta}x_0 + t_{n-2, \alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.\end{aligned}$$

The length of the confidence interval for  $\alpha + \beta x_0$  depends on the values of  $x_1, \dots, x_n$  through the value of  $(x_0 - \bar{x})^2/S_{xx}$ . It is clear that the length of the interval is shorter if  $x_0$  is near  $\bar{x}$  and minimized at  $x_0 = \bar{x}$ . Thus, in designing the experiment, the experimenter should choose the values  $x_1, \dots, x_n$  so that the value  $x_0$ , at which the mean is to be estimated, is at or near  $\bar{x}$ . It is only reasonable that we can estimate more precisely near the center of the data we observed.

A type of inference we have not discussed until now is *prediction* of an, as yet, unobserved random variable  $Y$ , a type of inference that is of interest in a regression setting. For example, suppose that  $x$  is a college applicant's measure of high school performance. A college admissions office might want to use  $x$  to predict  $Y$ , the student's grade point average after one year of college. Clearly,  $Y$  has not been observed yet since the student has not even been admitted! The college has data on former students,  $(x_1, y_1), \dots, (x_n, y_n)$ , giving their high school performances and one-year GPAs. These data might be used to predict the new student's GPA.

**DEFINITION 12.2.1:** A  $100(1 - \alpha)\%$  *prediction interval* for an unobserved random variable  $Y$  based on the observed data  $\mathbf{X}$  is a random interval  $[L(\mathbf{X}), U(\mathbf{X})]$  with the property that

$$P_\theta(L(\mathbf{X}) \leq Y \leq U(\mathbf{X})) \geq 1 - \alpha,$$

for all values of the parameter  $\theta$ .

Note the similarity in the definitions of a prediction interval and a confidence interval. The difference is that a prediction interval is an interval on a random variable, rather than a parameter. Intuitively, since a random variable is more variable than a parameter (which is constant), we expect a prediction interval to be wider than a confidence interval of the same level. In the special case of linear regression, we see that this is the case.

We assume that the new observation  $Y_0$  to be taken at  $x = x_0$  has a  $n(\alpha + \beta x_0, \sigma^2)$  distribution, independent of the previous data,  $(x_1, Y_1), \dots, (x_n, Y_n)$ . The estimators  $\hat{\alpha}$ ,  $\hat{\beta}$ , and  $S^2$  are calculated from the previous data and, thus,  $Y_0$  is independent of  $\hat{\alpha}$ ,  $\hat{\beta}$ , and  $S^2$ . Using (12.2.34), we find that  $Y_0 - (\hat{\alpha} + \hat{\beta}x_0)$  has a normal distribution with mean  $E(Y_0 - (\hat{\alpha} + \hat{\beta}x_0)) = \alpha + \beta x_0 - (\alpha + \beta x_0) = 0$ , and variance

$$\begin{aligned}\text{Var}(Y_0 - (\hat{\alpha} + \hat{\beta}x_0)) &= \text{Var } Y_0 + \text{Var } (\hat{\alpha} + \hat{\beta}x_0) \\ &= \sigma^2 + \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right).\end{aligned}$$

Using the independence of  $S^2$  and  $Y_0 - (\hat{\alpha} + \hat{\beta}x_0)$ , we see that

$$T = \frac{Y_0 - (\hat{\alpha} + \hat{\beta}x_0)}{S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2},$$

which can be rearranged in the usual way to obtain the  $100(1-\alpha)\%$  prediction interval,

$$(12.2.36) \quad \begin{aligned} \hat{\alpha} + \hat{\beta}x_0 - t_{n-2,\alpha/2}S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} &< Y_0 \\ &< \hat{\alpha} + \hat{\beta}x_0 + t_{n-2,\alpha/2}S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}. \end{aligned}$$

Since the endpoints of this interval depend only on the observed data, (12.2.36) defines a prediction interval for the new observation  $Y_0$ .

### 12.2.6 Simultaneous Estimation and Confidence Bands

In the previous section we looked at prediction at a single value  $x_0$ . In some circumstances, however, there may be interest in prediction at many  $x_0$ s. For example, in the previously mentioned grade point average prediction problem, an admissions office probably has interest in predicting the grade point average of many applicants, which naturally leads to prediction at many  $x_0$ s.

The problem encountered is the (by now) familiar problem of simultaneous inference. That is, how do we control the overall confidence level for the simultaneous inference? In the previous section, we saw that a  $1 - \alpha$  confidence interval for the mean of the  $Y$  population associated with  $x_0$ , that is,  $E(Y|x_0) = \alpha + \beta x_0$ , is given by

$$\begin{aligned} \hat{\alpha} + \hat{\beta}x_0 - t_{n-2,\alpha/2}S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \\ &< \alpha + \beta x_0 < \hat{\alpha} + \hat{\beta}x_0 + t_{n-2,\alpha/2}S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}. \end{aligned}$$

Now suppose that we want to make an inference about the  $Y$  population mean at a number of  $x_0$  values. For example, we might want intervals for  $E(Y|x_{0i})$ ,  $i = 1, \dots, m$ . We know that if we set up  $m$  intervals as above, each at level  $1 - \alpha$ , the overall inference will not be at the  $1 - \alpha$  level.

A simple and reasonably good solution is to use the Bonferroni Inequality, as used in Example 11.2.4. Using the inequality, we can state that the probability is at least  $1 - \alpha$  that

$$(12.2.37) \quad \begin{aligned} \hat{\alpha} + \hat{\beta}x_{0i} - t_{n-2,\alpha/(2m)}S\sqrt{\frac{1}{n} + \frac{(x_{0i} - \bar{x})^2}{S_{xx}}} &< \alpha + \beta x_{0i} \\ &< \hat{\alpha} + \hat{\beta}x_{0i} + t_{n-2,\alpha/(2m)}S\sqrt{\frac{1}{n} + \frac{(x_{0i} - \bar{x})^2}{S_{xx}}}, \end{aligned}$$

simultaneously for  $i = 1, \dots, m$ . (See Exercise 12.17.)

We can take simultaneous inference in regression one step further. Realize that our assumption about the population regression line implies that the equation  $E(Y|x) = \alpha + \beta x$  holds for all  $x$ ; hence, we should be able to make inferences at all  $x$ . Thus, we want to make a statement like (12.2.37), but we want it to hold for all  $x$ . As might be expected, as he did for the ANOVA, Scheffé derived a solution for this problem. We summarize the result, for the case of simple linear regression, in the following theorem.

**THEOREM 12.2.2:** Under the conditional normal regression model (12.2.17), the probability is at least  $1 - \alpha$  that

(12.2.38)

$$\hat{\alpha} + \hat{\beta}x - M_\alpha S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} < \alpha + \beta x < \hat{\alpha} + \hat{\beta}x + M_\alpha S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}},$$

simultaneously for all  $x$ , where  $M_\alpha = \sqrt{2F_{2,n-2,\alpha}}$ .

*Proof:* By rearranging terms, it should be clear that the conclusion of the theorem is true if we can find a constant  $M_\alpha$  that satisfies

$$P\left(\frac{((\hat{\alpha} + \hat{\beta}x) - (\alpha + \beta x))^2}{S^2 \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right]} \leq M_\alpha^2, \text{ for all } x\right) = 1 - \alpha$$

or, equivalently,

$$P\left(\max_x \frac{((\hat{\alpha} + \hat{\beta}x) - (\alpha + \beta x))^2}{S^2 \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right]} \leq M_\alpha^2\right) = 1 - \alpha.$$

The parameterization given in Exercise 12.9, which results in independent estimators for  $\alpha$  and  $\beta$ , makes the above maximization easier. Write

$$\begin{aligned}\hat{\alpha} + \hat{\beta}x &= \bar{Y} + \hat{\beta}(x - \bar{x}), \\ \alpha + \beta x &= \mu_{\bar{Y}} + \beta(x - \bar{x}), \quad (\mu_{\bar{Y}} = E\bar{Y} = \alpha + \beta\bar{x})\end{aligned}$$

and, for notational convenience, define  $t = x - \bar{x}$ . We then have

$$\frac{((\hat{\alpha} + \hat{\beta}x) - (\alpha + \beta x))^2}{S^2 \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right]} = \frac{((\bar{Y} - \mu_{\bar{Y}}) + (\hat{\beta} - \beta)t)^2}{S^2 \left[ \frac{1}{n} + \frac{t^2}{S_{xx}} \right]},$$

and we want to find  $M_\alpha$  to satisfy

$$P \left( \max_t \frac{\left( (\bar{Y} - \mu_{\bar{Y}}) + (\hat{\beta} - \beta)t \right)^2}{S^2 \left[ \frac{1}{n} + \frac{t^2}{S_{xx}} \right]} \leq M_\alpha^2 \right) = 1 - \alpha.$$

Note that  $S^2$  plays no role in the maximization, merely being a constant. Applying the result of Exercise 12.18, a direct application of calculus, we obtain

$$(12.2.39) \quad \begin{aligned} \max_t \frac{\left( (\bar{Y} - \mu_{\bar{Y}}) + (\hat{\beta} - \beta)t \right)^2}{S^2 \left[ \frac{1}{n} + \frac{t^2}{S_{xx}} \right]} &= \frac{n(\bar{Y} - \mu_{\bar{Y}})^2 + S_{xx}(\hat{\beta} - \beta)^2}{S^2} \\ &= \frac{\frac{(\bar{Y} - \mu_{\bar{Y}})^2}{\sigma^2/n} + \frac{(\hat{\beta} - \beta)^2}{\sigma^2/S_{xx}}}{S^2/\sigma^2}. \end{aligned}$$

(multiply by  $\sigma^2/\sigma^2$ )

From Theorem 12.2.1 and Exercise 12.9, we see that this last expression is the quotient of independent chi squared random variables, the denominator being divided by its degrees of freedom. The numerator is the sum of two independent random variables, each of which has a  $\chi_1^2$  distribution. Thus the numerator is distributed as  $\chi_2^2$ , the distribution of the quotient is

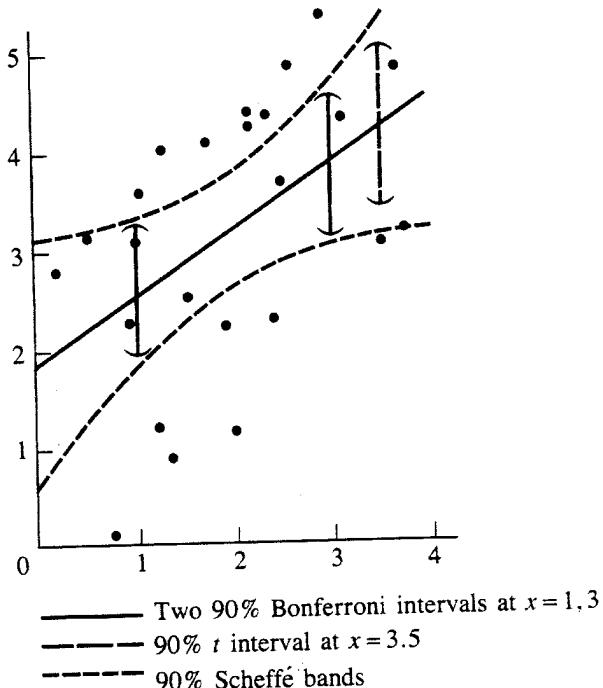
$$\frac{\frac{(\bar{Y} - \mu_{\bar{Y}})^2}{\sigma^2/n} + \frac{(\hat{\beta} - \beta)^2}{\sigma^2/S_{xx}}}{S^2/\sigma^2} \sim 2F_{2,n-2},$$

and

$$P \left( \max_t \frac{\left( (\bar{Y} - \mu_{\bar{Y}}) + (\hat{\beta} - \beta)t \right)^2}{S^2 \left[ \frac{1}{n} + \frac{t^2}{S_{xx}} \right]} \leq M_\alpha^2 \right) = 1 - \alpha$$

if  $M_\alpha = \sqrt{2F_{2,n-2}}$ , proving the theorem.  $\square$

Since (12.2.38) is true for all  $x$ , it actually gives a *confidence band* on the entire population regression line. That is, as a confidence interval covers a single-valued parameter, a confidence band covers an entire line with a band. An example of the Scheffé band is given in Figure 12.2.3 on page 580, along with two Bonferroni intervals and a single  $t$  interval. Notice that, although it is not the case in Figure 12.2.3, it is possible for the Bonferroni intervals to be *wider* than the Scheffé bands,



**FIGURE 12.2.3** Scheffé bands and Bonferroni and  $t$  intervals for data in Table 12.2.1

even though the Bonferroni inference (necessarily) pertains to fewer intervals. This will be the case whenever

$$t_{n-2,\alpha/(2m)} > 2F_{2,n-2,\alpha},$$

where  $m$  is defined as in (12.2.37). The inequality will always be satisfied for large enough  $m$ , so there will always be a point where it pays to switch from Bonferroni to Scheffé, even if there is interest in only a finite number of  $x$ s. This “phenomenon,” that we seem to get something for nothing, occurs because the Bonferroni Inequality is an all-purpose bound while the Scheffé band is an exact solution for the problem at hand. (The actual coverage probability for the Bonferroni intervals is higher than  $1 - \alpha$ .)

There are many variations on the Scheffé band. Some variations have different shapes and some guarantee coverage for only a particular interval of  $x$  values. See the *Miscellanea* section for a discussion of these alternate bands.

In theory, the proof of Theorem 12.2.2, with suitable modifications, can result in simultaneous prediction intervals. (In fact, the maximization of the function in Exercise 12.18 gives the result almost immediately.) The problem, however, is that the resulting statistic does not have a particularly nice distribution.

Finally, we note a problem about using procedures like the Scheffé band to make inferences at  $x$  values that are outside the range of the observed  $x$ s. Such procedures are based on the assumption that we *know* the population regression function is linear for all  $x$ . Although it may be reasonable to assume the regression function is linear over the range of  $x$ s observed, *extrapolation* to  $x$ s outside the observed range is usually unwise. (Since there are no data outside the observed range, we cannot check

if the regression becomes nonlinear.) This caveat also applies to the procedures in Section 12.2.5.

## 12.3 Regression with Errors in Variables

Regression with *errors in variables* (*EIV*), also known as the *measurement error model*, is so fundamentally different from the simple linear regression of the previous sections that it is probably best thought of as a completely different topic. It is presented as a generalization of the usual regression model mainly for traditional reasons. However, the problems that arise with this model are very different.

The models of this section are generalizations of simple linear regression in that we will work with models of the form

$$(12.3.1) \quad Y_i = \alpha + \beta x_i + \epsilon_i,$$

but now we do not assume that the  $x$ s are known. Instead, we can measure a random variable whose mean is  $x_i$ . (In keeping with our notational conventions, we will speak of measuring a random variable  $X_i$  whose mean is not  $x_i$  but  $\xi_i$ .)

The intention here is to illustrate different approaches to the EIV model, showing some of the standard solutions and the (sometimes) unexpected difficulties that arise. For a more thorough introduction to this problem, there is the review article by Anderson (1984b) and a comprehensive text by Fuller (1987). Kendall and Stuart (1979) also treat this topic in some detail.

In the general EIV model we assume that we observe pairs  $(x_i, y_i)$ , sampled from random variables  $(X_i, Y_i)$  whose means satisfy the linear relationship

$$(12.3.2) \quad EY_i = \alpha + \beta(EX_i).$$

If we define

$$EY_i = \eta_i \quad \text{and} \quad EX_i = \xi_i,$$

then the relationship (12.3.2) becomes

$$(12.3.3) \quad \eta_i = \alpha + \beta\xi_i,$$

a linear relationship between the means of the random variables.

The variables  $\xi_i$  and  $\eta_i$  are sometimes called *latent variables*, a term that refers to quantities that cannot be directly measured. Latent variables may be not only impossible to measure directly, but impossible to measure *at all*. For example, the IQ of a person is impossible to measure. We can measure a score on an IQ test but we can never measure the variable IQ. Relationships between IQ and other variables, however, are often hypothesized.

The model specified in (12.3.2) really makes no distinction between  $X$  and  $Y$ . If we are interested in a regression, however, there should be a reason for choosing

$Y$  as the response and  $X$  as the predictor. Keeping this specification in mind, of regressing  $Y$  on  $X$ , we define the *errors in variables model* or *measurement error model* as the following.

Observe pairs  $(X_i, Y_i), i = 1, \dots, n$ , according to

$$(12.3.4) \quad \begin{aligned} Y_i &= \alpha + \beta \xi_i + \epsilon_i, & \epsilon_i &\sim n(0, \sigma_\epsilon^2), \\ X_i &= \xi_i + \delta_i, & \delta_i &\sim n(0, \sigma_\delta^2). \end{aligned}$$

Note that the assumption of normality, although common, is not necessary. Other distributions can be used. In fact, some of the problems encountered with this model are caused by the normality assumption. (See, for example, Solari (1969).)

**Example 12.3.1:** The EIV regression model arises fairly naturally in situations where the  $x$  variable is observed along with the  $y$  variable (rather than being controlled). For example, in the 1800s the Scottish physicist J. D. Forbes tried to use measurements on the boiling temperature of water to estimate altitude above sea level. To do this, he simultaneously measured boiling temperature and atmospheric pressure (from which altitude can be obtained). Since barometers were quite fragile in the 1800s, it would be useful to estimate pressure, or more precisely,  $\log(\text{pressure})$ , from temperature. The data observed at 9 locales are

Boiling point ( $^{\circ}\text{F}$ )	$\log(\text{Pressure})$ ( $\log(\text{Hg})$ )
194.5	1.3179
197.9	1.3502
199.4	1.3646
200.9	1.3782
201.4	1.3806
203.6	1.4004
209.5	1.4547
210.7	1.4630
212.2	1.4780

and an EIV model is reasonable for this situation. ||

A number of special cases of the model (12.3.4) have already been seen. If  $\delta_i = 0$ , then the model becomes simple linear regression (since there is now no measurement error, we can directly observe the  $\xi_i$ s). If  $\alpha = 0$ , then we have

$$\begin{aligned} Y_i &\sim n(\eta_i, \sigma_\epsilon^2), & i &= 1, \dots, n, \\ X_i &\sim n(\xi_i, \sigma_\delta^2), & i &= 1, \dots, n, \end{aligned}$$

where, possibly,  $\sigma_\delta^2 \neq \sigma_\epsilon^2$ , a version of the Behrens–Fisher problem (previously mentioned in Section 11.2.1).

### 12.3.1 Functional and Structural Relationships

There are two different types of relationship that can be specified in the EIV model, one that specifies a *functional* linear relationship and one describing a *structural* linear relationship. The different relationship specifications can lead to different estimators with different properties. As said by Moran (1971), “This is not very happy terminology, but we will stick to it because the distinction is essential...” Some interpretations of this terminology are given in the *Miscellanea* section. For now we merely present the two models.

#### *Linear functional relationship model*

This is the model as presented in (12.3.4) where we have random variables  $X_i$  and  $Y_i$ , with  $EX_i = \xi_i$ ,  $EY_i = \eta_i$ , and we assume the *functional relationship*

$$\eta_i = \alpha + \beta\xi_i.$$

We observe pairs  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ , according to

$$(12.3.5) \quad Y_i = \alpha + \beta\xi_i + \epsilon_i, \quad \epsilon_i \sim n(0, \sigma_\epsilon^2),$$

$$X_i = \xi_i + \delta_i, \quad \delta_i \sim n(0, \sigma_\delta^2),$$

where  $\xi_i$ s are fixed, unknown parameters and the  $\epsilon_i$ s and  $\delta_i$ s are independent. The parameters of main interest are  $\alpha$  and  $\beta$  and inference on these parameters is made using the joint distribution of  $((X_1, Y_1), \dots, (X_n, Y_n))$ , *conditional on*  $\xi_1, \dots, \xi_n$ .

#### *Linear structural relationship model*

This model can be thought of as an extension of the functional relationship model, extended through the following hierarchy. As in the functional relationship model, we have random variables  $X_i$  and  $Y_i$ , with  $EX_i = \xi_i$ ,  $EY_i = \eta_i$ , and we assume the *functional relationship*  $\eta_i = \alpha + \beta\xi_i$ . But now we assume that the parameters  $\xi_1, \dots, \xi_n$  are themselves a random sample from a common population. Thus, conditional on  $\xi_1, \dots, \xi_n$ , we observe pairs  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ , according to

$$(12.3.6) \quad Y_i = \alpha + \beta\xi_i + \epsilon_i, \quad \epsilon_i \sim n(0, \sigma_\epsilon^2),$$

$$X_i = \xi_i + \delta_i, \quad \delta_i \sim n(0, \sigma_\delta^2),$$

and, also,

$$\xi_i \sim \text{iid } n(\xi, \sigma_\xi^2).$$

As before, the  $\epsilon_i$ s and  $\delta_i$ s are independent and they are also independent of the  $\xi_i$ s. As in the functional relationship model, the parameters of main interest are  $\alpha$  and  $\beta$ . Here, however, the inference on these parameters is made using the joint distribution of  $((X_1, Y_1), \dots, (X_n, Y_n))$ , *unconditional on*  $\xi_1, \dots, \xi_n$ . (That is,  $\xi_1, \dots, \xi_n$  are integrated out according to the distribution in (12.3.6).)

The two models are quite similar in that statistical properties of estimators in one model (for example, consistency) often carry over into the other model. More precisely, estimators that are consistent in the functional model are also consistent in the structural model (Nussbaum (1976) or Gleser (1983)). This makes sense, as the functional model is a “conditional version” of the structural model. Estimators that are consistent in the functional model must be so for all values of the  $\xi_i$ s, so are necessarily consistent in the structural model, which averages over the  $\xi_i$ s. The converse implication is false. However, there is a useful implication which goes from the structural to the functional relationship model. If a parameter is not *identifiable* in the structural model, it is also not identifiable in the functional model. (See Definition 11.2.1.)

As we shall see, the models share similar problems and, in certain situations, similar likelihood solutions. It is probably easier to do statistical theory in the structural model, while the functional model often seems to be the more reasonable model for many situations. Thus, the underlying similarities come in handy.

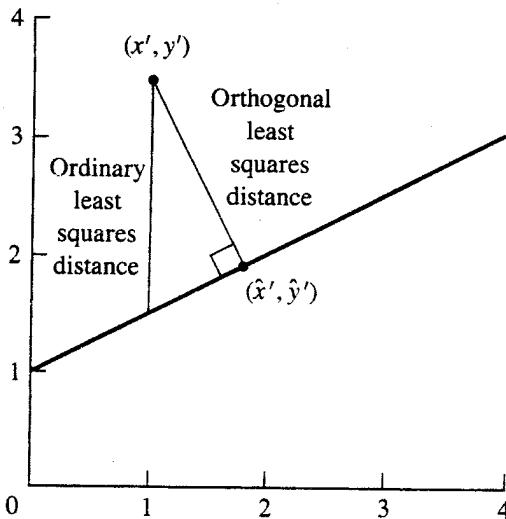
As already mentioned, one of the major differences in the models is in the inferences about  $\alpha$  and  $\beta$ , the parameters that describe the regression relationship. This difference is of utmost importance and cannot be stressed too often. In the functional relationship model, this inference is made conditional on  $\xi_1, \dots, \xi_n$ , using the joint distribution of  $X$  and  $Y$  conditional on  $\xi_1, \dots, \xi_n$ . On the other hand, in the structural relationship model, this inference is made unconditional on  $\xi_1, \dots, \xi_n$ , using the marginal distribution of  $X$  and  $Y$  with  $\xi_1, \dots, \xi_n$  integrated out.

### 12.3.2 A Least Squares Solution

As in Section 12.2.1, we forget statistics for a while and try to find the “best” line through the observed points  $(x_i, y_i), i = 1, \dots, n$ . Previously, when it was assumed that the  $x$ s were measured without error, it made sense to consider minimization of vertical distances. This distance measure implicitly assumes that the  $x$  value is correct and results in *ordinary least squares*. Here, however, there is no reason to consider vertical distances since the  $x$ s now have error associated with them. In fact, statistically speaking, ordinary least squares has some problems in EIV models (see the *Miscellanea* section).

One way to take account of the fact that the  $x$ s also have error in their measurement is to perform *orthogonal least squares*, that is, to find the line that minimizes orthogonal (perpendicular to the line) distances, rather than vertical distances (see Figure 12.3.1). This distance measure does not favor the  $x$  variable, as does ordinary least squares, but rather treats both variables equitably. It is also known as the method of *total least squares*.

Referring to Figure 12.3.1, for a particular data point  $(x', y')$ , the point on a line  $y = a + bx$  that is closest when we measure distance orthogonally is given by



**FIGURE 12.3.1** Distance minimized by orthogonal least squares

(Exercise 12.19)

$$(12.3.7) \quad \hat{x}' = \frac{by' + x' - ab}{1 + b^2}, \quad \hat{y}' = a + \frac{b}{1 + b^2}(by' + x' - ab).$$

Now assume that we have data  $(x_i, y_i), i = 1, \dots, n$ . The squared distance between an observed point  $(x_i, y_i)$  and the closest point on the line  $y = a + bx$  is  $(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2$ , where  $\hat{x}_i$  and  $\hat{y}_i$  are defined by (12.3.7). The *total least squares problem* is to minimize, over all  $a$  and  $b$ , the quantity

$$\sum_{i=1}^n ((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2).$$

It is straightforward to establish that we have

$$\begin{aligned}
 & \sum_{i=1}^n ((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2) \\
 &= \sum_{i=1}^n \left( \frac{b^2}{(1+b^2)^2} [y_i - (a + bx_i)]^2 + \frac{1}{(1+b^2)^2} [y_i - (a + bx_i)]^2 \right) \\
 (12.3.8) \quad &= \frac{1}{1+b^2} \sum_{i=1}^n (y_i - (a + bx_i))^2.
 \end{aligned}$$

For fixed  $b$ , the term in front of the sum is a constant. Thus, the minimizing choice of  $a$  in the sum is  $a = \bar{y} - b\bar{x}$ , just as in (12.2.4). Substituting back into (12.3.8), the total least squares solution is the one that minimizes, over all  $b$ ,

$$(12.3.9) \quad \frac{1}{1+b^2} \sum_{i=1}^n ((y_i - \bar{y}) - b(x_i - \bar{x}))^2.$$

As in (12.2.2), we define the sums of squares and cross-products by

$$(12.3.10) \quad S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Expanding the square and summing shows that (12.3.9) becomes

$$\frac{1}{1+b^2} [S_{yy} - 2bS_{xy} + b^2 S_{xx}].$$

Standard calculus methods will give the minimum (Exercise 12.20), and we find the orthogonal least squares line given by  $y = a + bx$ , with

$$(12.3.11) \quad a = \bar{y} - b\bar{x}, \quad b = \frac{-(S_{xx} - S_{yy}) + \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}}{2S_{xy}}.$$

As might be expected, this line is different from the least squares line. In fact, as we shall see, this line always lies between the ordinary regression of  $y$  on  $x$  and the ordinary regression of  $x$  on  $y$ . This is illustrated in Figure 12.3.2 where the data in Table 12.2.1 were used to calculate the orthogonal least squares line  $\hat{y} = -.49 + 1.88x$ .

In simple linear regression we saw that, under normality, the ordinary least squares solutions for  $\alpha$  and  $\beta$  were the same as the MLEs. Here, the orthogonal least squares solution is the MLE only in a special case, when we make certain assumptions about the parameters.

The difficulties to be encountered with likelihood estimation once again illustrate the differences between a mathematical solution and a statistical solution. We obtained a mathematical least squares solution to the line fitting problem without much difficulty. This will not happen with the likelihood solution.

### 12.3.3 Maximum Likelihood Estimation

We first consider the maximum likelihood solution of the functional linear relationship model, the situation for the structural relationship model being similar and, in some respects, easier. Using the normality assumption, the functional relationship model can be expressed as

$$Y_i \sim n(\alpha + \beta\xi_i, \sigma_e^2) \quad \text{and} \quad X_i \sim n(\xi_i, \sigma_\delta^2), \quad i = 1, \dots, n,$$

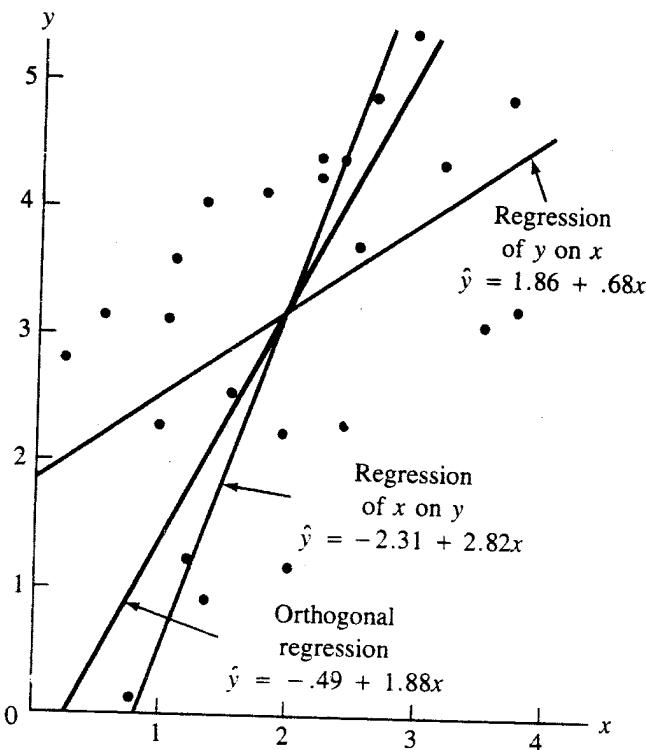


FIGURE 12.3.2 Three regression lines for data in Table 12.2.1

where the  $X_i$ s and  $Y_i$ s are independent. Given observations  $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \dots, (x_n, y_n))$ , the likelihood function is

$$L(\alpha, \beta, \xi_1, \dots, \xi_n, \sigma_\delta^2, \sigma_\epsilon^2 | (\mathbf{x}, \mathbf{y}))$$

(12.3.12)

$$= \frac{1}{(2\pi)^n} \frac{1}{(\sigma_\delta^2 \sigma_\epsilon^2)^{n/2}} \exp \left[ - \sum_{i=1}^n \frac{(x_i - \xi_i)^2}{2\sigma_\delta^2} \right] \exp \left[ - \sum_{i=1}^n \frac{(y_i - (\alpha + \beta\xi_i))^2}{2\sigma_\epsilon^2} \right].$$

The problem with this likelihood function is that it does not have a finite maximum. To see this, take the parameter configuration  $\xi_i = x_i$  and then let  $\sigma_\delta^2 \rightarrow 0$ . The value of the function goes to infinity, showing that there is no maximum likelihood solution. In fact, Solari (1969) has shown that if the equations defining the first derivative of  $L$  are set equal to zero and solved, the result is a saddle point, not a maximum. Notice that as long as we have total control over the parameters, we can always force the likelihood function to infinity. In particular, we can always take a variance to zero, while keeping the exponential term bounded.

We will make the common assumption, which is not only reasonable but also alleviates many problems, that  $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$ , where  $\lambda > 0$  is fixed and known. (See Kendall and Stuart (1979) for a discussion of other assumptions on the variances.) This assumption is one of the least restrictive, saying that we know only the ratio of the variances, not the individual values. Moreover, the resulting model is relatively well behaved.

Under this assumption, we can write the likelihood function as

$$(12.3.13) \quad L(\alpha, \beta, \xi_1, \dots, \xi_n, \sigma_\delta^2 | (\mathbf{x}, \mathbf{y})) = \frac{1}{(2\pi)^n} \frac{\lambda^{n/2}}{(\sigma_\delta^2)^n} \exp \left[ - \sum_{i=1}^n \frac{(x_i - \xi_i)^2 + \lambda(y_i - (\alpha + \beta\xi_i))^2}{2\sigma_\delta^2} \right],$$

which we can now maximize. We will perform the maximization in stages, making sure that, at each step, we have a maximum before proceeding to the next step. By examining the function (12.3.13), we can determine a reasonable order of maximization.

First, for each value of  $\alpha$ ,  $\beta$ , and  $\sigma_\delta^2$ , to maximize  $L$  with respect to  $\xi_1, \dots, \xi_n$  we minimize  $\sum_{i=1}^n ((x_i - \xi_i)^2 + \lambda(y_i - (\alpha + \beta\xi_i))^2)$ . (See Exercise 12.21 for details.) For each  $i$ , we have a quadratic in  $\xi_i$  and the minimum is attained at

$$\xi_i^* = \frac{x_i + \lambda\beta(y_i - \alpha)}{1 + \lambda\beta^2}.$$

On substituting back we get

$$\sum_{i=1}^n ((x_i - \xi_i^*)^2 + \lambda(y_i - (\alpha + \beta\xi_i^*))^2) = \frac{\lambda}{1 + \lambda\beta^2} \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2.$$

The likelihood function now becomes

$$(12.3.14) \quad \max_{\xi_1, \dots, \xi_n} L(\alpha, \beta, \xi_1, \dots, \xi_n, \sigma_\delta^2 | (\mathbf{x}, \mathbf{y})) = \frac{1}{(2\pi)^n} \frac{\lambda^{n/2}}{(\sigma_\delta^2)^n} \exp \left\{ - \frac{1}{2\sigma_\delta^2} \left[ \frac{\lambda}{1 + \lambda\beta^2} \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2 \right] \right\}.$$

Now, we can maximize with respect to  $\alpha$  and  $\beta$ , but a little work will show that we have already done this in the orthogonal least squares solution! Yes, there is somewhat of a correspondence between orthogonal least squares and maximum likelihood in the EIV model and we are about to exploit it. Define

$$(12.3.15) \quad \alpha^* = \sqrt{\lambda}\alpha, \quad \beta^* = \sqrt{\lambda}\beta, \quad y_i^* = \sqrt{\lambda}y_i, \quad i = 1, \dots, n.$$

The exponent of (12.3.14) becomes

$$\frac{\lambda}{1 + \lambda\beta^2} \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2 = \frac{1}{1 + \beta^{*2}} \sum_{i=1}^n (y_i^* - (\alpha^* + \beta^* x_i))^2,$$

which is identical to the expression in the orthogonal least squares problem. From (12.3.11) we know the minimizing values of  $\alpha^*$  and  $\beta^*$  and using (12.3.15) we obtain our MLEs for the slope and intercept

$$(12.3.16) \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad \text{and} \quad \hat{\beta} = \frac{-(S_{xx} - \lambda S_{yy}) + \sqrt{(S_{xx} - \lambda S_{yy})^2 + 4\lambda S_{xy}^2}}{2\lambda S_{xy}}.$$

It is clear from the formula that, at  $\lambda = 1$ , the MLEs agree with the orthogonal least squares solutions. This makes sense. The orthogonal least squares solution treated  $x$  and  $y$  as having the same magnitude of error and this translates into a variance ratio of 1. Carrying this argument further, we can relate this solution to ordinary least squares or maximum likelihood when the  $x$ s are assumed to be fixed. If the  $x$ s are fixed, their variance is zero and hence  $\lambda = 0$ . The maximum likelihood solution for general  $\lambda$  does reduce to ordinary least squares in this case. This relationship, among others, is explored in Exercise 12.22.

Putting (12.3.16) together with (12.3.14), we now have almost completely maximized the likelihood. We have

$$(12.3.17) \quad \begin{aligned} \max_{\alpha, \beta, \xi_1, \dots, \xi_n} L(\alpha, \beta, \xi_1, \dots, \xi_n, \sigma_\delta^2 | (x, y)) \\ = \frac{1}{(2\pi)^n (\sigma_\delta^2)^n} \exp \left[ -\frac{1}{2\sigma_\delta^2} \frac{\lambda}{1 + \lambda\hat{\beta}^2} \sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x_i))^2 \right]. \end{aligned}$$

Now maximizing  $L$  with respect to  $\sigma_\delta^2$  is very similar to finding the MLE of  $\sigma^2$  in ordinary normal sampling (Example 7.2.8), the major difference being the exponent of  $n$ , rather than  $n/2$ , on  $\sigma_\delta^2$ . The details are left to Exercise 12.23. The resulting MLE for  $\sigma_\delta^2$  is

$$(12.3.18) \quad \hat{\sigma}_\delta^2 = \frac{1}{2n} \frac{\lambda}{1 + \lambda\hat{\beta}^2} \sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x_i))^2.$$

From the properties of MLEs, it follows that the MLE of  $\sigma_\epsilon^2$  is given by  $\hat{\sigma}_\epsilon^2 = \hat{\sigma}_\delta^2/\lambda$  and  $\hat{\xi}_i = \hat{\alpha} + \hat{\beta}x_i$ . Although the  $\hat{\xi}_i$ s are not usually of interest, they can sometimes be useful if prediction is desired. Also, the  $\hat{\xi}_i$ s are useful in examining the adequacy of the fit (see Fuller (1987)).

It is interesting to note that although  $\hat{\alpha}$  and  $\hat{\beta}$  are consistent estimators,  $\sigma_\delta^2$  is not. More precisely, as  $n \rightarrow \infty$ ,

$$\hat{\alpha} \rightarrow \alpha \text{ in probability,}$$

$$\hat{\beta} \rightarrow \beta \text{ in probability,}$$

but

$$\hat{\sigma}_\delta^2 \rightarrow \frac{1}{2}\sigma_\delta^2 \text{ in probability.}$$

General results on consistency in EIV functional relationship models have been obtained by Gleser (1981).

We now turn to the linear structural relationship model. Recall that here we assume that we observe pairs  $(X_i, Y_i), i = 1, \dots, n$ , according to

$$\begin{aligned} Y_i &\sim n(\alpha + \beta\xi_i, \sigma_\epsilon^2), \\ X_i &\sim n(\xi_i, \sigma_\delta^2), \\ \xi_i &\sim n(\xi, \sigma_\xi^2), \end{aligned}$$

where the  $\xi_i$ s are independent and, given the  $\xi_i$ s, the  $X_i$ s and  $Y_i$ s are independent. As mentioned before, inference about  $\alpha$  and  $\beta$  will be made from the marginal distribution of  $X_i$  and  $Y_i$ , that is, the distribution obtained by integrating out  $\xi_i$ . If we integrate out  $\xi_i$ , we obtain the marginal distribution of  $(X_i, Y_i)$  (Exercise 12.24),

$$(12.3.19) \quad (X_i, Y_i) \sim \text{bivariate normal}(\xi, \alpha + \beta\xi, \sigma_\delta^2 + \sigma_\xi^2, \sigma_\epsilon^2 + \beta^2\sigma_\xi^2, \beta\sigma_\xi^2).$$

Notice the similarity of the correlation structure to that of the RCB ANOVA in Section 11.3.1. There, conditional on blocks, the observations were uncorrelated but, unconditionally, there was correlation (the intraclass correlation). Here, the functional relationship model, which is conditional on the  $\xi_i$ s, has uncorrelated observations, but the structural relationship model, where we infer unconditional on the  $\xi_i$ s, has correlated observations. The  $\xi_i$ s are playing a role similar to blocks and the correlation that appears here is similar to the intraclass correlation. (In fact, it is identical to the intraclass correlation if  $\beta = 1$  and  $\sigma_\delta^2 = \sigma_\epsilon^2$ .)

To proceed with likelihood estimation in this case, given observations  $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \dots, (x_n, y_n))$ , the likelihood function is that of a bivariate normal, as was encountered in Exercise 7.16. There, it was seen that the likelihood estimators in the bivariate normal could be found by equating sample quantities to population quantities. Hence, to find the MLEs of  $\alpha, \beta, \xi, \sigma_\epsilon^2, \sigma_\delta^2$ , and  $\sigma_\xi^2$ , we solve

$$\begin{aligned} \bar{y} &= \hat{\alpha} + \hat{\beta}\hat{\xi}, \\ \bar{x} &= \hat{\xi}, \\ (12.3.20) \quad \frac{1}{n}S_{yy} &= \hat{\sigma}_\epsilon^2 + \hat{\beta}^2\hat{\sigma}_\xi^2, \\ \frac{1}{n}S_{xx} &= \hat{\sigma}_\delta^2 + \hat{\sigma}_\xi^2, \\ \frac{1}{n}S_{xy} &= \hat{\beta}\hat{\sigma}_\xi^2. \end{aligned}$$

Note that we have five equations, but there are six unknowns, so the system is indeterminate. That is, the system of equations does not have a unique solution and there is no unique value of the parameter vector  $(\alpha, \beta, \xi, \sigma_\epsilon^2, \sigma_\delta^2, \sigma_\xi^2)$  that maximizes the likelihood.

Before we go on, realize that the variances of  $X_i$  and  $Y_i$  here are different from the variances in the functional relationship model. There, we were working conditional on  $\xi_1, \dots, \xi_n$  and here we are working marginally with respect to the  $\xi_i$ s. So, for example, in the functional relationship model we write  $\text{Var } X_i = \sigma_\delta^2$  (where it is understood that this variance is conditional on  $\xi_1, \dots, \xi_n$ ) while in the structural model we write  $\text{Var } X_i = \sigma_\delta^2 + \sigma_\xi^2$  (where it is understood that this variance is unconditional on  $\xi_1, \dots, \xi_n$ ). This should not be a source of confusion.

A solution to the equations in (12.3.20) imply a restriction on  $\hat{\beta}$ , a restriction that we have already encountered in the functional relationship case (Exercise 12.22). From the above equations involving the variances and covariance, it is straightforward to deduce that

$$\begin{aligned}\hat{\sigma}_\delta^2 &\geq 0 \quad \text{only if } S_{xx} \geq \frac{1}{\hat{\beta}} S_{xy}, \\ \hat{\sigma}_\epsilon^2 &\geq 0 \quad \text{only if } S_{yy} \geq \hat{\beta} S_{xy},\end{aligned}$$

which together imply that

$$\frac{|S_{xy}|}{S_{xx}} \leq |\hat{\beta}| \leq \frac{S_{yy}}{|S_{xy}|}.$$

(The bounds on  $\hat{\beta}$  are established in Exercise 12.27.)

We now address the identifiability problem in the structural relationship case, a problem that can be expected since, in (12.3.19) we have more parameters than are needed to specify the distribution. To make the structural linear relationship model identifiable, we must make an assumption that reduces the number of parameters to five. It fortunately happens that the assumption about variances made for the functional relationship solves the identifiability problem here. Thus, we assume that  $\sigma_\delta^2 = \lambda \sigma_\epsilon^2$ , where  $\lambda$  is known. This reduces the number of unknown parameters to five and makes the model identifiable. (See Exercise 12.26.) More restrictive assumptions, such as assuming that  $\sigma_\delta^2$  is known, may lead to MLEs of variances that have the value zero. Kendall and Stuart (1979) have a full discussion of this.

Once we assume that  $\sigma_\delta^2 = \lambda \sigma_\epsilon^2$ , the maximum likelihood estimates for  $\hat{\alpha}$  and  $\hat{\beta}$  in this model are the same as in the functional relationship model and are given by (12.3.16). The variance estimates are different, however, and are given by

$$\begin{aligned}\hat{\sigma}_\delta^2 &= \frac{1}{n} \left( S_{xx} - \frac{S_{xy}}{\hat{\beta}} \right), \\ (12.3.21) \quad \hat{\sigma}_\epsilon^2 &= \frac{\hat{\sigma}_\delta^2}{\lambda} = \frac{1}{n} (S_{yy} - \hat{\beta} S_{xy}),\end{aligned}$$

$$\hat{\sigma}_\xi^2 = \frac{1}{n} \frac{S_{xy}}{\hat{\beta}}.$$

(Exercise 12.28 shows this and also explores the relationship between variance estimates here and in the functional model.) Note that, in contrast to what happened in the functional relationship model, these estimators are all consistent in the linear structural relationship model (when  $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$ ).

### 12.3.4 Confidence Sets

As might be expected, the construction of confidence sets in the EIV model is a difficult task. A complete treatment of the subject needs machinery that we have not developed. In particular, we will concentrate here only on confidence sets for the slope,  $\beta$ .

As a first attack, we could use the approximate likelihood method of Section 9.4.1 to construct approximate confidence intervals. In practice this is probably what is most often done and is not totally unreasonable. However, these approximate intervals cannot maintain a nominal  $1 - \alpha$  confidence level. In fact, results of Gleser and Hwang (1987) yield the rather unsettling result that any interval estimator of the slope whose length is *always finite* will have confidence coefficient equal to zero!

For definiteness, in the remainder of this section we will assume that we are in the structural relationship case of the EIV model. The confidence set results presented are valid in both the structural and functional cases and, in particular, the formulas remain the same. We continue to assume that  $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$ , where  $\lambda$  is known.

Gleser and Hwang (1987) identify the parameter

$$\tau^2 = \frac{\sigma_\xi^2}{\sigma_\delta^2}$$

as determining the amount of information potentially available in the data to determine the slope  $\beta$ . They show that, as  $\tau^2 \rightarrow 0$ , the coverage probability of any finite-length confidence interval on  $\beta$  must also go to 0. To see why this is plausible, note that  $\tau^2 = 0$  implies that the  $\xi_i$ s do not vary and it would be impossible to fit a unique straight line.

An approximate confidence interval for  $\beta$  can be constructed by using the fact that the estimator

$$\hat{\sigma}_\beta^2 = \frac{(1 + \lambda\hat{\beta}^2)^2(S_{xx}S_{yy} - S_{xy}^2)}{(S_{xx} - \lambda S_{yy})^2 + 4\lambda S_{xy}^2}$$

is a consistent estimator of  $\sigma_\beta^2$ , the true variance of  $\hat{\beta}$ . Hence, using the CLT together with Slutsky's Theorem (Section 5.3.3), it can be shown that the interval

$$\hat{\beta} - \frac{z_{\alpha/2}\hat{\sigma}_\beta}{\sqrt{n}} \leq \beta \leq \hat{\beta} + \frac{z_{\alpha/2}\hat{\sigma}_\beta}{\sqrt{n}}$$

is an approximate  $1 - \alpha$  confidence interval for  $\beta$ . However, since it has finite length, it cannot maintain  $1 - \alpha$  coverage for all parameter values.

Gleser (1987) considers a modification of this interval and reports the infimum of its coverage probabilities as a function of  $\tau^2$ . Gleser's modification,  $C_G(\hat{\beta})$ , is

$$(12.3.22) \quad \hat{\beta} - \frac{t_{n-2,\alpha/2}\hat{\sigma}_\beta}{\sqrt{n-2}} \leq \beta \leq \hat{\beta} + \frac{t_{n-2,\alpha/2}\hat{\sigma}_\beta}{\sqrt{n-2}}.$$

Again using the CLT together with Slutsky's Theorem, it can be shown that this is an approximate  $1 - \alpha$  confidence interval for  $\beta$ . Since this interval also has finite length, it also cannot maintain  $1 - \alpha$  coverage for all parameter values. Gleser does some finite-sample numerical calculations and gives bounds on the infima of the coverage probabilities as a function of  $\tau^2$ . For reasonable values of  $n$  ( $\geq 10$ ), the coverage probability of a nominal 90% interval will be at least 80% if  $\tau^2 \geq .25$ . As  $\tau^2$  or  $n$  increases, this performance improves.

In contrast to  $C_G(\hat{\beta})$  of (12.3.22), which has finite length but no guaranteed coverage probability, we now look at an exact confidence set that, as it must, has infinite length. The set, known as the Creasy–Williams confidence set, is due to Creasy (1956) and Williams (1959) and is based on the fact (see Exercise 12.29) that if  $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$ , then

$$\text{Cov}(\beta\lambda Y_i + X_i, Y_i - \beta X_i) = 0.$$

Define  $r_\lambda(\beta)$  to be the sample correlation coefficient between  $\beta\lambda Y_i + X_i$  and  $Y_i - \beta X_i$ , that is,

$$(12.3.23) \quad r_\lambda(\beta) = \frac{\sum_{i=1}^n ((\beta\lambda y_i + x_i) - (\beta\lambda\bar{y} + \bar{x}))((y_i - \beta x_i) - (\bar{y} - \beta\bar{x}))}{\sqrt{\sum_{i=1}^n ((\beta\lambda y_i + x_i) - (\beta\lambda\bar{y} + \bar{x}))^2 \sum_{i=1}^n ((y_i - \beta x_i) - (\bar{y} - \beta\bar{x}))^2}} \\ = \frac{\beta\lambda S_{yy} + (1 - \beta^2\lambda)S_{xy} - \beta S_{xx}}{\sqrt{(\beta^2\lambda^2 S_{yy} + 2\beta\lambda S_{xy} + S_{xx})(S_{yy} - 2\beta S_{xy} + \beta^2 S_{xx})}}.$$

Since  $\beta\lambda Y_i + X_i$  and  $Y_i - \beta X_i$  are bivariate normal with correlation zero, it follows (Exercise 12.10) that

$$\frac{\sqrt{n-2}r_\lambda(\beta)}{\sqrt{1-r_\lambda^2(\beta)}} \sim t_{n-2},$$

for any value of  $\beta$ . Thus, we have identified a pivotal quantity and we conclude that the set

$$(12.3.24) \quad \left\{ \beta : \frac{(n-2)r_\lambda^2(\beta)}{1-r_\lambda^2(\beta)} \leq F_{1,n-2,\alpha} \right\}$$

is a  $1 - \alpha$  confidence set for  $\beta$  (see Exercise 12.29).

Although this confidence set is a  $1 - \alpha$  set, it suffers from defects similar to those of Fieller's intervals. The function describing the set (12.3.24) has two minima, where the function is zero. The confidence set can consist of two finite disjoint intervals, one finite and two infinite disjoint intervals, or the whole real line. For example, the graph of the  $F$  statistic function for the data in Table 12.2.1 with  $\lambda = 1$  is in Figure 12.3.3. The confidence set is all the  $\beta$ s where the function is less than or equal to  $F_{1,22,\alpha}$ . For  $\alpha = .05$  and  $F_{1,22,.05} = 4.30$ , the confidence set is  $[-1.13, -.14] \cup [.89, 7.38]$ . For  $\alpha = .01$  and  $F_{1,22,.01} = 7.95$ , the confidence set is  $(-\infty, -18.18] \cup [-1.68, .06] \cup [.60, \infty)$ .

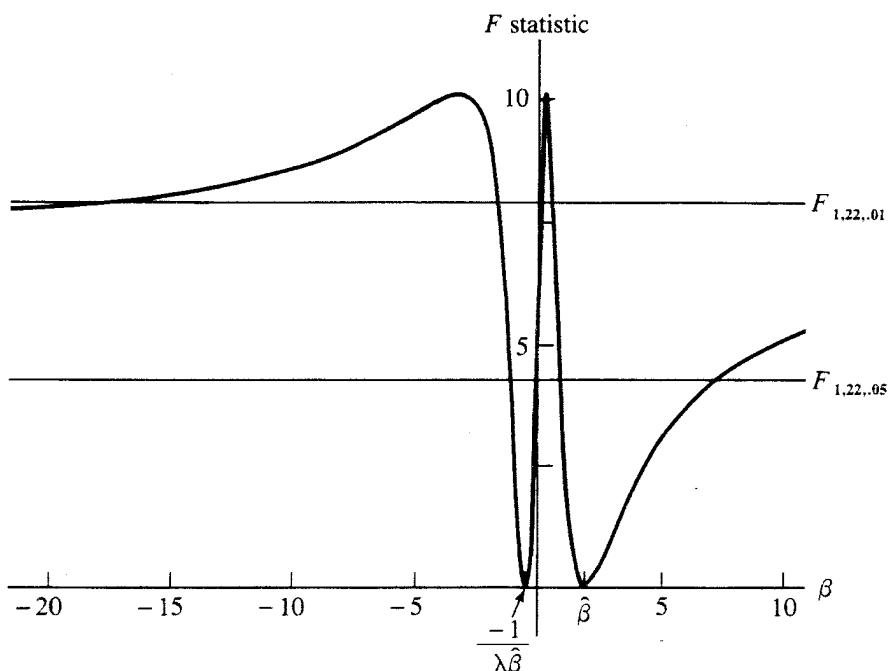


FIGURE 12.3.3  $F$  statistic defining Creasy–Williams confidence set

Furthermore, for every value of  $\beta$ ,  $-r_\lambda(\beta) = r_\lambda(-1/(\lambda\beta))$  (see Exercise 12.30) so that if  $\beta$  is in the confidence set, so is  $-1/(\lambda\beta)$ . Using this confidence set, we cannot distinguish  $\beta$  from  $-1/(\lambda\beta)$  and this confidence set always contains both positive and negative values. We can never determine the sign of the slope from this confidence set!

The confidence set given in (12.3.24) is not exactly the one discussed by Creasy (1956) but rather a modification. She was actually interested in estimating  $\phi$ , the angle that  $\beta$  makes with the  $x$ -axis, that is,  $\beta = \tan(\phi)$ , and confidence sets there have fewer problems. Estimation of  $\phi$  is perhaps more natural in EIV models (see, for example, Anderson (1976)) but we seem to be more inclined to estimate  $\alpha$  and  $\beta$ .

Most of the other standard statistical analyses that can be done in the ordinary linear regression case have analogues in EIV models. For example, we can test hypotheses about  $\beta$  or estimate values of  $EY_i$ . More about these topics can be found in Fuller (1987) or Kendall and Stuart (1979).

**EXERCISES**

- 12.1** In Section 12.2.1, we found the least squares estimators of  $\alpha$  and  $\beta$  by a two-stage minimization. This minimization can also be done using partial derivatives.

a. Compute  $\frac{\partial \text{RSS}}{\partial c}$  and  $\frac{\partial \text{RSS}}{\partial d}$  and set them equal to zero. Show that the resulting two equations can be written as

$$nc + \left( \sum_{i=1}^n x_i \right) d = \sum_{i=1}^n y_i$$

and

$$\left( \sum_{i=1}^n x_i \right) c + \left( \sum_{i=1}^n x_i^2 \right) d = \sum_{i=1}^n x_i y_i.$$

(These equations are called the *normal equations* for this minimization problem.)

- b. Show that  $c = a$  and  $d = b$  are the solutions to the normal equations.  
 c. Check the second partial derivative condition to verify that the point  $c = a$  and  $d = b$  is indeed the minimum of RSS.
- 12.2** Suppose  $n$  is an even number. The values of the predictor variable,  $x_1, \dots, x_n$ , all must be chosen to be in the interval  $[e, f]$ . Show that the choice that maximizes  $S_{xx}$  is to choose half of the  $x_i$  equal to  $e$  and the other half equal to  $f$ . (This was the choice mentioned in Section 12.2.2 that minimizes  $\text{Var } b$ .)
- 12.3** There are other reasonable measures, besides RSS, for measuring the fit of a line to a set of data. The *least absolute deviation line* is given by the values of  $c$  and  $d$  that minimize

$$\sum_{i=1}^n |y_i - (c + dx_i)|.$$

The least absolute deviation line is harder to compute than the least squares line and has some undesirable properties. In particular, the least absolute deviation line is not always uniquely defined.

- a. Show that, for a data set with three observations,  $(x_1, y_1)$ ,  $(x_1, y_2)$ , and  $(x_3, y_3)$  (note the first two  $x$ s are the same), any line that goes through  $(x_3, y_3)$  and lies between  $(x_1, y_1)$  and  $(x_1, y_2)$  is a least absolute deviation line.  
 b. For three individuals, measurements are taken on heart rate ( $x$ , in beats per minute) and oxygen consumption ( $y$ , in ml/kg). The  $(x, y)$  pairs are  $(127, 14.4)$ ,  $(127, 11.9)$ , and  $(136, 17.9)$ . Calculate the slope and intercept of the least squares line and the range of the least absolute deviation lines.
- 12.4** Observations  $(x_i, Y_i)$ ,  $i = 1, \dots, n$ , are collected according to the model

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

where  $E\epsilon_i = 0$ ,  $\text{Var } \epsilon_i = \sigma^2$ , and  $\text{Cov}(\epsilon_i, \epsilon_j) = 0$  if  $i \neq j$ . Find the best linear, unbiased estimator of  $\alpha$ .

- 12.5** Show that in the conditional normal model for simple linear regression, the MLE of  $\sigma^2$  is given by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

**12.6** Consider the residuals  $\hat{\epsilon}_1, \dots, \hat{\epsilon}_n$  defined in Section 12.2.4 by  $\hat{\epsilon}_i = Y_i - \hat{\alpha} - \hat{\beta}x_i$ .

- a. Show that  $E\hat{\epsilon}_i = 0$ .
- b. Verify that

$$\text{Var } \hat{\epsilon}_i = \text{Var } Y_i + \text{Var } \hat{\alpha} + x_i^2 \text{Var } \hat{\beta} - 2\text{Cov}(Y_i, \hat{\alpha}) - 2x_i \text{Cov}(Y_i, \hat{\beta}) + 2x_i \text{Cov}(\hat{\alpha}, \hat{\beta}).$$

- c. Use Lemma 12.2.1 to show that

$$\text{Cov}(Y_i, \hat{\alpha}) = \sigma^2 \left( \frac{1}{n} + \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right) \quad \text{and} \quad \text{Cov}(Y_i, \hat{\beta}) = \sigma^2 \frac{x_i - \bar{x}}{S_{xx}},$$

and use these to verify (12.2.23).

**12.7** Fill in the details about the distribution of  $\hat{\alpha}$  left out of the proof of Theorem 12.2.1.

- a. Show that the estimator  $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$  can be expressed as  $\hat{\alpha} = \sum_{i=1}^n c_i Y_i$ , where

$$c_i = \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}.$$

- b. Verify that

$$E\hat{\alpha} = \alpha \quad \text{and} \quad \text{Var } \hat{\alpha} = \sigma^2 \left[ \frac{1}{n S_{xx}} \sum_{i=1}^n x_i^2 \right].$$

- c. Verify that

$$\text{Cov}(\hat{\alpha}, \hat{\beta}) = \frac{-\sigma^2 \bar{x}}{S_{xx}}.$$

**12.8** Verify the claim in Theorem 12.2.1, that  $\hat{\epsilon}_i$  is uncorrelated with  $\hat{\alpha}$  and  $\hat{\beta}$ . (Show that  $\hat{\epsilon}_i = \sum e_j Y_j$ , where the  $e_j$ s are given by (12.2.25). Then, using the facts that we can write  $\hat{\alpha} = \sum c_j Y_j$  and  $\hat{\beta} = \sum d_j Y_j$ , verify that  $\sum e_j c_j = \sum e_j d_j = 0$  and apply Lemma 12.2.1.)

**12.9** Observations  $(x_i, Y_i), i = 1, \dots, n$ , are made according to the model

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

where  $x_1, \dots, x_n$  are fixed constants and  $\epsilon_1, \dots, \epsilon_n$  are iid  $n(0, \sigma^2)$ . The model is then reparameterized as

$$Y_i = \alpha' + \beta'(x_i - \bar{x}) + \epsilon_i.$$

Let  $\hat{\alpha}$  and  $\hat{\beta}$  denote the MLEs of  $\alpha$  and  $\beta$ , respectively, and  $\hat{\alpha}'$  and  $\hat{\beta}'$  denote the MLEs of  $\alpha'$  and  $\beta'$ , respectively.

a. Show that  $\hat{\beta}' = \hat{\beta}$ .

b. Show that  $\hat{\alpha}' \neq \hat{\alpha}$ . In fact, show that  $\hat{\alpha}' = \bar{Y}$ . Find the distribution of  $\hat{\alpha}'$ .

c. Show that  $\hat{\alpha}'$  and  $\hat{\beta}'$  are uncorrelated and, hence, independent under normality.

- 12.10** Observations  $(X_i, Y_i), i = 1, \dots, n$ , are made from a bivariate normal population with parameters  $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ , and the model

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

is going to be fit.

- a. Argue that the hypothesis  $H_0: \beta = 0$  is true if and only if the hypothesis  $H_0: \rho = 0$  is true. (See (12.2.20).)
- b. Show algebraically that

$$\frac{\hat{\beta}}{S/\sqrt{S_{xx}}} = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}},$$

where  $r$  is the sample correlation coefficient, the MLE of  $\rho$ .

c. Show how to test  $H_0: \rho = 0$ , given only  $r^2$  and  $n$ , using Student's  $t$  with  $n - 2$  degrees of freedom (see (12.2.28)). (Fisher derived an approximate confidence interval for  $\rho$ , using a variance-stabilizing transformation. See Stuart and Ord (1987).)

- 12.11** a. Illustrate the partitioning of the sum of squares for simple linear regression by calculating the regression ANOVA table for the following data. Parents are often interested in predicting the eventual heights of their children. The following is a portion of the data taken from a study that might have been suggested by Galton's analysis.

Height (inches) at age two ( $x$ )	Height (inches) as an adult ( $y$ )
39	71
30	63
32	63
34	67
35	68
36	68
36	70
30	64

- b. Analytically establish the partitioning of the sum of squares for simple linear regression by verifying (12.2.31).
- c. Prove that the two expressions for the regression sum of squares are, in fact, equal. That is, show that

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \frac{S_{xy}^2}{S_{xx}}.$$

d. Show that the *coefficient of determination*,  $r^2$ , given by

$$r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

can be derived as either the square of the sample correlation coefficient of the  $n$  pairs  $(y_1, x_1), \dots, (y_n, x_n)$  or of the  $n$  pairs  $(y_1, \hat{y}_1), \dots, (y_n, \hat{y}_n)$ .

**12.12** We obtain observations  $Y_1, \dots, Y_n$  which can be described by the relationship

$$Y_i = \theta x_i^2 + \epsilon_i,$$

where  $x_1, \dots, x_n$  are fixed constants and  $\epsilon_1, \dots, \epsilon_n$  are iid  $n(0, \sigma^2)$ .

- a. Find the least squares estimator of  $\theta$ .
- b. Find the MLE of  $\theta$ .
- c. Find the best unbiased estimator of  $\theta$ .

**12.13** Observations  $Y_1, \dots, Y_n$  are made according to the model

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

where  $x_1, \dots, x_n$  are fixed constants and  $\epsilon_1, \dots, \epsilon_n$  are iid  $n(0, \sigma^2)$ . Let  $\hat{\alpha}$  and  $\hat{\beta}$  denote MLEs of  $\alpha$  and  $\beta$ .

- a. Suppose we assume that  $x_1, \dots, x_n$  are observed values of iid random variables  $X_1, \dots, X_n$  with distribution  $n(\mu_X, \sigma_X^2)$ . Prove that when we take expectations over the joint distribution of  $X$  and  $Y$  we still get  $E\hat{\alpha} = \alpha$  and  $E\hat{\beta} = \beta$ .
- b. The phenomenon of part (a) does not carry over to the covariance. Calculate the unconditional covariance of  $\hat{\alpha}$  and  $\hat{\beta}$  (using the joint distribution of  $X$  and  $Y$ ).
- 12.14** We observe random variables  $Y_1, \dots, Y_n$  that are mutually independent, each with a normal distribution with variance  $\sigma^2$ . Furthermore,  $EY_i = \beta x_i$ , where  $\beta$  is an unknown parameter and  $x_1, \dots, x_n$  are fixed constants not all equal to zero.
- a. Find the MLE of  $\beta$ . Compute its mean and variance.
- b. Compute the Cramér–Rao Lower Bound for the variance of an unbiased estimator of  $\beta$ .
- c. Find a best unbiased estimator of  $\beta$ .
- d. If you could place the values  $x_1, \dots, x_n$  anywhere within a given nondegenerate closed interval  $[A, B]$ , where would you place these values? Justify your answer.
- e. For a given positive value  $r$ , the *maximum probability estimator of  $\beta$  with respect to  $r$*  is the value of  $D$  that maximizes the integral

$$\int_{D-r}^{D+r} f(y_1, \dots, y_n | \beta) d\beta,$$

where  $f(y_1, \dots, y_n | \beta)$  is the joint pdf of  $Y_1, \dots, Y_n$ . Find this estimator.

- 12.15** An ecologist takes data  $(x_i, Y_i), i = 1, \dots, n$ , where  $x_i$  is the size of an area and  $Y_i$  is the number of moss plants in the area. We model the data by  $Y_i \sim \text{Poisson}(\theta x_i)$ ,  $Y_i$ s independent.
- a. Show that the least squares estimator of  $\theta$  is  $\sum x_i Y_i / \sum x_i^2$  and has variance  $\theta \sum x_i^3 / (\sum x_i^2)^2$ . Also, compute its bias.
  - b. Show that the MLE of  $\theta$  is  $\sum Y_i / \sum x_i$  and has variance  $\theta / \sum x_i$ . Compute its bias.

- c. Find a best unbiased estimator of  $\theta$  and show that its variance attains the Cramér-Rao Lower Bound.
- 12.16** Usually we think of prediction in a regression setting. But it also makes sense in simple random sampling from a single population. Let  $Y_1, \dots, Y_n$  be a random sample from a  $n(\mu, \sigma^2)$  population. The sample mean and variance from this sample are  $\bar{Y}$  and  $S^2$ . We consider predicting a new, independent observation,  $Y_0$ , from this population with an interval of the form
- $$\bar{Y} - tSb < Y_0 < \bar{Y} + tSb,$$
- where  $t$  is a  $t$  distribution percentile and  $b$  is a constant. What are the values of  $t$  and  $b$  that should be used to produce a  $100(1 - \alpha)\%$  prediction interval?
- 12.17** Verify that the simultaneous confidence intervals in (12.2.37) have the claimed coverage probability.
- 12.18** a. Prove that if  $a$ ,  $b$ ,  $c$ , and  $d$  are constants, with  $c > 0$  and  $d > 0$ , then

$$\max_t \frac{(a + bt)^2}{c + dt^2} = \frac{a^2}{c} + \frac{b^2}{d}.$$

- b. Use part (a) to verify equation (12.2.39) and hence fill in the gap in Theorem 12.2.2.  
 c. Use part (a) to find a Scheffé-type simultaneous band using the prediction intervals of (12.2.36). That is, rewriting the prediction intervals as was done in Theorem 12.2.2, show

$$\max_t \frac{-((\bar{Y} - \mu_{\bar{Y}}) + (\hat{\beta} - \beta)t)^2}{S^2 \left[ 1 + \frac{1}{n} + \frac{t^2}{S_{xx}} \right]} = \frac{\frac{n}{n+1}(\bar{Y} - \mu_{\bar{Y}})^2 + S_{xx}(\hat{\beta} - \beta)^2}{S^2}.$$

- d. The distribution of the maximum is not easy to write down, but we could approximate it. Approximate the statistic by using moment matching, as done in Example 7.2.3.  
**12.19** Verify the expressions in (12.3.7). (*Hint:* Use the Pythagorean Theorem.)  
**12.20** Show that the extrema of

$$f(b) = \frac{1}{1 + b^2} [S_{yy} - 2bS_{xy} + b^2S_{xx}]$$

are given by

$$b = \frac{-(S_{xx} - S_{yy}) \pm \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}}{2S_{xy}}.$$

- Show that the “+” solution gives the minimum of  $f(b)$ .
- 12.21** In maximizing the likelihood (12.3.13), we first minimized, for each value of  $\alpha$ ,  $\beta$ , and  $\sigma_\delta^2$ , the function

$$f(\xi_1, \dots, \xi_n) = \sum_{i=1}^n ((x_i - \xi_i)^2 + \lambda(y_i - (\alpha + \beta\xi_i))^2)$$

with respect to  $\xi_1, \dots, \xi_n$ .

a. Prove that this function is minimized at

$$\xi_i^* = \frac{x_i + \lambda\beta(y_i - \alpha)}{1 + \lambda\beta^2}.$$

b. Show that the function

$$D_\lambda((x, y), (\xi, \alpha + \beta\xi)) = (x - \xi)^2 + \lambda(y - (\alpha + \beta\xi))^2$$

defines a *metric* between the points  $(x, y)$  and  $(\xi, \alpha + \beta\xi)$ . A *metric* is a distance measure, a function  $D$  that measures the distance between two points  $A$  and  $B$ . A metric satisfies the following four properties:

- i.  $D(A, A) = 0$ .
- ii.  $D(A, B) > 0$  if  $A \neq B$ .
- iii.  $D(A, B) = D(B, A)$  (reflexive).
- iv.  $D(A, B) \leq D(A, C) + D(C, B)$  (triangle inequality).

**12.22** Consider the MLE of the slope in the EIV model

$$\hat{\beta}(\lambda) = \frac{-(S_{xx} - \lambda S_{yy}) + \sqrt{(S_{xx} - \lambda S_{yy})^2 + 4\lambda S_{xy}^2}}{2\lambda S_{xy}},$$

where  $\lambda = \sigma_\delta^2/\sigma_\epsilon^2$  is assumed known.

- a. Show that  $\lim_{\lambda \rightarrow 0} \hat{\beta}(\lambda) = S_{xy}/S_{xx}$ , the slope of the ordinary regression of  $y$  on  $x$ .
- b. Show that  $\lim_{\lambda \rightarrow \infty} \hat{\beta}(\lambda) = S_{yy}/S_{xy}$ , the reciprocal of the slope of the ordinary regression of  $x$  on  $y$ .
- c. Show that  $\hat{\beta}(\lambda)$  is, in fact, monotone in  $\lambda$  and is increasing if  $S_{xy} > 0$  and decreasing if  $S_{xy} < 0$ .
- d. Show that the orthogonal least squares line ( $\lambda = 1$ ) is always between the lines given by the ordinary regressions of  $y$  on  $x$  and of  $x$  on  $y$ .
- e. The following data were collected in a study to examine the relationship between brain weight and body weight in a number of animal species.

Species	Body weight (kg) ( $x$ )	Brain weight (g) ( $y$ )
Arctic fox	3.385	44.50
Owl monkey	.480	15.50
Mountain beaver	1.350	8.10
Guinea pig	1.040	5.50
Chinchilla	.425	6.40
Ground squirrel	.101	4.00
Tree hyrax	2.000	12.30
Big brown bat	.023	.30

Calculate the MLE of the slope assuming the EIV model. Also, calculate the least squares slopes of the regressions of  $y$  on  $x$  and of  $x$  on  $y$ , and show how these quantities bound the MLE.

**12.23** In the EIV functional relationship model, where  $\lambda = \sigma_\delta^2/\sigma_\epsilon^2$  is assumed known, show that the MLE of  $\sigma_\delta^2$  is given by (12.3.18).

- 12.24** Show that in the linear structural relationship model (12.3.6), if we integrate out  $\xi_i$ , the marginal distribution of  $(X_i, Y_i)$  is given by (12.3.19).
- 12.25** Consider a linear structural relationship model where we assume that  $\xi_i$  has an improper distribution,  $\xi_i \sim \text{uniform}(-\infty, \infty)$ .
- Show that for each  $i$ ,

$$\begin{aligned} & \int_{-\infty}^{\infty} \frac{1}{(2\pi)} \frac{1}{\sigma_\delta \sigma_\epsilon} \exp \left[ - \left( \frac{(x_i - \xi_i)^2}{2\sigma_\delta^2} \right) \right] \exp \left[ - \left( \frac{(y_i - (\alpha + \beta\xi_i))^2}{2\sigma_\epsilon^2} \right) \right] d\xi_i \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\beta^2 \sigma_\delta^2 + \sigma_\epsilon^2}} \exp \left[ - \frac{1}{2} \frac{(y_i - (\alpha + \beta x_i))^2}{\beta^2 \sigma_\delta^2 + \sigma_\epsilon^2} \right]. \end{aligned}$$

(Completing the square in the exponential makes the integration easy.)

- The result of the integration in part (a) looks like a pdf and, if we consider it a pdf of  $Y$  conditional on  $X$ , then we seem to have a linear relationship between  $X$  and  $Y$ . Thus, it is sometimes said that this “limiting case” of the structural relationship leads to simple linear regression and ordinary least squares. Explain why this interpretation of the above function is wrong.
- 12.26** Verify the nonidentifiability problems in the structural relationship model in the following ways.
- Produce two different sets of parameters that give the same marginal distribution to  $(X_i, Y_i)$ .
  - Show that there are at least two distinct parameter vectors that yield the same solution to the equations given in (12.3.20).
- 12.27** In the structural relationship model, the solution to the equations in (12.3.20) implies a restriction on  $\hat{\beta}$ , the same restriction seen in the functional relationship case (Exercise 12.22).
- Show that in (12.3.20), the MLE of  $\sigma_\delta^2$  is nonnegative only if  $S_{xx} \geq (1/\hat{\beta})S_{xy}$ . Also, the MLE of  $\sigma_\epsilon^2$  is nonnegative only if  $S_{yy} \geq \hat{\beta}S_{xy}$ .
  - Show that the restrictions in part (a), together with the rest of the equations in (12.3.20), imply that

$$\frac{|S_{xy}|}{S_{xx}} \leq |\hat{\beta}| \leq \frac{S_{yy}}{|S_{xy}|}.$$

- 12.28** a. Derive the MLEs for  $(\alpha, \beta, \sigma_\epsilon^2, \sigma_\delta^2, \sigma_\xi^2)$  in the structural relationship model, by solving the equations (12.3.20) under the assumption that  $\sigma_\delta^2 = \lambda \sigma_\epsilon^2$ .
- b. Calculate the MLEs for  $(\alpha, \beta, \sigma_\epsilon^2, \sigma_\delta^2, \sigma_\xi^2)$  for the data of Exercise 12.22 by assuming the structural relationship model holds and that  $\sigma_\delta^2 = \lambda \sigma_\epsilon^2$ .
- c. Verify the relationship between variance estimates in the functional and structural relationship models. In particular, show that

$$\widehat{\text{Var}} X_i(\text{structural}) = 2 \widehat{\text{Var}} X_i(\text{functional}).$$

That is, verify

$$\left( S_{xx} - \frac{S_{xy}}{\hat{\beta}} \right) = \frac{\lambda}{1 + \lambda \hat{\beta}^2} \sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta} x_i))^2.$$

d. Verify the following equality which is implicit in the MLE variance estimates given in (12.3.21). Show that

$$S_{xx} - \frac{S_{xy}}{\hat{\beta}} = \lambda(S_{yy} - \hat{\beta}S_{xy}).$$

**12.29** a. Show that for random variables  $X$  and  $Y$  and constants  $a, b, c, d$ ,

$$\text{Cov}(aY + bX, cY + dX) = ac\text{Var } Y + (bc + ad)\text{Cov}(X, Y) + bd\text{Var } X.$$

b. Use the result in part (a) to verify that in the structural relationship model with  $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$ ,

$$\text{Cov}(\beta\lambda Y_i + X_i, Y_i - \beta X_i) = 0,$$

the identity on which the Creasy–Williams confidence set is based.

c. Use the results of part (b) to show that

$$\frac{\sqrt{(n-2)r_\lambda(\beta)}}{\sqrt{1-r_\lambda^2(\beta)}} \sim t_{n-2},$$

for any value of  $\beta$ , where  $r_\lambda(\beta)$  is given in (12.3.23). Also, show that the confidence set defined in (12.3.24) has constant coverage probability equal to  $1 - \alpha$ .

**12.30** Verify the following facts about  $\hat{\beta}$  (the MLE of  $\beta$  when we assume  $\sigma_\delta^2 = \lambda\sigma_\epsilon^2$ ),  $r_\lambda(\beta)$  of (12.3.23), and  $C_\lambda(\hat{\beta})$ , the Creasy–Williams confidence set of (12.3.24).

- a.  $\hat{\beta}$  and  $-1/(\lambda\hat{\beta})$  are the two roots of the quadratic equation defining the zeros of the first derivative of the likelihood function (12.3.14).
- b.  $r_\lambda(\beta) = -r_\lambda(-1/(\lambda\beta))$  for every  $\beta$ .
- c. If  $\beta \in C_\lambda(\hat{\beta})$ , then  $-1/(\lambda\beta) \in C_\lambda(\hat{\beta})$ .

**12.31** There is an interesting connection between the Creasy–Williams confidence set of (12.3.24) and the interval  $C_G(\hat{\beta})$  of (12.3.22).

a. Show that

$$C_G(\hat{\beta}) = \left\{ \beta : \frac{(\beta - \hat{\beta})^2}{\sigma_\beta^2/(n-2)} \leq F_{1,n-2,\alpha} \right\},$$

where  $\hat{\beta}$  is the MLE of  $\beta$  and  $\sigma_\beta^2$  is the previously defined consistent estimator of  $\sigma_\beta^2$ .

b. Show that the Creasy–Williams set can be written in the form

$$\left\{ \beta : \frac{(\beta - \hat{\beta})^2}{\sigma_\beta^2/(n-2)} \left[ \frac{(1 + \lambda\beta\hat{\beta})^2}{(1 + \lambda\beta^2)^2} \right] \leq F_{1,n-2,\alpha} \right\}.$$

Hence  $C_G(\hat{\beta})$  can be derived by replacing the term in square brackets with 1, its probability limit. (In deriving this representation, the fact that  $\hat{\beta}$  and  $-1/(\lambda\hat{\beta})$  are roots of the numerator of  $r_\lambda(\beta)$  is of great help. In particular, the fact that

$$\frac{r_\lambda^2(\beta)}{1 - r_\lambda^2(\beta)} = \frac{\lambda^2 S_{xy}^2 (\beta - \hat{\beta})^2 (\beta + (1/\lambda\hat{\beta}))^2}{(1 + \lambda\beta^2)^2 (S_{xx} S_{yy} - S_{xy}^2)}$$

is straightforward to establish.)

## Miscellanea

---

### Shapes of Confidence Bands

Confidence bands come in many shapes, not just the *hyperbolic* shape defined by the Scheffé band. For example, Gafarian (1964) showed how to construct a *straight-line* band over a finite interval. Gafarian-type bands allow statements of the form

$$P(\hat{\alpha} + \hat{\beta}x - d_\alpha \leq \alpha + \beta x \leq \hat{\alpha} + \hat{\beta}x + d_\alpha, \text{ for all } x \in [a, b]) = 1 - \alpha.$$

Gafarian gave tables of  $d_\alpha$ . A finite-width band must, necessarily, apply only to a finite range of  $x$ . Any band of level  $1 - \alpha$  must have infinite length as  $|x| \rightarrow \infty$ .

Casella and Strawderman (1980), among others, showed how to construct Scheffé-type bands over finite intervals, thereby reducing width while maintaining the same confidence as the infinite Scheffé band. Naiman (1983) compared performance of straight-line and Scheffé bands over finite intervals. Under his criterion, one of average width, the Scheffé band is superior. In some cases, an experimenter might be more comfortable with the interpretation of a straight-line band, however.

Shapes other than straight-line and hyperbolic are possible. Piegorsch (1985) investigated and characterized the shapes that are admissible in the sense that their probability statements cannot be improved upon. He obtained "growth conditions" that must be satisfied by an admissible band.

### The Meaning of Functional and Structural

The names *functional* and *structural* are, in themselves, a prime source of confusion in the EIV model. Kendall and Stuart (1979) give a detailed discussion of these concepts, distinguishing among relationships between *mathematical* (nonrandom) variables and relationships between random variables. One way to see the relationship is to write the models in a hierarchy in which the structural relationship model is obtained by putting a distribution on the parameters of the functional model:

$$\begin{array}{ll} \text{Functional} & \left\{ \begin{array}{ll} E(Y_i|\xi_i) = \alpha + \beta\xi_i + \epsilon_i, & \epsilon_i \sim n(0, \sigma_\epsilon^2) \\ E(X_i|\xi_i) = \xi_i + \delta_i, & \delta_i \sim n(0, \sigma_\delta^2) \\ \xi_i \sim n(\xi, \sigma_\xi^2) & \end{array} \right. \\ \text{relationship} & \text{model} \\ \text{model} & \end{array} \quad \begin{array}{l} \text{Structural} \\ \text{relationship} \\ \text{model} \end{array}$$

The difference in the words may be understood through the following distinction, not a universally accepted one. In the subject of calculus, for example, we often see the equation  $y = f(x)$ , an equation that describes a *functional* relationship, that is, a relationship that is *assumed to exist between variables*. Thus, using the idea that a functional relationship is an assumed relationship between two variables, the equation  $\eta_i = \alpha + \beta\xi_i$ , where  $\eta_i = E(Y_i|\xi_i)$ , is a functional (hypothesized) relationship in either the functional or structural relationship model.

On the other hand, a *structural* relationship is a relationship that *arises from the hypothesized structure of the problem*. Thus, in the structural relationship model, the relationship  $\eta = EY_i = \alpha + \beta\xi = \alpha + \beta EX_i$  can be deduced from the structure of the model; hence it is a structural relationship.

To make these ideas clearer, consider the case of simple linear regression where we assume that there is no error in the  $xs$ . The equation  $E(Y_i|x_i) = \alpha + \beta x_i$  is a functional relationship, a relationship that is hypothesized to exist between  $E(Y_i|x_i)$  and  $x_i$ . We can, however, also do simple linear regression under the assumption that the pair  $(X_i, Y_i)$  has a

bivariate normal distribution and we operate conditional on the  $x_i$ s. In this case, the relationship  $E(Y_i|x_i) = \alpha + \beta x_i$  follows from the structure of the hypothesized model, hence is a structural relationship.

Notice that, using these meanings, the distinction in terminology becomes a matter of taste. In any model we can deduce structural relations from functional relations and vice versa. The important distinction is whether the nuisance parameters, the  $\xi_i$ s, are integrated out before inference is done.

### **Consistency of Ordinary Least Squares in EIV Models**

In general it is not a good idea to use the ordinary least squares estimator to estimate the slope in EIV regression. This is because the estimator is inconsistent. Suppose that we assume a linear structural relationship (12.3.6). We have

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \\ &\rightarrow \frac{\text{Cov}(X, Y)}{\text{Var } X} \quad (\text{as } n \rightarrow \infty, \text{ using the WLLN}) \\ &= \beta \frac{\sigma_\xi^2}{\sigma_\delta^2 + \sigma_\xi^2}, \quad (\text{from (12.3.19)})\end{aligned}$$

showing that  $\hat{\beta}$  cannot be consistent. The same type of result can be obtained in the functional relationship case.

The behavior of  $\hat{\beta}$  in EIV models is treated in Cochran (1968). Carroll, Gallo, and Gleser (1985) and Gleser, Carroll, and Gallo (1987) investigated conditions under which functions of the ordinary least squares estimator are consistent.

### **Instrumental Variables in EIV Models**

The concept of instrumental variables goes back to, at least, Wald (1940), who constructed a consistent estimator of the slope with their help. To see what an instrumental variable is, write the EIV model in the form

$$Y_i = \alpha + \beta \xi_i + \epsilon_i,$$

$$X_i = \xi_i + \delta_i,$$

and do some algebra to get

$$Y_i = \alpha + \beta X_i + [\epsilon_i - \beta \delta_i].$$

An *instrumental variable*,  $Z_i$ , is a random variable that predicts  $X_i$  well, but is uncorrelated with  $\nu_i = \epsilon_i - \beta \delta_i$ . If such a variable can be identified, it can be used to improve predictions. In particular, it can be used to construct a consistent estimator of  $\beta$ .

Wald (1940) showed that, under fairly general conditions, the estimator

$$\hat{\beta}_W = \frac{\bar{Y}_{(1)} - \bar{Y}_{(2)}}{\bar{X}_{(1)} - \bar{X}_{(2)}}$$

is a consistent estimator of  $\beta$  in identifiable models, where the subscripts refer to two groupings of the data. A variable  $Z_i$ , which takes on only two values to define the grouping, is an instrumental variable. See Moran (1971) for a discussion of Wald's estimator.

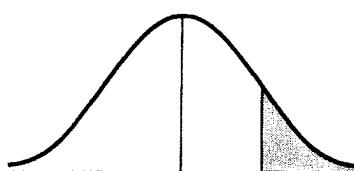
Although instrumental variables can be of great help, there can be some problems associated with their use. For example, Feldstein (1974) showed instances where the use of instrumental variables can be detrimental. Moran (1971) discussed the difficulty of verifying the conditions needed to ensure consistency of a simple estimator like  $\hat{\beta}_W$ . Fuller (1987) provided an in-depth discussion of instrumental variables. A model proposed by Berkson (1950) exploited a correlation structure similar to that used with instrumental variables.

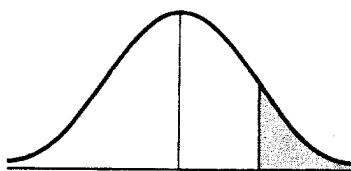


# Tables

1. Normal distribution, right-hand tail probabilities
2. Cutoff points for Student's  $t$  distribution, right-hand tail probabilities
3. Cutoff points for the chi squared distribution, right-hand tail probabilities
4. Cutoff points for the  $F$  distribution, right-hand tail probabilities

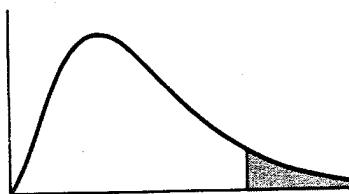
**TABLE 1** Normal distribution, right-hand tail probabilities



**TABLE 2** Cutoff points for Student's *t* distribution, right-hand tail probabilities

df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1,000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
$\infty$	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291

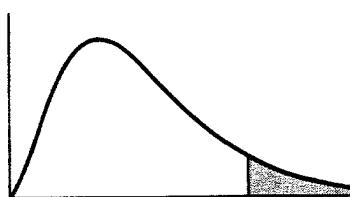
TABLE 3 Cutoff points for the chi squared distribution, right-hand tail probabilities



df	.9995	.999	.9975	.995	.990	.975	.950	.900	.750	.500
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.10	0.45
2	0.00	0.00	0.01	0.01	0.02	0.05	0.10	0.21	0.58	1.39
3	0.02	0.02	0.04	0.07	0.11	0.22	0.35	0.58	1.21	2.37
4	0.06	0.09	0.14	0.21	0.30	0.48	0.71	1.06	1.92	3.36
5	0.16	0.21	0.31	0.41	0.55	0.83	1.15	1.61	2.67	4.35
6	0.30	0.38	0.53	0.68	0.87	1.24	1.64	2.20	3.45	5.35
7	0.48	0.60	0.79	0.99	1.24	1.69	2.17	2.83	4.25	6.35
8	0.71	0.86	1.10	1.34	1.65	2.18	2.73	3.49	5.07	7.34
9	0.97	1.15	1.45	1.73	2.09	2.70	3.33	4.17	5.90	8.34
10	1.26	1.48	1.83	2.16	2.56	3.25	3.94	4.87	6.74	9.34
11	1.59	1.83	2.23	2.60	3.05	3.82	4.57	5.58	7.58	10.34
12	1.93	2.21	2.66	3.07	3.57	4.40	5.23	6.30	8.44	11.34
13	2.31	2.62	3.11	3.57	4.11	5.01	5.89	7.04	9.30	12.34
14	2.70	3.04	3.58	4.07	4.66	5.63	6.57	7.79	10.17	13.34
15	3.11	3.48	4.07	4.60	5.23	6.26	7.26	8.55	11.04	14.34
16	3.54	3.94	4.57	5.14	5.81	6.91	7.96	9.31	11.91	15.34
17	3.98	4.42	5.09	5.70	6.41	7.56	8.67	10.09	12.79	16.34
18	4.44	4.90	5.62	6.26	7.01	8.23	9.39	10.86	13.68	17.34
19	4.91	5.41	6.17	6.84	7.63	8.91	10.12	11.65	14.56	18.34
20	5.40	5.92	6.72	7.43	8.26	9.59	10.85	12.44	15.45	19.34
21	5.90	6.45	7.29	8.03	8.90	10.28	11.59	13.24	16.34	20.34
22	6.40	6.98	7.86	8.64	9.54	10.98	12.34	14.04	17.24	21.34
23	6.92	7.53	8.45	9.26	10.20	11.69	13.09	14.85	18.14	22.34
24	7.45	8.08	9.04	9.89	10.86	12.40	13.85	15.66	19.04	23.34
25	7.99	8.65	9.65	10.52	11.52	13.12	14.61	16.47	19.94	24.34
26	8.54	9.22	10.26	11.16	12.20	13.84	15.38	17.29	20.84	25.34
27	9.09	9.80	10.87	11.81	12.88	14.57	16.15	18.11	21.75	26.34
28	9.66	10.39	11.50	12.46	13.56	15.31	16.93	18.94	22.66	27.34
29	10.23	10.99	12.13	13.12	14.26	16.05	17.71	19.77	23.57	28.34
30	10.80	11.59	12.76	13.79	14.95	16.79	18.49	20.60	24.48	29.34
40	16.91	17.92	19.42	20.71	22.16	24.43	26.51	29.05	33.66	39.34
50	23.46	24.67	26.46	27.99	29.71	32.36	34.76	37.69	42.94	49.33
60	30.34	31.74	33.79	35.53	37.48	40.48	43.19	46.46	52.29	59.33
80	44.79	46.52	49.04	51.17	53.54	57.15	60.39	64.28	71.14	79.33
100	59.90	61.92	64.86	67.33	70.06	74.22	77.93	82.36	90.13	99.33

TABLE 3 (continued)

df	.250	.200	.150	.100	.050	.025	.020	.010	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2

**TABLE 4** Cutoff points for the  $F$  distribution, right-hand tail probabilities

Denominator df	$\alpha$	Numerator df									
		1	2	3	4	5	6	7	8	9	10
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
	.050	161.45	199.50	215.71	224.58	230.10	233.99	236.77	238.88	240.54	241.88
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5	6055.8
2	.001	405284	500000	540379	562500	576405	585937	592873	598144	602284	605621
	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
3	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39	999.40
	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86	129.25
	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
5	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05
6	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
7	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92
	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
8	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41

TABLE 4 (continued)

Denominator df	$\alpha$	Numerator df									
		12	15	20	25	30	40	50	60	120	1,000
1	.100	60.71	61.22	61.74	62.05	62.26	62.53	62.69	62.79	63.06	63.30
	.050	243.91	245.95	248.01	249.26	250.10	251.14	251.77	252.20	253.25	254.11
	.025	976.71	984.87	993.10	998.08	1001.4	1005.6	1008.1	1009.8	1014.0	1017.7
	.010	6106.3	6157.3	6208.7	6239.8	6260.6	6286.8	6302.5	6313.0	6339.4	6362.7
2	.001	610668	615764	620908	624017	626099	628712	630285	631337	633972	636301
	.100	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.47	9.48	9.49
	.050	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.48	19.49	19.49
	.025	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.48	39.49	39.50
3	.010	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.48	99.49	99.50
	.001	999.42	999.43	999.45	999.46	999.47	999.47	999.48	999.48	999.49	999.50
	.100	5.22	5.20	5.18	5.17	5.17	5.16	5.15	5.15	5.14	5.13
	.050	8.74	8.70	8.66	8.63	8.62	8.59	8.58	8.57	8.55	8.53
4	.025	14.34	14.25	14.17	14.12	14.08	14.04	14.01	13.99	13.95	13.91
	.010	27.05	26.87	26.69	26.58	26.50	26.41	26.35	26.32	26.22	26.14
	.001	128.32	127.37	126.42	125.84	125.45	124.96	124.66	124.47	123.97	123.53
	.100	3.90	3.87	3.84	3.83	3.82	3.80	3.80	3.79	3.78	3.76
5	.050	5.91	5.86	5.80	5.77	5.75	5.72	5.70	5.69	5.66	5.63
	.025	8.75	8.66	8.56	8.50	8.46	8.41	8.38	8.36	8.31	8.26
	.010	14.37	14.20	14.02	13.91	13.84	13.75	13.69	13.65	13.56	13.47
	.001	47.41	46.76	46.10	45.70	45.43	45.09	44.88	44.75	44.40	44.09
6	.100	3.27	3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.12	3.11
	.050	4.68	4.62	4.56	4.52	4.50	4.46	4.44	4.43	4.40	4.37
	.025	6.52	6.43	6.33	6.27	6.23	6.18	6.14	6.12	6.07	6.02
	.010	9.89	9.72	9.55	9.45	9.38	9.29	9.24	9.20	9.11	9.03
7	.001	26.42	25.91	25.39	25.08	24.87	24.60	24.44	24.33	24.06	23.82
	.100	2.90	2.87	2.84	2.81	2.80	2.78	2.77	2.76	2.74	2.72
	.050	4.00	3.94	3.87	3.83	3.81	3.77	3.75	3.74	3.70	3.67
	.025	5.37	5.27	5.17	5.11	5.07	5.01	4.98	4.96	4.90	4.86
8	.010	7.72	7.56	7.40	7.30	7.23	7.14	7.09	7.06	6.97	6.89
	.001	17.99	17.56	17.12	16.85	16.67	16.44	16.31	16.21	15.98	15.77

(continued)

TABLE 4 (Continued) Cutoff points for the  $F$  distribution, right-hand tail probabilities

Denominator df	$\alpha$	Numerator df									
		1	2	3	4	5	6	7	8	9	10
7	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08
8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54
9	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89
10	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
	.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75
11	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
	.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53
	.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92
12	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
	.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
	.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29
13	.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
	.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25
	.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
	.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80

TABLE 4 (Continued)

Denominator df	$\alpha$	Numerator df									
		12	15	20	25	30	40	50	60	120	1,000
7	.100	2.67	2.63	2.59	2.57	2.56	2.54	2.52	2.51	2.49	2.47
	.050	3.57	3.51	3.44	3.40	3.38	3.34	3.32	3.30	3.27	3.23
	.025	4.67	4.57	4.47	4.40	4.36	4.31	4.28	4.25	4.20	4.15
	.010	6.47	6.31	6.16	6.06	5.99	5.91	5.86	5.82	5.74	5.66
	.001	13.71	13.32	12.93	12.69	12.53	12.33	12.20	12.12	11.91	11.72
8	.100	2.50	2.46	2.42	2.40	2.38	2.36	2.35	2.34	2.32	2.30
	.050	3.28	3.22	3.15	3.11	3.08	3.04	3.02	3.01	2.97	2.93
	.025	4.20	4.10	4.00	3.94	3.89	3.84	3.81	3.78	3.73	3.68
	.010	5.67	5.52	5.36	5.26	5.20	5.12	5.07	5.03	4.95	4.87
	.001	11.19	10.84	10.48	10.26	10.11	9.92	9.80	9.73	9.53	9.36
9	.100	2.38	2.34	2.30	2.27	2.25	2.23	2.22	2.21	2.18	2.16
	.050	3.07	3.01	2.94	2.89	2.86	2.83	2.80	2.79	2.75	2.71
	.025	3.87	3.77	3.67	3.60	3.56	3.51	3.47	3.45	3.39	3.34
	.010	5.11	4.96	4.81	4.71	4.65	4.57	4.52	4.48	4.40	4.32
	.001	9.57	9.24	8.90	8.69	8.55	8.37	8.26	8.19	8.00	7.84
10	.100	2.28	2.24	2.20	2.17	2.16	2.13	2.12	2.11	2.08	2.06
	.050	2.91	2.85	2.77	2.73	2.70	2.66	2.64	2.62	2.58	2.54
	.025	3.62	3.52	3.42	3.35	3.31	3.26	3.22	3.20	3.14	3.09
	.010	4.71	4.56	4.41	4.31	4.25	4.17	4.12	4.08	4.00	3.92
	.001	8.45	8.13	7.80	7.60	7.47	7.30	7.19	7.12	6.94	6.78
11	.100	2.21	2.17	2.12	2.10	2.08	2.05	2.04	2.03	2.00	1.98
	.050	2.79	2.72	2.65	2.60	2.57	2.53	2.51	2.49	2.45	2.41
	.025	3.43	3.33	3.23	3.16	3.12	3.06	3.03	3.00	2.94	2.89
	.010	4.40	4.25	4.10	4.01	3.94	3.86	3.81	3.78	3.69	3.61
	.001	7.63	7.32	7.01	6.81	6.68	6.52	6.42	6.35	6.18	6.02
12	.100	2.15	2.10	2.06	2.03	2.01	1.99	1.97	1.96	1.93	1.91
	.050	2.69	2.62	2.54	2.50	2.47	2.43	2.40	2.38	2.34	2.30
	.025	3.28	3.18	3.07	3.01	2.96	2.91	2.87	2.85	2.79	2.73
	.010	4.16	4.01	3.86	3.76	3.70	3.62	3.57	3.54	3.45	3.37
	.001	7.00	6.71	6.40	6.22	6.09	5.93	5.83	5.76	5.59	5.44
13	.100	2.10	2.05	2.01	1.98	1.96	1.93	1.92	1.90	1.88	1.85
	.050	2.60	2.53	2.46	2.41	2.38	2.34	2.31	2.30	2.25	2.21
	.025	3.15	3.05	2.95	2.88	2.84	2.78	2.74	2.72	2.66	2.60
	.010	3.96	3.82	3.66	3.57	3.51	3.43	3.38	3.34	3.25	3.18
	.001	6.52	6.23	5.93	5.75	5.63	5.47	5.37	5.30	5.14	4.99

(continued)

**TABLE 4 (Continued)** Cutoff points for the *F* distribution, right-hand tail probabilities

Denominator df	$\alpha$	Numerator df									
		1	2	3	4	5	6	7	8	9	10
14	.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
	.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15
	.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
	.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40
15	.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
	.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
	.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08
16	.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
	.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99
	.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
	.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81
17	.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
	.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92
	.010	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59
	.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58
18	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39
19	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22
20	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08

TABLE 4 (Continued)

Denominator df	$\alpha$	Numerator df									
		12	15	20	25	30	40	50	60	120	1,000
14	.100	2.05	2.01	1.96	1.93	1.91	1.89	1.87	1.86	1.83	1.80
	.050	2.53	2.46	2.39	2.34	2.31	2.27	2.24	2.22	2.18	2.14
	.025	3.05	2.95	2.84	2.78	2.73	2.67	2.64	2.61	2.55	2.50
	.010	3.80	3.66	3.51	3.41	3.35	3.27	3.22	3.18	3.09	3.02
	.001	6.13	5.85	5.56	5.38	5.25	5.10	5.00	4.94	4.77	4.62
15	.100	2.02	1.97	1.92	1.89	1.87	1.85	1.83	1.82	1.79	1.76
	.050	2.48	2.40	2.33	2.28	2.25	2.20	2.18	2.16	2.11	2.07
	.025	2.96	2.86	2.76	2.69	2.64	2.59	2.55	2.52	2.46	2.40
	.010	3.67	3.52	3.37	3.28	3.21	3.13	3.08	3.05	2.96	2.88
	.001	5.81	5.54	5.25	5.07	4.95	4.80	4.70	4.64	4.47	4.33
16	.100	1.99	1.94	1.89	1.86	1.84	1.81	1.79	1.78	1.75	1.72
	.050	2.42	2.35	2.28	2.23	2.19	2.15	2.12	2.11	2.06	2.02
	.025	2.89	2.79	2.68	2.61	2.57	2.51	2.47	2.45	2.38	2.32
	.010	3.55	3.41	3.26	3.16	3.10	3.02	2.97	2.93	2.84	2.76
	.001	5.55	5.27	4.99	4.82	4.70	4.54	4.45	4.39	4.23	4.08
17	.100	1.96	1.91	1.86	1.83	1.81	1.78	1.76	1.75	1.72	1.69
	.050	2.38	2.31	2.23	2.18	2.15	2.10	2.08	2.06	2.01	1.97
	.025	2.82	2.72	2.62	2.55	2.50	2.44	2.41	2.38	2.32	2.26
	.010	3.46	3.31	3.16	3.07	3.00	2.92	2.87	2.83	2.75	2.66
	.001	5.32	5.05	4.78	4.60	4.48	4.33	4.24	4.18	4.02	3.87
18	.100	1.93	1.89	1.84	1.80	1.78	1.75	1.74	1.72	1.69	1.66
	.050	2.34	2.27	2.19	2.14	2.11	2.06	2.04	2.02	1.97	1.92
	.025	2.77	2.67	2.56	2.49	2.44	2.38	2.35	2.32	2.26	2.20
	.010	3.37	3.23	3.08	2.98	2.92	2.84	2.78	2.75	2.66	2.58
	.001	5.13	4.87	4.59	4.42	4.30	4.15	4.06	4.00	3.84	3.69
19	.100	1.91	1.86	1.81	1.78	1.76	1.73	1.71	1.70	1.67	1.64
	.050	2.31	2.23	2.16	2.11	2.07	2.03	2.00	1.98	1.93	1.88
	.025	2.72	2.62	2.51	2.44	2.39	2.33	2.30	2.27	2.20	2.14
	.010	3.30	3.15	3.00	2.91	2.84	2.76	2.71	2.67	2.58	2.50
	.001	4.97	4.70	4.43	4.26	4.14	3.99	3.90	3.84	3.68	3.53
20	.100	1.89	1.84	1.79	1.76	1.74	1.71	1.69	1.68	1.64	1.61
	.050	2.28	2.20	2.12	2.07	2.04	1.99	1.97	1.95	1.90	1.85
	.025	2.68	2.57	2.46	2.40	2.35	2.29	2.25	2.22	2.16	2.09
	.010	3.23	3.09	2.94	2.84	2.78	2.69	2.64	2.61	2.52	2.43
	.001	4.82	4.56	4.29	4.12	4.00	3.86	3.77	3.70	3.54	3.40

(continued)

**TABLE 4 (Continued)** Cutoff points for the *F* distribution, right-hand tail probabilities

Denominator df	$\alpha$	Numerator df									
		1	2	3	4	5	6	7	8	9	10
21	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73
	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95
22	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70
	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83
23	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67
	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73
24	.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
	.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64
	.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
	.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64
25	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
	.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
	.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
	.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56
26	.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
	.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59
	.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
	.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48
27	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
	.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57
	.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
	.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41

TABLE 4 (Continued)

Denominator	df	$\alpha$	Numerator df									
			12	15	20	25	30	40	50	60	1,000	
21	.100		1.87	1.83	1.78	1.74	1.72	1.69	1.67	1.66	1.62	1.59
	.050		2.25	2.18	2.10	2.05	2.01	1.96	1.94	1.92	1.87	1.82
	.025		2.64	2.53	2.42	2.36	2.31	2.25	2.21	2.18	2.11	2.05
	.010		3.17	3.03	2.88	2.79	2.72	2.64	2.58	2.55	2.46	2.37
	.001		4.70	4.44	4.17	4.00	3.88	3.74	3.64	3.58	3.42	3.28
22	.100		1.86	1.81	1.76	1.73	1.70	1.67	1.65	1.64	1.60	1.57
	.050		2.23	2.15	2.07	2.02	1.98	1.94	1.91	1.89	1.84	1.79
	.025		2.60	2.50	2.39	2.32	2.27	2.21	2.17	2.14	2.08	2.01
	.010		3.12	2.98	2.83	2.73	2.67	2.58	2.53	2.50	2.40	2.32
	.001		4.58	4.33	4.06	3.89	3.78	3.63	3.54	3.48	3.32	3.17
23	.100		1.84	1.80	1.74	1.71	1.69	1.66	1.64	1.62	1.59	1.55
	.050		2.20	2.13	2.05	2.00	1.96	1.91	1.88	1.86	1.81	1.76
	.025		2.57	2.47	2.36	2.29	2.24	2.18	2.14	2.11	2.04	1.98
	.010		3.07	2.93	2.78	2.69	2.62	2.54	2.48	2.45	2.35	2.27
	.001		4.48	4.23	3.96	3.79	3.68	3.53	3.44	3.38	3.22	3.08
24	.100		1.83	1.78	1.73	1.70	1.67	1.64	1.62	1.61	1.57	1.54
	.050		2.18	2.11	2.03	1.97	1.94	1.89	1.86	1.84	1.79	1.74
	.025		2.54	2.44	2.33	2.26	2.21	2.15	2.11	2.08	2.01	1.94
	.010		3.03	2.89	2.74	2.64	2.58	2.49	2.44	2.40	2.31	2.22
	.001		4.39	4.14	3.87	3.71	3.59	3.45	3.36	3.29	3.14	2.99
25	.100		1.82	1.77	1.72	1.68	1.66	1.63	1.61	1.59	1.56	1.52
	.050		2.16	2.09	2.01	1.96	1.92	1.87	1.84	1.82	1.77	1.72
	.025		2.51	2.41	2.30	2.23	2.18	2.12	2.08	2.05	1.98	1.91
	.010		2.99	2.85	2.70	2.60	2.54	2.45	2.40	2.36	2.27	2.18
	.001		4.31	4.06	3.79	3.63	3.52	3.37	3.28	3.22	3.06	2.91
26	.100		1.81	1.76	1.71	1.67	1.65	1.61	1.59	1.58	1.54	1.51
	.050		2.15	2.07	1.99	1.94	1.90	1.85	1.82	1.80	1.75	1.70
	.025		2.49	2.39	2.28	2.21	2.16	2.09	2.05	2.03	1.95	1.89
	.010		2.96	2.81	2.66	2.57	2.50	2.42	2.36	2.33	2.23	2.14
	.001		4.24	3.99	3.72	3.56	3.44	3.30	3.21	3.15	2.99	2.84
27	.100		1.80	1.75	1.70	1.66	1.64	1.60	1.58	1.57	1.53	1.50
	.050		2.13	2.06	1.97	1.92	1.88	1.84	1.81	1.79	1.73	1.68
	.025		2.47	2.36	2.25	2.18	2.13	2.07	2.03	2.00	1.93	1.86
	.010		2.93	2.78	2.63	2.54	2.47	2.38	2.33	2.29	2.20	2.11
	.001		4.17	3.92	3.66	3.49	3.38	3.23	3.14	3.08	2.92	2.78

(continued)

TABLE 4 (Continued) Cutoff points for the  $F$  distribution, right-hand tail probabilities

Denominator df	$\alpha$	Numerator df									
		1	2	3	4	5	6	7	8	9	10
28	.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
	.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55
	.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
	.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35
29	.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
	.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53
	.010	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
	.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29
30	.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
	.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51
	.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
	.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24
40	.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
	.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
	.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
	.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87
50	.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73
	.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
	.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32
	.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70
	.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82	3.67
60	.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
	.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27
	.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
	.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54
100	.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66
	.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93
	.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18
	.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50
	.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44	3.30

TABLE 4 (Continued)

Denominator df	$\alpha$	Numerator df									
		12	15	20	25	30	40	50	60	120	1,000
28	.100	1.79	1.74	1.69	1.65	1.63	1.59	1.57	1.56	1.52	1.48
	.050	2.12	2.04	1.96	1.91	1.87	1.82	1.79	1.77	1.71	1.66
	.025	2.45	2.34	2.23	2.16	2.11	2.05	2.01	1.98	1.91	1.84
	.010	2.90	2.75	2.60	2.51	2.44	2.35	2.30	2.26	2.17	2.08
	.001	4.11	3.86	3.60	3.43	3.32	3.18	3.09	3.02	2.86	2.72
29	.100	1.78	1.73	1.68	1.64	1.62	1.58	1.56	1.55	1.51	1.47
	.050	2.10	2.03	1.94	1.89	1.85	1.81	1.77	1.75	1.70	1.65
	.025	2.43	2.32	2.21	2.14	2.09	2.03	1.99	1.96	1.89	1.82
	.010	2.87	2.73	2.57	2.48	2.41	2.33	2.27	2.23	2.14	2.05
	.001	4.05	3.80	3.54	3.38	3.27	3.12	3.03	2.97	2.81	2.66
30	.100	1.77	1.72	1.67	1.63	1.61	1.57	1.55	1.54	1.50	1.46
	.050	2.09	2.01	1.93	1.88	1.84	1.79	1.76	1.74	1.68	1.63
	.025	2.41	2.31	2.20	2.12	2.07	2.01	1.97	1.94	1.87	1.80
	.010	2.84	2.70	2.55	2.45	2.39	2.30	2.25	2.21	2.11	2.02
	.001	4.00	3.75	3.49	3.33	3.22	3.07	2.98	2.92	2.76	2.61
40	.100	1.71	1.66	1.61	1.57	1.54	1.51	1.48	1.47	1.42	1.38
	.050	2.00	1.92	1.84	1.78	1.74	1.69	1.66	1.64	1.58	1.52
	.025	2.29	2.18	2.07	1.99	1.94	1.88	1.83	1.80	1.72	1.65
	.010	2.66	2.52	2.37	2.27	2.20	2.11	2.06	2.02	1.92	1.82
	.001	3.64	3.40	3.14	2.98	2.87	2.73	2.64	2.57	2.41	2.25
50	.100	1.68	1.63	1.57	1.53	1.50	1.46	1.44	1.42	1.38	1.33
	.050	1.95	1.87	1.78	1.73	1.69	1.63	1.60	1.58	1.51	1.45
	.025	2.22	2.11	1.99	1.92	1.87	1.80	1.75	1.72	1.64	1.56
	.010	2.56	2.42	2.27	2.17	2.10	2.01	1.95	1.91	1.80	1.70
	.001	3.44	3.20	2.95	2.79	2.68	2.53	2.44	2.38	2.21	2.05
60	.100	1.66	1.60	1.54	1.50	1.48	1.44	1.41	1.40	1.35	1.30
	.050	1.92	1.84	1.75	1.69	1.65	1.59	1.56	1.53	1.47	1.40
	.025	2.17	2.06	1.94	1.87	1.82	1.74	1.70	1.67	1.58	1.49
	.010	2.50	2.35	2.20	2.10	2.03	1.94	1.88	1.84	1.73	1.62
	.001	3.32	3.08	2.83	2.67	2.55	2.41	2.32	2.25	2.08	1.92
100	.100	1.61	1.56	1.49	1.45	1.42	1.38	1.35	1.34	1.28	1.22
	.050	1.85	1.77	1.68	1.62	1.57	1.52	1.48	1.45	1.38	1.30
	.025	2.08	1.97	1.85	1.77	1.71	1.64	1.59	1.56	1.46	1.36
	.010	2.37	2.22	2.07	1.97	1.89	1.80	1.74	1.69	1.57	1.45
	.001	3.07	2.84	2.59	2.43	2.32	2.17	2.08	2.01	1.83	1.64

(continued)

**TABLE 4 (Continued)** Cutoff points for the *F* distribution, right-hand tail probabilities

Denominator df	$\alpha$	Numerator df									
		1	2	3	4	5	6	7	8	9	10
200	.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63
	.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88
	.025	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11
	.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41
	.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26	3.12
1,000	.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64	1.61
	.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84
	.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13	2.06
	.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34
	.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13	2.99

TABLE 4 (Continued)

Denominator	df	$\alpha$	Numerator df									
			12	15	20	25	30	40	50	60	1,000	
200	.100		1.58	1.52	1.46	1.41	1.38	1.34	1.31	1.29	1.23	1.16
	.050		1.80	1.72	1.62	1.56	1.52	1.46	1.41	1.39	1.30	1.21
	.025		2.01	1.90	1.78	1.70	1.64	1.56	1.51	1.47	1.37	1.25
	.010		2.27	2.13	1.97	1.87	1.79	1.69	1.63	1.58	1.45	1.30
1,000	.001		2.90	2.67	2.42	2.26	2.15	2.00	1.90	1.83	1.64	1.43
	.100		1.55	1.49	1.43	1.38	1.35	1.30	1.27	1.25	1.18	1.08
	.050		1.76	1.68	1.58	1.52	1.47	1.41	1.36	1.33	1.24	1.11
	.025		1.96	1.85	1.72	1.64	1.58	1.50	1.45	1.41	1.29	1.13
	.010		2.20	2.06	1.90	1.79	1.72	1.61	1.54	1.50	1.35	1.16
	.001		2.77	2.54	2.30	2.14	2.02	1.87	1.77	1.69	1.49	1.22

# Table of Common Distributions

## Discrete Distributions

---

### Bernoulli( $p$ )

pmf  $P(X = x|p) = p^x(1 - p)^{1-x}; \quad x = 0, 1; \quad 0 \leq p \leq 1$

mean and variance  $EX = p, \quad \text{Var } X = p(1 - p)$

mgf  $M_X(t) = (1 - p) + pe^t$

---

### Binomial( $n, p$ )

pmf  $P(X = x|n, p) = \binom{n}{x} p^x(1 - p)^{n-x}; \quad x = 0, 1, 2, \dots, n; \quad 0 \leq p \leq 1$

mean and variance  $EX = np, \quad \text{Var } X = np(1 - p)$

mgf  $M_X(t) = [pe^t + (1 - p)]^n$

notes Related to Binomial Theorem (Theorem 3.1.1). The *multinomial* distribution (Definition 4.6.1) is a multivariate version of the binomial distribution.

---

### Discrete Uniform

pmf  $P(X = x|N) = \frac{1}{N}; \quad x = 1, 2, \dots, N; \quad N = 1, 2, \dots$

mean and variance  $EX = \frac{N+1}{2}, \quad \text{Var } X = \frac{(N+1)(N-1)}{12}$

mgf  $M_X(t) = \frac{1}{N} \sum_{i=1}^N e^{it}$

**Geometric( $p$ )**

*pmf*  $P(X = x|p) = p(1 - p)^{x-1}; \quad x = 1, 2, \dots; \quad 0 \leq p \leq 1$

*mean and variance*  $EX = \frac{1}{p}, \quad \text{Var } X = \frac{1-p}{p^2}$

*mgf*  $M_X(t) = \frac{pe^t}{1-(1-p)e^t}, \quad t < -\log(1-p)$

*notes*  $Y = X - 1$  is negative binomial(1,  $p$ ). The distribution is *memoryless*:  $P(X > s|X > t) = P(X > s - t)$ .

**Hypergeometric**

*pmf*  $P(X = x|N, M, K) = \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}}; \quad x = 0, 1, 2, \dots, K;$   
 $M - (N - K) \leq x \leq M; \quad N, M, K \geq 0$

*mean and variance*  $EX = \frac{KM}{N}, \quad \text{Var } X = \frac{KM}{N} \frac{(N-M)(N-K)}{N(N-1)}$

*notes* If  $K \ll M$  and  $N$ , the range  $x = 0, 1, 2, \dots, K$  will be appropriate.

**Negative binomial( $r, p$ )**

*pmf*  $P(X = x|r, p) = \binom{r+x-1}{x} p^r (1-p)^x; \quad x = 0, 1, \dots; \quad 0 \leq p \leq 1$

*mean and variance*  $EX = \frac{r(1-p)}{p}, \quad \text{Var } X = \frac{r(1-p)}{p^2}$

*mgf*  $M_X(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r, \quad t < -\log(1-p)$

*notes* An alternate form of the pmf is given by  $P(Y = y|r, p) = \binom{y-1}{r-1} p^r (1-p)^{y-r}$ ,  $y = r, r+1, \dots$ . The random variable  $Y = X + r$ . The negative binomial can be derived as a gamma mixture of Poissons. (See Exercise 4.34.)

**Poisson( $\lambda$ )**

*pmf*  $P(X = x|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!}; \quad x = 0, 1, \dots; \quad 0 \leq \lambda < \infty$

*mean and variance*  $EX = \lambda, \quad \text{Var } X = \lambda$

*mgf*  $M_X(t) = e^{\lambda(e^t - 1)}$

## Continuous Distributions

---

**Beta( $\alpha, \beta$ )**

*pdf*  $f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 \leq x \leq 1, \quad \alpha > 0, \quad \beta > 0$

*mean and variance*  $EX = \frac{\alpha}{\alpha+\beta}, \quad \text{Var } X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

*mgf*  $M_X(t) = 1 + \sum_{k=1}^{\infty} \left( \prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{t^k}{k!}$

*notes* The constant in the beta pdf can be defined in terms of gamma functions,  $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ . Equation (3.2.18) gives a general expression for the moments.

---

**Cauchy( $\theta, \sigma$ )**

*pdf*  $f(x|\theta, \sigma) = \frac{1}{\pi\sigma} \frac{1}{1 + \left(\frac{x-\theta}{\sigma}\right)^2}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty, \quad \sigma > 0$

*mean and variance* do not exist

*mgf* does not exist

*notes* Special case of Student's  $t$ , when degrees of freedom = 1. Also, if  $X$  and  $Y$  are independent  $n(0, 1)$ ,  $X/Y$  is Cauchy.

---

**Chi squared**

*pdf*  $f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}; \quad 0 \leq x < \infty; \quad p = 1, 2, \dots$

*mean and variance*  $EX = p, \quad \text{Var } X = 2p$

*mgf*  $M_X(t) = \left(\frac{1}{1-2t}\right)^{p/2}, \quad t < \frac{1}{2}$

*notes* Special case of the gamma distribution.

---

**Double exponential( $\mu, \sigma$ )**

*pdf*  $f(x|\mu, \sigma) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$

*mean and variance*  $EX = \mu, \quad \text{Var } X = 2\sigma^2$

<i>mgf</i>	$M_X(t) = \frac{e^{\mu t}}{1-(\sigma t)^2}, \quad  t  < \frac{1}{\sigma}$
<i>notes</i>	Also known as the <i>Laplace</i> distribution.

---

**Exponential( $\beta$ )**

*pdf*  $f(x|\beta) = \frac{1}{\beta} e^{-x/\beta}, \quad 0 \leq x < \infty, \quad \beta > 0$

*mean and variance*  $EX = \beta, \quad \text{Var } X = \beta^2$

*mgf*  $M_X(t) = \frac{1}{1-\beta t}, \quad t < \frac{1}{\beta}$

*notes* Special case of the gamma distribution. Has the *memoryless* property. Has many special cases:  $Y = X^{1/\gamma}$  is *Weibull*,  $Y = \sqrt{2X/\beta}$  is *Rayleigh*,  $Y = \alpha - \gamma \log(X/\beta)$  is *Gumbel*.

---

**F**

*pdf*  $f(x|\nu_1, \nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{x^{(\nu_1-2)/2}}{\left(1+\left(\frac{\nu_1}{\nu_2}\right)x\right)^{(\nu_1+\nu_2)/2}}, \quad 0 \leq x < \infty;$   
 $\nu_1, \nu_2 = 1, \dots$

*mean and variance*  $EX = \frac{\nu_2}{\nu_2-2}, \quad \nu_2 > 2,$

$$\text{Var } X = 2 \left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)}, \quad \nu_2 > 4$$

*moments*  $(\text{mgf does not exist}) \quad EX^n = \frac{\Gamma\left(\frac{\nu_1+2n}{2}\right)\Gamma\left(\frac{\nu_2-2n}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_2}{\nu_1}\right)^n, \quad n < \frac{\nu_2}{2}$

*notes* Related to chi squared ( $F_{\nu_1, \nu_2} = \left(\frac{\chi_{\nu_1}^2}{\nu_1}\right) / \left(\frac{\chi_{\nu_2}^2}{\nu_2}\right)$ , where the  $\chi^2$ s are independent) and *t* ( $F_{1,\nu} = t_{\nu}^2$ ).

---

**Gamma( $\alpha, \beta$ )**

*pdf*  $f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, \quad 0 \leq x < \infty, \quad \alpha, \beta > 0$

*mean and variance*  $EX = \alpha\beta, \quad \text{Var } X = \alpha\beta^2$

*mgf*  $M_X(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}, \quad t < \frac{1}{\beta}$

*notes* Some special cases are exponential ( $\alpha = 1$ ) and chi squared ( $\alpha = p/2, \beta = 2$ ). If  $\alpha = \frac{3}{2}$ ,  $Y = \sqrt{X/\beta}$  is *Maxwell*.  $Y = 1/X$  has the *inverted gamma distribution*. Can also be related to the Poisson (Example 3.2.1).

**Logistic( $\mu, \beta$ )**

<i>pdf</i>	$f(x \mu, \beta) = \frac{1}{\beta} \frac{e^{-(x-\mu)/\beta}}{[1+e^{-(x-\mu)/\beta}]^2}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty,$ $\beta > 0$
<i>mean and variance</i>	$EX = \mu, \quad \text{Var } X = \frac{\pi^2 \beta^2}{3}$
<i>mgf</i>	$M_X(t) = e^{\mu t} \Gamma(1 - \beta t) \Gamma(1 + \beta t), \quad  t  < \frac{1}{\beta}$
<i>notes</i>	The cdf is given by $F(x \mu, \beta) = \frac{1}{1+e^{-(x-\mu)/\beta}}$ .

**Lognormal( $\mu, \sigma^2$ )**

<i>pdf</i>	$f(x \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \frac{e^{-(\log x - \mu)^2/(2\sigma^2)}}{x}, \quad 0 < x < \infty, \quad -\infty < \mu < \infty,$ $\sigma > 0$
<i>mean and variance</i>	$EX = e^{\mu + (\sigma^2/2)}, \quad \text{Var } X = e^{2(\mu + \sigma^2)} - e^{2\mu + \sigma^2}$
<i>moments</i> ( <i>mgf does not exist</i> )	$EX^n = e^{n\mu + n^2\sigma^2/2}$
<i>notes</i>	Example 2.3.5 gives another distribution with the same moments.

**Normal( $\mu, \sigma^2$ )**

<i>pdf</i>	$f(x \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty,$ $-\infty < \mu < \infty, \quad \sigma > 0$
<i>mean and variance</i>	$EX = \mu, \quad \text{Var } X = \sigma^2$
<i>mgf</i>	$M_X(t) = e^{\mu t + \sigma^2 t^2/2}$
<i>notes</i>	Sometimes called the <i>Gaussian</i> distribution.

**Pareto( $\alpha, \beta$ )**

<i>pdf</i>	$f(x \alpha, \beta) = \frac{\beta\alpha^\beta}{x^{\beta+1}}, \quad \alpha < x < \infty, \quad \alpha > 0, \quad \beta > 0$
<i>mean and variance</i>	$EX = \frac{\beta\alpha}{\beta-1}, \quad \beta > 1,$ $\text{Var } X = \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}, \quad \beta > 2$
<i>mgf</i>	does not exist

***t***

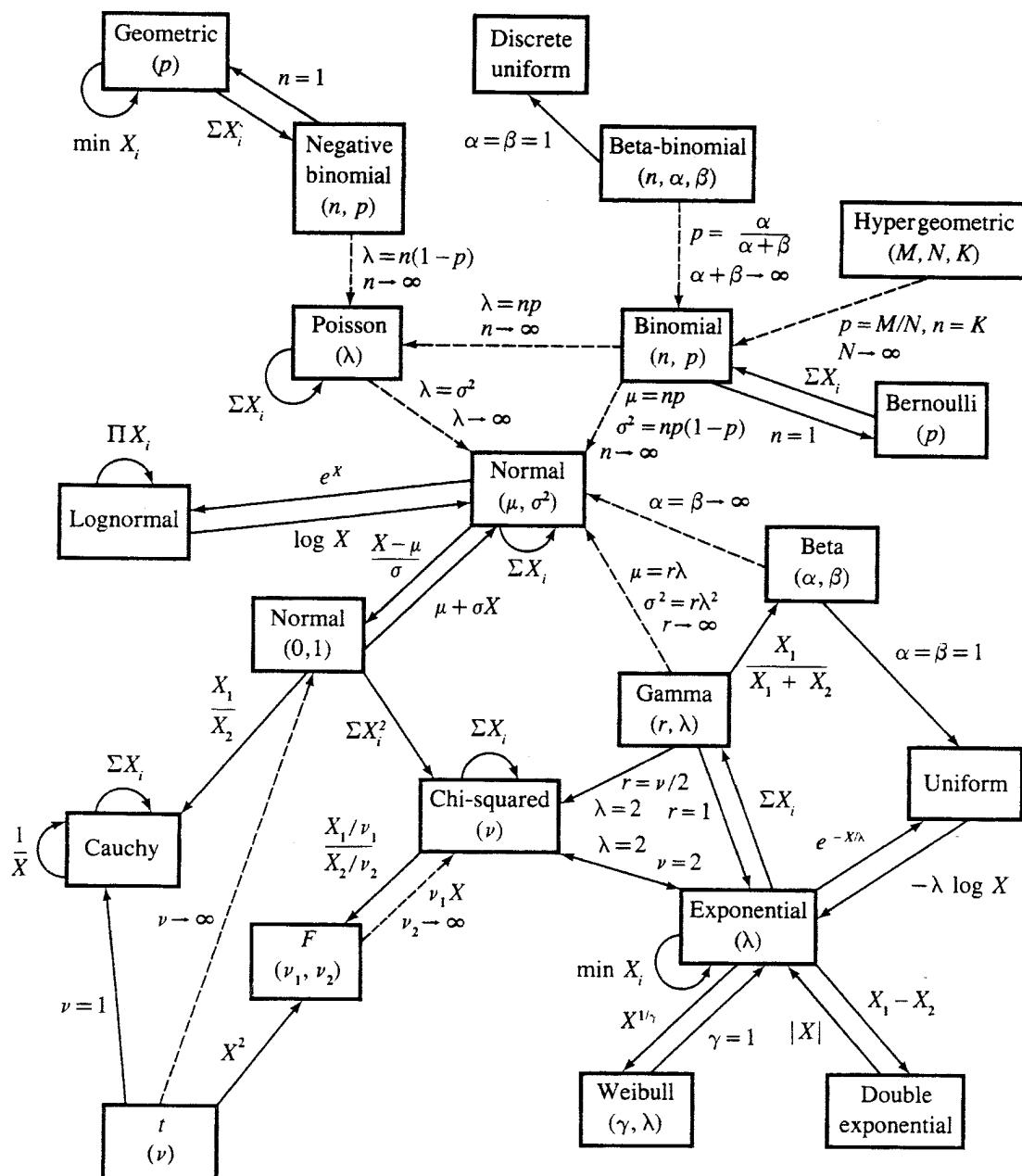
<i>pdf</i>	$f(x \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\nu\pi}} \frac{1}{(1+(\frac{x^2}{\nu}))^{(\nu+1)/2}}, \quad -\infty < x < \infty, \quad \nu = 1, \dots$
<i>mean and variance</i>	$EX = 0, \quad \nu > 1,$ $\text{Var } X = \frac{\nu}{\nu-2}, \quad \nu > 2$
<i>moments (mgf does not exist)</i>	$EX^n = \frac{\Gamma(\frac{n+1}{2})\Gamma(\frac{\nu-n}{2})}{\sqrt{\pi}\Gamma(\frac{\nu}{2})} \nu^{n/2}$ if $n < \nu$ and even, $EX^n = 0$ if $n < \nu$ and odd.
<i>notes</i>	Related to $F$ ( $F_{1,\nu} = t_\nu^2$ ).

**Uniform( $a, b$ )**

<i>pdf</i>	$f(x a, b) = \frac{1}{b-a}, \quad a \leq x \leq b$
<i>mean and variance</i>	$EX = \frac{b+a}{2}, \quad \text{Var } X = \frac{(b-a)^2}{12}$
<i>mgf</i>	$M_X(t) = \frac{e^{bt}-e^{at}}{(b-a)t}$
<i>notes</i>	If $a = 0$ and $b = 1$ , this is a special case of the beta ( $\alpha = \beta = 1$ ).

**Weibull( $\gamma, \beta$ )**

<i>pdf</i>	$f(x \gamma, \beta) = \frac{\gamma}{\beta} x^{\gamma-1} e^{-x^\gamma/\beta}, \quad 0 \leq x < \infty, \quad \gamma > 0, \quad \beta > 0$
<i>mean and variance</i>	$EX = \beta^{1/\gamma} \Gamma\left(1 + \frac{1}{\gamma}\right), \quad \text{Var } X = \beta^{2/\gamma} \left[ \Gamma\left(1 + \frac{2}{\gamma}\right) - \Gamma^2\left(1 + \frac{1}{\gamma}\right) \right]$
<i>moments</i>	$EX^n = \beta^{n/\gamma} \Gamma\left(1 + \frac{n}{\gamma}\right)$
<i>notes</i>	The mgf exists only for $\gamma \geq 1$ . Its form is not very useful. A special case is exponential ( $\gamma = 1$ ).



**Relationships among common distributions.** Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

# References

- Anderson, T. W. (1976). Estimation of Linear Functional Relationships: Approximate Distributions and Connection with Simultaneous Equations in Econometrics (with discussion). *Journal of the Royal Statistical Society, Series B* 38, 1–36.
- Anderson, T. W. (1984a). *An Introduction to Multivariate Statistical Analysis*, 2nd edition. New York: Wiley.
- Anderson, T. W. (1984b). Estimating Linear Statistical Relationships. *Annals of Statistics* 12, 1–45.
- Barlow, R. and Proschan, F. (1975). *Statistical Theory of Life Testing*. New York: Holt, Rinehart and Winston.
- Barnard, G. A. (1949). Statistical Inference (with discussion). *Journal of the Royal Statistical Society, Series B* 11, 115–139.
- Barnard, G. A. (1980). Pivotal Inference and the Bayesian Controversy (with discussion). *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.). Valencia: University Press.
- Barr, D. R. and Zehna, P. W. (1983). *Probability: Modeling Uncertainty*. Reading, Massachusetts: Addison-Wesley.
- Basu, D. (1959). The Family of Ancillary Statistics. *Sankhyā, Series A* 21, 247–256.
- Bechhofer, R. E. (1954). A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with Known Variances. *Annals of Mathematical Statistics* 25, 16–39.
- Berg, C. (1988). The Cube of a Normal Distribution Is Indeterminate. *Annals of Probability* 16, 910–913.
- Berger, J. O. (1975). Minimax Estimation of Location Vectors for a Wide Class of Densities. *Annals of Statistics* 3, 1318–1328.
- Berger, J. O. (1976). Admissible Minimax Estimation of a Multivariate Normal Mean with Arbitrary Quadratic Loss. *Annals of Statistics* 4, 223–226.
- Berger, J. O. (1984). The Robust Bayesian Viewpoint (with discussion). *Robustness of Bayesian Analysis* (J. Kadane, ed.), 63–144. Amsterdam: North-Holland.
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. New York: Springer-Verlag.
- Berger, J. O. and Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence (with discussion). *Journal of the American Statistical Association* 82, 112–122.
- Berger, J. O. and Wolpert, R. W. (1984). *The Likelihood Principle*. Institute of Mathematical Statistics Lecture Notes—Monograph Series. Hayward, California: IMS.

- Berger, R. L. (1982). Multiparameter Hypothesis Testing and Acceptance Sampling. *Technometrics* 24, 295–300.
- Berkson, J. (1950). Are There Two Regressions? *Journal of the American Statistical Association* 45, 164–180.
- Berliner, L. M. (1983). Improving on Inadmissible Estimators in the Control Problem. *Annals of Statistics* 11, 814–826.
- Betteley, I. G. (1977). The Addition Law for Expectations. *American Statistician* 31, 33–35.
- Bickel, P. J. and Doksum, K. A. (1981). An Analysis of Transformations Revisited. *Journal of the American Statistical Association* 76, 296–311.
- Bickel, P. J. and Mallows, C. L. (1988). A Note on Unbiased Bayes Estimates. *American Statistician* 42, 132–134.
- Birnbaum, A. (1962). On the Foundations of Statistical Inference (with discussion). *Journal of the American Statistical Association* 57, 269–306.
- Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Massachusetts: MIT Press.
- Blackwell, D. and Girshick, M. A. (1954). *Theory of Games and Statistical Decisions*. New York: Wiley.
- Blyth, C. R. (1951). On Minimax Statistical Decision Procedures and Their Admissibility. *Annals of Mathematical Statistics* 22, 22–42.
- Blyth, C. R. (1986). Approximate Binomial Confidence Limits. *Journal of the American Statistical Association* 81, 843–855; correction 84, 636.
- Blyth, C. R. and Still, H. A. (1983). Binomial Confidence Intervals. *Journal of the American Statistical Association* 78, 108–116.
- Bondar, J. V. and Milnes, P. (1981). Amenability: A Survey for Statistical Applications of Hunt–Stein and Related Conditions on Groups. *Z. Wahrsch. verw. Gebiete* 57, 103–128.
- Box, G. E. P. (1954). Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems I: Effect of Inequality of Variance in the One-Way Classification. *Annals of Mathematical Statistics* 25, 290–302.
- Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations (with discussion). *Journal of the Royal Statistical Society, Series B* 26, 211–252.
- Box, G. E. P. and Cox, D. R. (1982). An Analysis of Transformations Revisited, Rebutted. *Journal of the American Statistical Association* 77, 209–210.
- Box, G. E. P., Hunter, J. S., and Hunter, W. G. (1978). *Statistics for Experimenters*. New York: Wiley.
- Brewster, J. F. and Zidek, J. V. (1974). Improving on Equivariant Estimators. *Annals of Statistics* 2, 21–38.
- Brown, L. D. (1966). On the Admissibility of Invariant Estimators of One or More Location Parameters. *Annals of Mathematical Statistics* 37, 1087–1136.
- Brown, L. D. (1971). Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary-Value Problems. *Annals of Mathematical Statistics* 42, 855–903.
- Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics Lecture Notes - Monograph Series. Hayward, California: IMS.
- Brown, L. D. (1988). Lecture Notes. Department of Mathematics, Cornell University. Ithaca, New York.
- Brown, L. D. (1990). An Ancillarity Paradox Which Appears in Multiple Linear Regression (with discussion). *Annals of Statistics* 18, 471–538.
- Brown, L. D. and Purves, R. (1973). Measurable Selections of Extrema. *Annals of Statistics* 1, 902–912.

- Buehler, R. J. (1982). Some Ancillary Statistics and Their Properties (with discussion). *Journal of the American Statistical Association* 77, 581-594.
- Campbell, C. and Joiner, B. L. (1973). How to Get the Answer Without Being Sure You've Asked the Question. *American Statistician* 27, 229-231.
- Carmer, S. G. and Walker, W. M. (1982). Baby Bear's Dilemma: A Statistical Tale. *Agronomy Journal* 74, 122-124.
- Carroll, R. J., Gallo, P., and Gleser, L. J. (1985). Comparison of Least Squares and Errors-in-Variables Regression, with Special Reference to Randomized Analysis of Covariance. *Journal of the American Statistical Association* 80, 929-932.
- Casella, G. (1985). An Introduction to Empirical Bayes Data Analysis. *American Statistician* 39, 83-87.
- Casella, G. (1986). Refining Binomial Confidence Intervals. *Canadian Journal of Statistics* 14, 113-129.
- Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem (with discussion). *Journal of the American Statistical Association* 82, 106-111.
- Casella, G. and Hwang, J. T. (1983). Empirical Bayes Confidence Sets for the Mean of a Multivariate Distribution. *Journal of the American Statistical Association* 78, 688-698.
- Casella, G. and Hwang, J. T. (1987). Employing Vague Prior Information in the Construction of Confidence Sets. *Journal of Multivariate Analysis* 21, 79-104.
- Casella, G. and Strawderman, W. E. (1980). Confidence Bands for Linear Regression with Restricted Predictor Variables. *Journal of the American Statistical Association* 75, 862-868.
- Casella, G. and Strawderman, W. E. (1981). Estimating a Bounded Normal Mean. *Annals of Statistics* 9, 868-876.
- Chapman, D. G. and Robbins, H. (1951). Minimum Variance Estimation Without Regularity Assumptions. *Annals of Mathematical Statistics* 22, 581-586.
- Chung, K. L. (1974). *A Course in Probability Theory*. New York: Academic Press.
- Clopper, C. J. and Pearson, E. S. (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika* 26, 404-413. (Also in *The Selected Papers of E. S. Pearson*, Cambridge University Press, 1966.)
- Cochran, W. G. (1934). The Distribution of Quadratic Forms in a Normal System with Applications to the Analysis of Variance. *Proceedings of the Cambridge Philosophical Society* 30, 178-191.
- Cochran, W. G. (1968). Errors of Measurement in Statistics. *Technometrics* 10, 637-666.
- Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*, 2nd edition. New York: Wiley.
- Cornfield, J. and Tukey, J. (1956). Average Values of Mean Squares in Factorials. *Annals of Mathematical Statistics* 27, 907-949.
- Cox, D. R. (1958). Some Problems Connected with Statistical Inference. *Annals of Mathematical Statistics* 29, 357-372.
- Cox, D. R. (1971). The Choice Between Ancillary Statistics. *Journal of the Royal Statistical Society, Series B* 33, 251-255.
- Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton, New Jersey: Princeton University Press.
- Creasy, M. A. (1956). Confidence Limits for the Gradient in the Linear Functional Relationship. *Journal of the Royal Statistical Society, Series B* 18, 65-69.
- Crow, E. L. (1956). Confidence Intervals for a Proportion. *Biometrika* 43, 423-425.
- David, H. A. (1985). Bias of  $S^2$  Under Dependence. *American Statistician* 39, 201.

- Davidson, R. R. and Solomon, D. L. (1974). Moment-Type Estimation in the Exponential Family. *Communications in Statistics* 3, 1101–1108.
- deFinetti, B. (1972). *Probability, Induction, and Statistics*. London: Wiley.
- Dempster, A. P., Selwyn, M. R., Patel, C. M., and Roth, A. J. (1984). Statistical and Computational Aspects of Mixed Model Analysis. *Applied Statistics (Journal of the Royal Statistical Society, Series C)* 33, 203–214.
- Draper, N. R. and Smith H. (1981). *Applied Regression Analysis*, 2nd edition. New York: Wiley.
- Duling, D. R., Motten, A. G., and Mason, R. P. (1988). Generation and Evaluation of Isotropic ESR Spectrum Simulations. *Journal of Magnetic Resonance* 77, 504–511.
- Durbin, J. (1970). On Birnbaum's Theorem and the Relation Between Sufficiency, Conditionality, and Likelihood. *Journal of the American Statistical Association* 65, 395–398.
- Dynkin, E. B. (1951). Necessary and Sufficient Statistics for a Family of Probability Distributions. English translation in *Selected Translations in Mathematical Statistics and Probability* 1 (1961), 23–41.
- Eberhardt, K. R. and Fligner, M. A. (1977). A Comparison of Two Tests for Equality of Two Proportions. *American Statistician* 31, 151–155.
- Efron, B. F. and Hinkley, D. V. (1978). Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information. *Biometrika* 65, 457–487.
- Efron, B. F. and Morris, C. N. (1972). Limiting the Risk of Bayes and Empirical Bayes Estimators Part II: The Empirical Bayes Case. *Journal of the American Statistical Association* 67, 130–139.
- Efron, B. F. and Morris, C. N. (1973). Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach. *Journal of the American Statistical Association* 68, 117–130.
- Efron, B. F. and Morris, C. N. (1975). Data Analysis Using Stein's Estimator and Its Generalizations. *Journal of the American Statistical Association* 70, 311–319.
- Efron, B. F. and Morris, C. N. (1977). Stein's Paradox in Statistics. *Scientific American* 236(5), 119–127.
- Feldman, D. and Fox, M. (1968). Estimation of the Parameter  $n$  in the Binomial Distribution. *Journal of the American Statistical Association* 63, 150–158.
- Feldstein, M. (1974). Errors in Variables: A Consistent Estimator with Smaller MSE in Finite Samples. *Journal of the American Statistical Association* 69, 990–996.
- Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Volume I*. New York: Wiley.
- Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Volume II*. New York: Wiley.
- Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.
- Fieller, E. C. (1954). Some Problems in Interval Estimation. *Journal of the Royal Statistical Society, Series B* 16, 175–185.
- Fisher, R. A. (1925). Theory of Statistical Estimation. *Proceedings of the Cambridge Philosophical Society* 22, 700–725.
- Fisher, R. A. (1935). The Fiducial Argument in Statistical Inference. *Annals of Eugenics* 6, 391–398. (Also in R. A. Fisher, *Contributions to Mathematical Statistics*, New York: Wiley, 1950.)
- Fisher, R. A. (1939). The Comparison of Samples with Possibly Unequal Variances. *Annals of Eugenics* 9, 174–180. (Also in R. A. Fisher, *Contributions to Mathematical Statistics*, New York: Wiley, 1950.)
- Fraser, D. A. S. (1968). *The Structure of Inference*. New York: Wiley.

- Fraser, D. A. S. (1979). *Inference and Linear Models*. New York: McGraw-Hill.
- Freedman, D., Pisani, R., and Purves, R. (1978). *Statistics*. New York: Norton.
- Fuller, W. A. (1987). *Measurement Error Models*. New York: Wiley.
- Gafarian, A. V. (1964). Confidence Bands in Straight Line Regression. *Journal of the American Statistical Association* 59, 182–213.
- Gardner, M. (1961). *The Second Scientific American Book of Mathematical Puzzles and Diversions*. New York: Simon and Schuster.
- Garwood, F. (1936). Fiducial Limits for the Poisson Distribution. *Biometrika* 28, 437–442.
- Ghosh, B. K. (1979). A Comparison of Some Approximate Confidence Intervals for the Binomial Parameter. *Journal of the American Statistical Association* 74, 894–900.
- Ghosh, M. and Meeden, G. (1977). On the Non-Attainability of Chebychev Bounds. *American Statistician* 31, 35–36.
- Gianola, D. and Fernando, R. L. (1986). Bayesian Methods in Animal Breeding Theory. *Journal of Animal Science* 63, 217–244.
- Gilat, D. (1977). Monotonicity of a Power Function: An Elementary Probabilistic Proof. *American Statistician* 31, 91–93.
- Gleser, L. J. (1981). Estimation in a Multivariate Errors-in-Variables Regression Model: Large-Sample Results. *Annals of Statistics* 9, 24–44.
- Gleser, L. J. (1983). Functional, Structural, and Ultrastructural Errors-in-Variables Models. *Proceedings of the Business and Economic Statistics Section*, 57–66. Alexandria, Virginia: American Statistical Association.
- Gleser, L. J. (1987). Confidence Intervals for the Slope in a Linear Errors-in-Variables Regression Model. *Advances in Multivariate Statistical Analysis* (K. Gupta, ed.), 85–109. Dordrecht: D. Reidel.
- Gleser, L. J. (1989). The Gamma Distribution as a Mixture of Exponential Distributions. *American Statistician* 43, 115–117.
- Gleser, L. J., Carroll, R. J., and Gallo, P. (1987). The Limiting Distribution of Least Squares in an Errors-in-Variables Regression Model. *Annals of Statistics* 15, 220–233.
- Gleser, L. J. and Healy, J. D. (1976). Estimating the Mean of a Normal Distribution with Known Coefficient of Variation. *Journal of the American Statistical Association* 71, 977–981.
- Gleser, L. J. and Hwang, J. T. (1987). The Nonexistence of  $100(1-\alpha)\%$  Confidence Sets of Finite Expected Diameter in Errors-in-Variables and Related Models. *Annals of Statistics* 15, 1351–1362.
- Guenther, W. C. (1978). Some Easily Found Minimum Variance Unbiased Estimators. *American Statistician* 32, 29–33.
- Gupta, S. S. (1965). On Some Multiple Decision (Selection and Ranking) Rules. *Technometrics* 7, 225–245.
- Gupta, S. S. and Panchapakesan, S. (1979). *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*. New York: Wiley.
- Hall, W. J., Wijsman, R. A., and Ghosh, J. K. (1965). The Relationship Between Sufficiency and Invariance with Applications in Sequential Analysis. *Annals of Mathematical Statistics* 36, 575–614.
- Halmos, P. R. and Savage, L. J. (1949). Applications of the Radon–Nikodym Theorem to the Theory of Sufficient Statistics. *Annals of Mathematical Statistics* 20, 225–241.
- Hampel, F. R. (1974). The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association* 69, 383–393.
- Hardy, G. H., Littlewood, J. E., and Polya, G. (1952). *Inequalities*, 2nd edition. London: Cambridge University Press.

- Harville, D. A. (1981). Unbiased and Minimum-Variance Unbiased Estimation of Estimable Functions for Fixed Linear Models with Arbitrary Covariance Structure. *Annals of Statistics* 9, 633–637.
- Hayter, A. J. (1984). A Proof of the Conjecture that the Tukey–Kramer Multiple Comparison Procedure Is Conservative. *Annals of Statistics* 12, 61–75.
- Hinkley, D. V. and Rungger, G. (1984). The Analysis of Transformed Data (with discussion). *Journal of the American Statistical Association* 79, 302–320.
- Hocking, R. R. (1973). A Discussion of the Two-Way Mixed Model. *American Statistician* 27, 148–152.
- Hocking, R. R. (1985). *The Analysis of Linear Models*. Pacific Grove, California: Brooks/Cole.
- Huber, P. J. (1967). The Behaviour of Maximum Likelihood Estimates Under Nonstandard Conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics, and Probability* 1, 221–233. Berkeley, California: University of California Press.
- Hudson, H. M. (1978). A Natural Identity for Exponential Families with Applications in Multiparameter Estimation. *Annals of Statistics* 6, 473–484.
- Huzurbazar, V. S. (1949). On a Property of Distributions Admitting Sufficient Statistics. *Biometrika* 36, 71–74.
- Hwang, J. T. (1982). Improving on Standard Estimators in Discrete Exponential Families with Applications to Poisson and Negative Binomial Cases. *Annals of Statistics* 10, 857–867.
- Hwang, J. T. (1990). Fieller's Theorem and Resampling Plans. Technical Report. Statistics Center, Cornell University, Ithaca, New York.
- Hwang, J. T. and Casella, G. (1982). Minimax Confidence Sets for the Mean of a Multivariate Normal Distribution. *Annals of Statistics* 10, 868–881.
- James, W. and Stein, C. (1961). Estimation with Quadratic Loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1, 361–380. Berkeley, California: University of California Press.
- Johnson, N. L. and Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. New York: Houghton Mifflin.
- Johnson, N. L. and Kotz, S. (1970a). *Distributions in Statistics: Continuous Univariate Distributions* I. New York: Houghton Mifflin.
- Johnson, N. L. and Kotz, S. (1970b). *Distributions in Statistics: Continuous Univariate Distributions* II. New York: Houghton Mifflin.
- Joshi, V. M. (1967). Inadmissibility of the Usual Confidence Sets for the Mean of a Multivariate Normal Population. *Annals of Mathematical Statistics* 38, 1868–1875.
- Joshi, V. M. (1969). Admissibility of the Usual Confidence Sets for the Mean of a Univariate or Bivariate Normal Population. *Annals of Mathematical Statistics* 40, 1042–1067.
- Kalbfleisch, J. D. (1975). Sufficiency and Conditionality. *Biometrika* 62, 251–268.
- Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- Kelker, D. (1970). Distribution Theory of Spherical Distributions and a Location–Scale Parameter Generalization. *Sankhyā, Series A* 32, 419–430.
- Kendall, M. and Stuart, A. (1979). *The Advanced Theory of Statistics, Volume II: Inference and Relationship*, 4th edition. New York: Macmillan.
- Kiefer, J. C. (1957). Invariance, Minimax Sequential Estimation, and Continuous Time Processes. *Annals of Mathematical Statistics* 28, 573–601.
- Kirk, R. E. (1982). *Experimental Design: Procedures for the Behavioral Sciences*, 2nd edition. Pacific Grove, California: Brooks/Cole.
- Kotz, S., Johnson, N. L., and Read, C. B. (1982). *Encyclopedia of Statistical Sciences* (Nine Volumes). New York: Wiley.

- Leemis, L. M. (1986). Relationships Among Common Univariate Distributions. *American Statistician* 40, 143–146.
- Lehmann, E. L. (1981). An Interpretation of Completeness and Basu's Theorem. *Journal of the American Statistical Association* 76, 335–340.
- Lehmann, E. L. (1983). *Theory of Point Estimation*. New York: Wiley.
- Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd edition. New York: Wiley.
- Lehmann, E. L. and Scheffé, H. (1950, 1955, 1956). Completeness, Similar Regions, and Unbiased Estimation. *Sankhyā, Series A* 10, 305–340; 15, 219–236; correction 17, 250.
- Lindley, D. V. (1957). A Statistical Paradox. *Biometrika* 44, 187–192.
- Lindley, D. V. (1962). Discussion of the article by Stein. *Journal of the Royal Statistical Society, Series B* 24, 265–296.
- Lindley, D. V. and Phillips, L. D. (1976). Inference for a Bernoulli Process (a Bayesian View). *American Statistician* 30, 112–119.
- Lindley, D. V. and Smith, A. F. M. (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society, Series B* 34, 1–41.
- Maatta, J. M. and Casella, G. (1987). Conditional Properties of Interval Estimators of the Normal Variance. *Annals of Statistics* 15, 1372–1388.
- Madansky, A. (1962). More on Length of Confidence Intervals. *Journal of the American Statistical Association* 57, 586–589.
- Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. New York: Academic Press.
- Miller, R. G. (1974). The Jackknife—A Review. *Biometrika* 61, 1–15.
- Miller, R. G. (1981). *Simultaneous Statistical Inference*, 2nd edition. New York: Springer-Verlag.
- Montgomery, D. C. (1984). *Design and Analysis of Experiments*. New York: Wiley.
- Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the Theory of Statistics*, 3rd edition. New York: McGraw-Hill.
- Moors, J. J. A. (1981). Inadmissibility of Linearly Invariant Estimators in Truncated Parameter Spaces. *Journal of the American Statistical Association* 76, 910–915.
- Moran, P. A. P. (1971). Estimating Structural and Functional Relationships. *Journal of Multivariate Analysis* 1, 232–255.
- Morris, C. N. (1982). Natural Exponential Families with Quadratic Variance Functions. *Annals of Statistics* 10, 65–80.
- Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications (with discussion). *Journal of the American Statistical Association* 78, 47–65.
- Morrison, D. G. (1978). A Probability Model for Forced Binary Choices. *American Statistician* 32, 23–25.
- Naiman, D. Q. (1983) Comparing Scheffé-Type to Constant Width Confidence Bounds in Regression. *Journal of the American Statistical Association* 78, 906–912.
- Neter, J., Wasserman, W., and Kutner, M. H. (1985). *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*. Homewood, Illinois: R. D. Irwin.
- Noorbaloochi, S. and Meeden, G. (1983). Unbiasedness as the Dual of Being Bayes. *Journal of the American Statistical Association* 78, 619–623.
- Norton, R. M. (1984). The Double Exponential Distribution: Using Calculus to Find an MLE. *American Statistician* 38, 135–136.
- Nussbaum, M. (1976). Maximum Likelihood and Least Squares Estimation of Linear Functional Relationships. *Mathematische Operationsforschung und Statistik, Series Statistik* 7, 23–49.

- Olkin, I., Petkau, A. J., and Zidek, J. V. (1981). A Comparison of  $n$  Estimators for the Binomial Distribution. *Journal of the American Statistical Association* 76, 637–642.
- Piegorsch, W. W. (1985). Admissible and Optimal Confidence Bands in Simple Linear Regression. *Annals of Statistics* 13, 801–810.
- Pitman, E. J. G. (1939). The Estimation of the Location and Scale Parameters of a Continuous Population of Any Given Form. *Biometrika* 30, 200–215.
- Pratt, J. W. (1961). Length of Confidence Intervals. *Journal of the American Statistical Association* 56, 549–567.
- Quenouille, M. H. (1956). Notes on Bias in Estimation. *Biometrika* 43, 353–360.
- Robbins, H. (1977). A Fundamental Question of Practical Statistics (Letter to the Editor). *American Statistician* 31, 97.
- Robson, D. S. (1959). A Simple Method for Constructing Orthogonal Polynomials when the Independent Variable Is Unequally Spaced. *Biometrics* 15, 187–191.
- Romano, J. P. and Siegel, A. F. (1986). *Counterexamples in Probability and Statistics*. Pacific Grove, California: Wadsworth and Brooks/Cole.
- Rudin, W. (1976). *Principles of Real Analysis*. New York: McGraw-Hill.
- Ruppert, D. (1987). What Is Kurtosis? *American Statistician* 41, 1–5.
- Sacks, J. (1963). Generalized Bayes Solutions in Estimation Problems. *Annals of Mathematical Statistics* 34, 751–768.
- Samuels, M. L., Casella, G., and McCabe, G. P. (1988). Are Blocks Different from Random Factors? Technical Report, Department of Statistics, Purdue University. West Lafayette, Indiana.
- Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin* (now called *Biometrics*) 2, 110–114.
- Saw, J. G., Yang, M. C. K., and Mo, T. C. (1984). Chebychev's Inequality with Estimated Mean and Variance. *American Statistician* 38, 130–132.
- Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.
- Searle, S. R. (1971). *Linear Models*. New York: Wiley.
- Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. New York: Wiley.
- Searle, S. R. (1987). *Linear Models for Unbalanced Data*. New York: Wiley.
- Searle, S. R. and Pukelsheim, F. (1985). Establishing  $\chi^2$  Properties of Sums of Squares Using Induction. *American Statistician* 39, 301–303.
- Selvin, S. (1975). A Problem in Probability (Letter to the Editor). *American Statistician* 29, 67.
- Smith, A. F. M. (1983). Discussion of the article by DuMouchel and Harris. *Journal of the American Statistical Association* 78, 310–311.
- Snedecor, G. W. and Cochran, W. G. (1989). *Statistical Methods*, 8th edition. Ames, Iowa: Iowa State University Press.
- Solari, M. E. (1969). The "Maximum Likelihood Solution" of the Problem of Estimating a Linear Functional Relationship. *Journal of the Royal Statistical Society, Series B* 31, 372–375.
- Solomon, D. L. (1975). A Note on the Non-Equivalence of the Neyman-Pearson and Generalized Likelihood Ratio Tests for Testing a Simple Null Versus a Simple Alternative Hypothesis. *American Statistician* 29, 101–102.
- Solomon, D. L. (1983). The Spatial Distribution of Cabbage Butterfly Eggs. *Life Science Models, Volume 4* (H. Marcus-Roberts and M. Thompson, eds.), 350–366. New York: Springer-Verlag.
- Stein, C. (1955). Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1, 197–206. Berkeley, California: University of California Press.

- Stein, C. (1964). Inadmissibility of the Usual Estimator for the Variance of a Normal Distribution with Unknown Mean. *Annals of the Institute of Statistical Mathematics* 16, 155–160.
- Stein, C. (1973). Estimation of the Mean of a Multivariate Distribution. *Proceedings of the Prague Symposium on Asymptotic Statistics*, 345–381.
- Stein, C. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *Annals of Statistics* 9, 1135–1151.
- Sterne, T. E. (1954). Some Remarks on Confidence or Fiducial Limits. *Biometrika* 41, 275–278.
- Stigler, S. M. (1983). Who Discovered Bayes' Theorem? *American Statistician* 37, 290–296.
- Stigler, S. M. (1984). Kruskal's Proof of the Joint Distribution of  $\bar{X}$  and  $S^2$ . *American Statistician* 38, 134–135.
- Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, Massachusetts: Harvard University Press.
- Strawderman, W. E. (1971). Proper Bayes Minimax Estimators of the Multivariate Normal Mean. *Annals of Mathematical Statistics* 42, 385–388.
- Stuart, A. and Ord, K. J. (1987). *Kendall's Advanced Theory of Statistics, Volume I: Distribution Theory*, 5th edition. New York: Oxford University Press.
- Tardiff, R. M. (1981). L'Hospital's Rule and the Central Limit Theorem. *American Statistician* 35, 43.
- Tate, R. F. and Klett, G. W. (1959). Optimal Confidence Intervals for the Variance of a Normal Distribution. *Journal of the American Statistical Association* 54, 674–682.
- Thomas, L. C. (1984). *Games, Theory and Applications*. New York: Wiley.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.
- Vardeman, S. B. (1987). Discussion of the articles by Casella and Berger and Berger and Sellke. *Journal of the American Statistical Association* 82, 130–131.
- Wald, A. (1940). The Fitting of Straight Lines when Both Variables Are Subject to Error. *Annals of Mathematical Statistics* 11, 284–300.
- Wald, A. (1949). Note on the Consistency of the Maximum Likelihood Estimate. *Annals of Mathematical Statistics* 20, 595–601.
- Wald, A. (1950). *Statistical Decision Functions*. New York: Wiley.
- Widder, D. V. (1946). *The Laplace Transform*. Princeton, New Jersey: Princeton University Press.
- Wilks, S. S. (1938). Shortest Average Confidence Intervals from Large Samples. *Annals of Mathematical Statistics* 9, 166–175.
- Williams, E. J. (1959). *Regression Analysis*. New York: Wiley.
- Zehna, P. W. (1966). Invariance of Maximum Likelihood Estimators. *Annals of Mathematical Statistics* 37, 744.
- Zellner, A. (1986). Bayesian Estimation and Prediction Using Asymmetric Loss Functions. *Journal of the American Statistical Association* 81, 446–451.



# Author Index

- Anderson, T. W., 500, 581, 594  
Barlow, R., 99  
Barnard, G. A., 413  
Barr, D. R., 127  
Basu, D., 283  
Bechhofer, R. E., 552  
Berg, C., 81  
Berger, J. O., 267, 389, 402, 413,  
    427, 429, 438, 462, 482, 485,  
    496, 497, 499, 508  
Berger, R. L., 357, 389, 402  
Berkson, J., 605  
Berliner, L. M., 492  
Betteley, I. G., 84  
Bickel, P. J., 344, 543  
Birnbaum, A., 267  
Bishop, Y. M., 383  
Blackwell, D., 507  
Blyth, C. R., 445, 448, 485  
Boes, D. C., 416  
Bondar, J. V., 507  
Box, G. E. P., 512, 509, 543  
Brewster, J. F., 302, 336, 481  
Brown, L. D., 115, 219, 474, 499,  
    500  
Buehler, R. J., 283  
  
Campbell, C., 334  
Carmer, S. G., 552  
Carroll, R. J., 604  
Casella, G., 335, 389, 402, 437,  
    448, 490, 500, 506, 508, 530,  
    542, 553, 603  
Chapman, D. G., 315  
Chung, K. L., 34, 37, 65, 82, 84,  
    213, 215, 219, 243  
Clopper, C. J., 413  
  
Cochran, W. G., 509, 512, 551, 553,  
    604  
Cox, D. R., 260, 283, 543  
Cox, G. M., 509, 553  
Cramér, H., 441  
Creasy, M. A., 593, 594  
Crow, E. L., 448  
  
David, H. A., 244  
Davidson, R. R., 342  
DeFinetti, B., 9, 236  
Dempster, A. P., 508  
Doksum, K. A., 543  
Draper, N. R., 554  
Duling, D. R., 40  
Durbin, J., 272  
Dynkin, E. B., 283  
  
Eberhardt, K. R., 385  
Efron, B. F., 260, 326, 495, 499,  
    506  
  
Feldman, D., 293  
Feldstein, M., 605  
Feller, W., 65, 66, 83, 84, 96, 127,  
    213, 219, 236  
Ferguson, T. S., 392, 483  
Fernando, R. L., 508  
Fieller, E. C., 459  
Fienberg, S. E., 383  
Fisher, R. A., 283, 512  
Fligner, M. A., 385  
Fox, M., 293  
Fraser, D. A. S., 413  
Freedman, D., 555  
Fuller, W. A., 581, 589, 594, 605  
  
Gafarian, A. V., 603  
Gallo, P., 604

- Gardner, M., 43, 192  
 Garwood, F., 422  
 Ghosh, B. K., 385  
 Ghosh, J. K., 278, 354  
 Ghosh, M., 200  
 Gianola, D., 508  
 Gilat, D., 401  
 Girshick, M. A., 507  
 Gleser, L. J., 195, 339, 584, 590,  
     592, 593, 604  
 Graybill, F. A., 416  
 Guenther, W. C., 340  
 Gupta, S. S., 552
- Hall, W. J., 278, 354  
 Halmos, P. R., 250  
 Hampel, F. R., 229  
 Hardy, G. H., 178  
 Harville, D. A., 564  
 Hayter, A. J., 551  
 Healy, J. D., 339  
 Hinkley, D. V., 260, 283, 326, 543  
 Hocking, R. R., 530, 542, 564  
 Holland, P. W., 383  
 Hsu, J., 451  
 Huber, P. J., 325, 441  
 Hudson, H. M., 188  
 Hunter, J. S., 509  
 Hunter, W. G., 509  
 Huzurbazar, V. S., 343  
 Hwang, J. T., 189, 500, 508, 553,  
     592
- James, W., 497  
 Johnson, N. L., 112  
 Joiner, B. L., 334  
 Joshi, V. M., 458, 472, 500
- Kalbfleisch, J. D., 103, 272  
 Kelker, D., 227  
 Kendall, M., 315, 325, 381, 441,  
     512, 554, 581, 587, 591, 594,  
     603  
 Kiefer, J. C., 507  
 Kirk, R. E., 509, 547, 553  
 Klett, G. W., 455  
 Kotz, S., 112
- Leemis, L. M., 630
- Lehmann, E. L., 115, 247, 255, 263,  
     264, 282, 337, 343, 354, 372,  
     376, 402, 427, 496, 507  
 Lindley, D. V., 271, 389, 508, 553  
 Littlewood, J. E., 178
- Maatta, J. M., 437  
 Madansky, A., 437, 453  
 Mallows, C. L., 344  
 Marshall, A. W., 178  
 Mason, R. P., 40  
 McCabe, G. P., 530, 542, 545  
 Meeden, G., 200, 344  
 Miller, R. G., 341, 552  
 Miller, R. P., 11  
 Milnes, P., 507  
 Montgomery, D. C., 509, 553  
 Mood, A. M., 416  
 Moors, J. J. A., 334  
 Moran, P. A. P., 583, 605  
 Morris, C. N., 97, 495, 499, 506,  
     508, 553  
 Morrison, D. G., 199  
 Motten, A. G., 40
- Naiman, D. Q., 603  
 Neter, J., 554  
 Noorbaloochi, S., 344  
 Norton, R. M., 333  
 Nussbaum, M., 584
- Olkin, I., 178, 297  
 Ord, K. J., 123, 195
- Panchapakesan, S., 552  
 Pearson, E. S., 413  
 Phillips, L. D., 271  
 Piegorsch, W. W., 603  
 Pisani, R., 555  
 Pitman, E. J. G., 337  
 Polya, G., 178  
 Pratt, J. W., 436  
 Prentice, R. L., 103  
 Proschan, F., 99  
 Pukelsheim, F., 547  
 Purves, R., 474, 555
- Quenouille, M. H., 341
- Robbins, H., 315, 385
- Robson, D. S., 571  
 Romano, J. P., 81, 82  
 Rudin, W., 70, 437  
 Runger, G., 543  
 Ruppert, D., 80
- Sacks, J., 499  
 Samuels, M. L., 530, 542  
 Satterthwaite, F. E., 287  
 Savage, L. J., 250  
 Saw, J. G., 245  
 Scheffé, H., 255, 263, 523, 530  
 Searle, S. R., 176, 526, 530, 541,  
     542, 547, 551  
 Sellke, T., 389, 402  
 Selvin, S., 42  
 Siegel, A. F., 81, 82  
 Smith, A. F. M., 508  
 Smith, H., 554  
 Snedecor, G. W., 512, 551  
 Solari, M. E., 582, 587  
 Solomon, D. L., 157, 342, 390  
 Stein, C., 187, 302, 497  
 Sterne, T. E., 417, 448  
 Stigler, S. M., 21, 222, 243, 555  
 Still, H. A., 448  
 Strawderman, W. E., 490, 499, 603  
 Stuart, A., 123, 195, 315, 325, 381,  
     441, 512, 554, 581, 587, 591,  
     594, 603
- Tardiff, R. M., 240  
 Tate, R. F., 455  
 Thomas, L. C., 507  
 Tukey, J. W., 230, 545
- Vardeman, S. B., 469
- Wald, A., 325, 482, 502, 604  
 Walker, W. M., 552  
 Widder, D. V., 66  
 Wijsman, R. A., 278, 354  
 Wilks, S. S., 442, 444, 457  
 Williams, E. J., 593  
 Wolpert, R. W., 267, 413
- Zehna, P. W., 127  
 Zellner, A., 502  
 Zidek, J. V., 302, 336, 481

# Subject Index

Acceptance sampling, 345, 357  
hypergeometric, 88  
Action space, 461  
Admissible decision rule, 480  
absolute error loss  
binomial, 480  
Bayes, 482, 503  
complete class, 481  
sample mean, 485  
squared error loss  
mean, 485  
Analysis of variance (*see* ANOVA;  
Randomized complete block  
design)  
Ancillary statistic, 257, 283  
exponential, 263–264  
location family, 258, 280  
location-scale family, 282  
sample size, 280  
scale family, 258–259  
uniform, 257  
usefulness, 259–260  
ANOVA  
assumptions, 512  
Bonferroni Inequality, 523  
cell means model, 511  
confidence interval, 517  
contrasts, 513–514, 517–518, 522  
orthogonal, 527  
sum of squares, 526  
data snooping, 524, 547  
equality of variance, 512  
expected mean squares, 547  
*F* test, 522  
relation to *t* test, 545  
identifiability, 511  
levels of a factor, 511  
linear combinations, 516  
LRT, 545

ANOVA (*continued*)  
null hypothesis, 512  
oneway, 509–510  
other models, 553  
overparametrized model, 510–511  
relation to RCB, 541  
Scheffé's method, 523–524, 545  
selection and ranking, 552  
Stein estimation, 552  
sufficient statistic, 543  
sum of squares, 525  
*t* test, 516–517  
relation to *F* test, 545  
table, 527  
transformation of data, 512, 543  
treatment means, 510  
union-intersection test, 514, 518  
variance estimator, 516  
Approximation  
binomial by normal, 106  
binomial by Poisson, 66, 94  
hypergeometric by binomial, 122  
hypergeometric by Poisson, 122  
negative binomial by normal, 219  
Arithmetic mean, 183  
Asymptotic efficiency, 325  
Axiom of Continuity, 38  
Axiom of Countable Additivity, 8,  
38  
Axiom of Finite Additivity, 9, 38  
Basu's Theorem, 262  
converse, 282  
Bayes estimator  
absolute error loss, 475  
binomial, 476–477  
mean, 476  
Bernoulli, 298, 305  
empirical, 508

- Bayes estimator (*continued*)
  - hierarchical, 508
  - multivariate, 496
  - normal, 299, 335
  - normal variance, 335
  - relation to unbiased estimator, 343
  - robust, 508
  - squared error loss, 475
    - binomial, 476
    - mean, 476, 486
    - weird, 503
  - Bayes risk, 463, 473, 486
  - Bayes' Rule, 21
  - Bayes rule
    - decision theoretic, 473
    - construction, 474
    - limits, 482
  - Behrens–Fisher problem, 396, 512, 582–583
  - Bernoulli trials (*see* Distribution, Bernoulli)
  - Beta function (*see also* Distribution, beta), 107
  - Bias, 303
  - Binomial coefficients, 15
  - Binomial Theorem, 90
  - Birnbaum's Theorem, 269
  - Borel field, 6
    - countable, 7
    - generated, 34
    - properties, 37
    - trivial, 6
    - uncountable, 7
  - Breakdown value, 229, 242
  - Carleman's Condition, 82
  - Central Limit Theorem (CLT), 216, 218, 240, 243
    - Demoivre–Laplace, 219
  - Characteristic function, 83–84
    - convergence, 84
    - use in CLT, 219
  - Chi squared test of independence, 398
  - Cochran's Theorem, 526, 535, 551
  - Coefficient of determination, 574
  - Complete class, 481
    - Bayes rule, 482
    - essentially, 481
    - hypothesis testing, 484
    - minimal, 503
  - Neyman–Pearson Lemma, 484
    - point estimation, 483
    - sufficient statistic, 483
  - Complete statistic (*see* Completeness)
  - Completeness, 260
    - ancillarity, 262, 282–283
    - Basu's Theorem, 262
    - best unbiased estimator, 320
  - Completeness (*continued*)
    - binomial, 260
    - exponential, 263–264, 282
    - exponential family, 263
    - independence, 262
    - Poisson, 281
    - sufficiency, 262
    - uniform, 261
  - Concave function, 182
  - Conditional probability, 18
    - multiplication rule, 21
    - properties, 42
  - Conditionality Principle, 268
    - interpretations, 272
  - Confidence coefficient, 404, 429
    - relation to risk, 471
  - Confidence interval (*see also* Credible set), 403
    - ANOVA, 517
    - approximate, 443
      - binomial, 441–442, 444
      - maximum likelihood, 441–442
      - negative binomial, 445–446
      - Poisson, 443–444
    - asymptotic (see Confidence interval, approximate)
    - binomial
      - relation to  $F$  distribution, 449
    - construction, 417–418, 420, 448
    - decision theoretic, 470
      - difficulty, 472
      - mean, 471
      - normal, 502
    - expected length
      - normal, 432
    - exponential, 419
      - LRT, 409
      - pivotal, 414–415
    - Fieller, 459
    - invariant
      - location, 454–455
      - scale, 455
    - length
      - gamma, 432
      - normal, 432
      - optimal, 430–431
      - relation to false coverage, 436
      - relation to HPD, 432
      - unimodal, 430–431
    - LRT
      - binomial, 447
    - negative binomial
      - relation to  $F$  distribution, 449
    - normal, 403, 406
      - pivotal, 415
    - one-sided, 403
      - binomial, 411–412
      - normal, 411
    - Poisson, 421
    - ratio of means, 459
    - RCB, 537
  - Confidence interval (*continued*)
    - regression, 574–575, 577
    - relation to set estimator, 404–405
    - shortest
      - binomial, 417, 448
    - UMA
      - length dominance, 453
      - normal, 453
    - UMA unbiased
      - normal, 454
    - unbiased, 435
    - uniform, 405
    - variance, 455
  - Confidence procedure (*see also* Confidence interval), 458
  - Confidence set (*see* Confidence interval)
  - Bayes
    - relation to invariant, 440
    - difficulty, 451
    - EIV, 592–593
    - frequentist–Bayesian
      - interpretations, 422–423
    - invariant, 427
      - constant coverage probability, 438
      - normal, 427–428
      - optimal, 437
      - relation to Bayes, 440
      - relation to tests, 427
      - size, 438–439
    - multivariate, 500
    - pivotal
      - EIV, 593–594
    - randomized, 458
    - relation to hypothesis test, 407–408
    - relation to sufficient statistic, 458
    - size, 429
    - UMA, 433–434
      - normal, 435
      - relation to UMP test, 434
    - UMA unbiased, 435
      - normal, 436
      - relation to UMP unbiased test, 435–436
    - unbiased, 435
  - Conjugate family, 299
  - Consistency (*see* Point estimator, consistent), 214
  - Contingency table, 398
  - Continuity correction, 106–107
  - Contrasts, 513–514
  - Control problem, 492
  - Convergence
    - almost sure, 214–215
    - in distribution, 216, 220, 383
    - in probability, 213, 238, 323
    - strong law, 216

- Convergence (*continued*)**  
 weak law, 214  
 with probability one, 214
- Convex function**, 182–183
- Convolution**, 209–210
- Correlation** (*see also Covariance*),  
 161, 163  
 coefficient, 161
- Counting formulas**, 16
- Counting methods**, 14
- Covariance** (*see also Correlation*),  
 160  
 independence, 162  
 linear functions, 602
- Coverage probability**, 404  
 frequentist–Bayesian  
 interpretation, 425–426  
 relation to risk, 471
- Cox's paradox**, 454
- Cramér–Rao Lower Bound**, 308  
 necessary condition, 315
- Credible set**, 423  
 normal, 425  
 Poisson, 423  
 relation to confidence set  
 normal, 425–426, 451
- Critical region** (*see Hypothesis test, rejection region*)
- Cross product set**, 143
- Cumulative distribution function**  
 (cdf), 29  
 independence, 191  
 joint, 136–137  
 mixed, 31  
 monotone transformation, 49  
 properties, 30
- Cutoff point**, 362–363
- Decision rule**, 462  
 admissible (*see Admissible decision rule*)  
 behavioral, 392  
 comparison, 479  
 equalizer, 493  
 inadmissible (*see Inadmissible decision rule*)  
 invariant, 492  
 randomized, 392, 483
- Decision theory**, 461
- Degrees of freedom**, 102
- DeMorgan's Laws** (*see Set operation*), 6, 37
- Determinant**, 148, 177
- Disjoint sets**, 5  
 independence, 42
- Distribution**  
 Bernoulli, 89  
 beta, 107  
 product, 148  
 relation to binomial, 82  
 relation to  $F$ , 228
- Distribution** (*continued*)  
 relation to gamma, 193  
 beta-binomial, 195  
 binomial, 46, 89  
 exponential family, 113  
 mean, 55  
 mgf, 63  
 recursion relation, 94  
 relation to beta, 82  
 relation to multinomial, 173  
 relation to negative binomial,  
 123  
 relation to Poisson, 192  
 variance, 60
- bivariate normal, 167, 196–197
- bivariate  $t$ , 571
- Cauchy**, 109, 126  
 generation, 78  
 mean, 55  
 relation to normal, 152, 194
- chi squared, 102  
 properties, 222  
 relation to negative binomial,  
 82  
 relation to normal, 52
- discrete uniform, 85
- double exponential, 112
- exponential, 102  
 hazard function, 125  
 mean, 54  
 memoryless, 102  
 recursion relation, 71  
 relation to gamma, 195  
 relation to uniform, 49  
 variance, 59
- exponential family (*see Exponential family*)
- extreme value, 124
- $F$ , 227  
 reciprocal, 228  
 relation to beta, 228  
 relation to  $t$ , 228  
 folded normal, 124
- gamma, 100  
 mgf, 62  
 moments, 123  
 recursion relation, 186–187  
 relation to beta, 193  
 relation to exponential, 195  
 relation to inverted gamma, 50  
 relation to negative binomial,  
 123  
 relation to Poisson, 101, 123  
 sum, 175
- Gaussian, 103
- geometric, 98  
 cdf, 31  
 memoryless, 98  
 pmf, 34
- Gumbel, 124
- hypergeometric, 86
- Distribution** (*continued*)  
 inverted gamma, 335  
 relation to gamma, 50  
 location family (*see Location-scale family*)  
 location-scale family (*see Location-scale family*)  
 logarithmic series, 123  
 logistic  
 cdf, 31  
 hazard function, 125  
 pdf, 36
- lognormal, 110  
 mgf, 81  
 moments, 64
- Maxwell, 124
- mixture, 155  
 Bernoulli–beta, 157, 159  
 binomial–beta, 195  
 binomial–negative binomial,  
 156  
 binomial–Poisson, 153  
 binomial–Poisson–exponential,  
 155  
 chi squared–Poisson, 157  
 exponential, 195  
 logarithmic series–Poisson, 195  
 normal–gamma, 241  
 Poisson–gamma, 157, 194  
 $t$ – $F$ , 241
- multinomial, 172  
 relation to binomial, 173
- negative binomial, 95, 195  
 quadratic variance, 97  
 relation to binomial, 123  
 relation to chi squared, 82  
 relation to gamma, 123  
 relation to Poisson, 97
- noncentral chi squared, 157
- normal, 103  
 bivariate, 167, 196–197  
 conditional, 168, 196  
 exponential family, 113–115  
 linear combination, 175–176,  
 223  
 ratio, 152  
 recursion relation, 72  
 relation to Cauchy, 152, 194  
 relation to chi squared, 52  
 relation to  $t$ , 241  
 standard, 103  
 sum, 145–146, 149
- Pareto, 124
- Poisson, 92  
 mgf, 67  
 postulates, 126–127  
 recursion relation, 94, 186  
 relation to binomial, 192  
 relation to gamma, 101, 123  
 relation to negative binomial,  
 97, 123

- Distribution (*continued*)
  - sum, 147
  - Rayleigh, 124
  - scale family (*see* Location-scale family)
  - $t$ , 225–226
    - noncentral, 376, 391–392
    - relation to  $F$ , 228
    - relation to normal, 241
  - truncated, 123
  - uniform, 99
    - order statistic, 233
    - relation to exponential, 49
  - Weibull, 103, 124
    - hazard function, 125
- Dot notation, 516
- Empirical Bayes (*see* Point estimator, empirical Bayes)
- Empty set, 3
- Equalizer rule (*see* Minimax decision rule, equalizer)
- Equivariance, 278
- Errors in variables regression (EIV), 581
  - confidence set, 592–593
  - consistency, 589–590, 592, 604
  - functional relationship, 583, 603
  - identifiability, 584, 591
  - least squares (*see* Least squares estimate, EIV)
- MLE
  - functional, 589
  - relation to least squares, 600
  - structural, 590–591
- relation to RCB, 590
  - structural relationship, 583, 603
- Estimator (*see* Point estimator, Set estimator, etc.)
- Event, 2
- Evidence, 267
- Exchangeable, 236
- Expected value, 54
  - bivariate, 130–131, 134
  - properties, 191
  - conditional, 139–140
    - minimizing, 192
  - independence, 144, 174–175
  - iterated, 154
  - joint, 169
  - of maxima and minima, 84
  - minimizing property, 57
  - properties, 56
  - sum, 207
  - two calculation methods, 57
- Exponential family, 112, 125
  - moments, 125
  - natural parameter, 115
  - sums, 212
- Factorial, 15
- Factorization Theorem, 250
- Fiducial inference, 266
- Fieller's Theorem, 459
- Fisher information (*see* Information number)
- Fixed factor, 529–530
- Forced binary choice, 199
- Fundamental Theorem of Calculus, 35, 137
- Fundamental Theorem of Counting, 14
- Game theory, 507
- Gamma function (*See also* Distribution, gamma), 100
- Gauss–Markov Theorem, 564
- Generating function
  - characteristic (*see* Characteristic function)
  - cumulant, 83
  - factorial moment, 83
  - moment (*see* Moment generating function)
  - probability, 83
- Geometric mean, 183
- Geometric series
  - differentiation, 74
  - partial sum, 31
- Gosset, W. S., 225
- Grand mean, 516
- Group
  - location, 276–278, 301, 306, 336, 393, 427, 438, 440, 454–455, 493
  - location–scale, 302, 336, 428
  - scale, 302, 336, 353, 428, 438, 455, 494
  - of transformations, 274–275
- Harmonic mean, 183
- Hazard function, 103, 124–125
- Highest posterior density (HPD) region, 424
- Homoscedasticity, 512
- Hunt–Stein Theorem, 507–508
- Hwang's Lemma, 189
- Hypothesis, 345
  - alternative, 345
  - composite, 368
  - null, 345
  - one-sided, 370
  - research, 362
  - simple, 366, 368
  - two-sided, 370
- Hypothesis test, 345
  - abuse, 388
  - acceptance region, 346
  - asymptotic, 399
    - Bernoulli, 384, 385
  - Bayes, 401, 402
    - generalized zero–one loss, 477
- Hypothesis test (*continued*)
  - normal, 355, 478
  - p-value, 388–389
  - contingency table, 398
  - decision theoretic, 467
    - mean, 468, 469
  - intersection–union, 357
    - Bernoulli, 358, 380
    - level, 379
    - normal, 358, 380
  - invariant, 352, 494
    - ANOVA, 545
    - Bernoulli, 352
    - decision theoretic, 505–506
    - location family, 393
    - normal, 353, 363
    - RCB, 545
    - relation to confidence set, 428
    - relation to unbiased, 402
    - $t$  test (*see*  $t$  test)
  - level of test, 361
  - locally most powerful, 376
    - $t$ , 377–378
  - LRT (*see* Likelihood ratio test)
  - McNemar's test, 399–400
  - power, 359
    - binomial, 359
    - exponential family, 393
    - monotone, 401
    - normal, 360
    - relation to risk, 468
  - randomized, 369, 392, 400
  - rejection region, 346
  - relation to confidence set, 407–408
  - sample size
    - normal, 360–361
  - size of test, 361
    - asymptotic, 382
  - sufficiency, 367, 400
  - $t$  test (*see*  $t$  test)
  - type I error, 358, 361–362
  - type II error, 358, 361–362
  - UMP, 365
    - binomial, 368
    - normal, 369, 371, 394
    - randomized, 374
    - relation to decision theory, 479
    - relation to UMA set, 434
    - sufficiency, 368
    - uniform, 391
  - UMP unbiased
    - normal, 394
    - relation to UMA unbiased set, 435–436
  - unbiased, 364
    - normal, 374
    - relation to invariant, 402
  - union–intersection, 356
    - level, 378
    - normal, 356, 363

- Hypothesis test (continued)**  
size, 398
- Identifiability**, 511  
ANOVA, 511  
EIV, 584, 591  
RCB, 530
- Identity**  
beta, 199  
binomial coefficients, 39, 87  
chi squared, 188  
conditional covariance, 198  
covariance, 602  
expectation  
continuous, 78  
discrete, 79  
gamma, 199  
integration–by–parts, 187  
negative binomial, 189  
Poisson, 189  
Iid, 201  
Inadmissible decision rule, 480  
sample mean, 496  
squared error loss  
variance, 481
- Independence**  
completeness, 262  
covariance, 162  
disjoint, 42  
events, 22  
mutual, 25  
properties, 23  
expected value, 144, 174–175  
random variables, 141–142,  
150–151, 174, 201  
sums, 145
- Indicator function**, 114
- Inequality**  
Bonferroni's, 11, 13, 416  
ANOVA, 523  
regression, 577  
Bonferroni - Boole similarity, 13  
Boole, 11  
Cauchy–Schwarz, 164, 180  
Chebychev's, 184–185, 200, 214,  
245, 323–324  
covariance, 184  
Hölder's, 179, 559  
Jensen's, 182  
Liapounov's, 180  
logarithm, 39  
Markov's, 200  
means, 183  
mgf, 81  
Minkowski's, 180–181  
normal, 185–186, 198  
triangle, 181, 198
- Information number, 311, 326  
Instrumental variable, 604  
**Interchange**  
differentiation and integration
- Interchange (continued)**  
exponential family, 125, 393  
differentiation and summation,  
74–75  
integration and summation, 76
- Interval estimator** (*see* Confidence set; Confidence interval;  
Confidence procedure)
- Intraclass correlation**, 531–532
- Invariance** (*see also* Point estimator,  
invariant; Confidence interval,  
invariant; Decision rule,  
invariant; etc.)  
binomial, 274–275  
distribution, 275  
formal, 273  
location, 276  
measurement, 273, 279  
location, 276  
normal, 275–276, 278  
relation to minimaxity, 507–508  
relation to sufficiency, 278  
variance estimate, 277
- Invariance Principle**, 274  
interpretations, 278
- Invariant interval** (*see* Confidence interval, invariant)  
Inverse binomial sampling, 97
- Jackknife estimator** (*see* Point estimator, jackknife)
- Jacobian**, 148
- Karlin–Rubin Theorem, 370  
**Kernel**, 63
- Kolmogorov's Axioms, 7  
conditional probability, 19  
induced probability, 27, 29, 46
- Kruskal's proof, 222, 243  
**Kurtosis**, 80
- Laplace transform, 66, 263  
**Latent variables**, 581  
**Law of Large Numbers**  
Strong, 216  
Weak, 214
- Least squares estimate**, 559  
consistency  
EIV, 604  
EIV, 584  
intercept, 586  
relation to MLE, 600  
slope, 586  
of intercept, 559  
normal equations, 595  
of slope, 559
- Lebesgue's Dominated Convergence Theorem**, 70
- Lehmann–Scheffé Theorem**, 344  
**Leibnitz's Rule**, 69  
**Lévy Theorem**, 243
- Likelihood function**, 265  
induced, 293  
log likelihood, 291, 332  
negative binomial, 265  
sufficiency, 283
- Likelihood Principle**, 266  
binomial and negative binomial,  
271  
Birnbaum's Theorem, 269  
formal, 267, 269  
interpretations, 272  
sample size, 282
- Likelihood ratio test (LRT)**, 346  
ANOVA, 545  
asymptotic, 381  
beta, 388  
exponential, 348, 350, 362, 386  
MLE, 347  
multinomial, 382, 399–400  
Neyman–Pearson Lemma, 390  
normal, 347, 350, 362, 394  
nuisance parameter, 350  
two-sample, 396  
Pareto, 386  
sufficient statistic, 349  
*t* test (*see t* test)
- Lindeberg–Feller Condition**, 243
- Linear combination**, 514
- Location family** (*see* Location–scale family)  
pivotal quantity, 413
- Location–scale family**, 116  
location parameter, 116, 119  
pivotal quantity, 413  
scale parameter, 118–119  
standard pdf, 116, 119  
stochastic order, 391
- Log linear model**, 383
- Loss function**, 462  
absolute error, 464  
asymmetric, 502  
generalized zero–one, 468  
hypothesis testing, 468  
interval estimation, 470  
likelihood, 467  
LINEX, 502  
point estimation, 464  
squared error, 464  
zero–one, 468
- Mapping**, 45  
inverse, 45
- Maximum likelihood estimator (MLE)**, 289  
asymptotic efficiency, 325  
Bernoulli, 291  
approximate variance, 326–327  
binomial, 292, 297  
bivariate normal, 333  
calculus, 295–296  
consistency, 325

- Maximum likelihood estimator (MLE) (continued)**
- EIV, 589–590, 592
  - instability, 297
  - invariance, 293–294
  - LRT, 347
  - normal, 290, 295–296
  - regression (*see* Regression, MLE)
  - relation to method of moments, 342
  - Maximum probability estimator, 598
  - Maximum of quadratic functions, 519, 545, 562, 579, 599
  - Mean (*see* Expected value; Sample mean)
  - Mean squared error, 303, 465
  - Measurement error (*see* Errors in variables regression)
  - Median, 79
    - minimizing property, 79
  - Meré dice problem, 23, 91
  - Metric, 600
  - Minimal sufficient statistic (*see* Sufficient statistic, minimal)
  - Minimax decision rule, 487, 490, 504
    - admissible
      - unique, 491
    - Bayes, 488
      - equalizer, 488–489
    - binomial, 489
    - equalizer, 488–489
    - limit of Bayes, 505
    - relation to invariance, 507–508
    - squared error loss
      - binomial, 491–492
      - bounded mean, 490
      - mean, 491
    - Stein estimation, 496
  - Moment generating function (mgf), 61
    - convergence, 66
    - identically distributed, 65
    - independent random variables, 145
    - properties, 68
    - sample mean, 209
    - sum, 175
    - use in CLT, 217
  - Moments, 58
    - nonuniqueness, 64, 81
    - uniqueness, 82–83
  - Monotone likelihood ratio (MLR), 370
    - binomial, 390
    - Cauchy, 390–391
    - exponential family, 390
    - logistic, 390
    - noncentral  $t$ , 391–392
    - normal, 390
    - Poisson, 390
  - Monotone likelihood ratio (MLR) (*continued*)
    - stochastic order, 390–391
    - Multinomial coefficient, 173
    - Multinomial Theorem, 173
    - Multiple comparisons, 551
      - comparisonwise error rate, 546
      - experimentwise error rate, 546
      - protected LSD, 546
      - Tukey's  $Q$  method, 551
    - Multivariate estimation, 495
    - Mutually exclusive sets (*see* Disjoint sets)
    - Necessary statistic, 283
    - Neighborhood, 61
    - Neyman–Pearson Lemma, 366, 372
      - generalized, 372–373, 377
      - LRT, 390
    - Neyman shortest, 436
    - Normal equations, 595
    - Nuisance parameter, 301, 350, 354
    - Odds ratio, 327
    - One-to-one transformation, 48, 148
    - Onto transformation, 48, 148
    - Order statistics, 229
      - cdf, 231–232
      - joint cdf, 234
      - joint pdf, 234
      - pdf, 232
      - pmf, 231
      - spacings, 178
      - sufficient, 280
    - Orthogonal least squares, 584
    - Overparameterized model
      - ANOVA, 510–511
      - RCB, 528
    - p-value, 364
      - relation to Bayes test, 401–402
    - Pairwise disjoint sets, 5
    - Parallel system, 197
    - Parameter, 46
      - Parameter space, 461
    - Partition, 5, 21, 246, 255
    - Percentile, 230
    - Pitman Estimator (*see* Point estimator, invariant), 337
    - Pivotal quantity, 413
      - gamma, 414
      - general form, 414
      - location family, 413
      - location–scale family, 413
      - scale family, 413
    - Point estimate, 284
    - Point estimator, 284
      - approximation, 330
      - Bayes (*see* Bayes estimator)
      - consistent, 322
    - Point estimator (*continued*)
      - maximum likelihood estimator, 325
      - normal, 323
      - decision theoretic, 464
      - binomial, 465
      - variance, 465–466
      - empirical Bayes
        - multivariate, 506
        - normal, 335
      - invariant, 300, 494
        - formal, 300
        - location, 301, 336
        - location–scale, 301–302, 336
        - mean squared error, 306
        - measurement, 300
        - Pitman, 337
        - jackknife, 341
        - linear, 338, 561
      - maximum likelihood (*see* Maximum likelihood estimator)
      - mean squared error
        - Bernoulli, 305
        - variance, 304
      - method of moments, 285
      - binomial, 286
      - bivariate normal, 333
      - chi squared, 287
      - normal, 286
      - relation to MLE, 342
      - ratio, 331
      - sufficiency, 316
      - unbiased (*see* Unbiased estimator)
    - Poisson postulates, 126–127
    - Positive part, 497
    - Posterior distribution, 298
    - Power (*see* Hypothesis test, power)
    - Pratt's Theorem, 436
    - Prediction interval, 576–577
    - Prior distribution, 297, 463
      - conjugate
        - gamma, 335
        - Poisson, 335
      - decision theoretic, 473
      - improper, 439
      - invariant
        - location, 439–440
        - scale, 456
      - least favorable, 488
      - noninformative
        - location, 439
        - scale, 456
    - Probability density function (pdf), 35
      - bivariate transformation, 148
      - conditional, 139, 170
      - construction, 37
      - even, 79
      - joint, 133, 169
      - marginal, 134, 169
      - monotone transformation, 50

- Probability density function (pdf)  
*(continued)*  
 multivariate transformation, 177  
 piecewise monotone transformation, 51  
 properties, 36  
 sample, 202  
 symmetric, 80  
 truncated, 43  
 unimodal, 80
- Probability function, 7  
 induced, 27  
 properties, 9–11
- Probability integral transformation, 52  
 confidence interval, 418  
 discrete, 78, 420–421
- Probability of false coverage, 433
- Probability mass function (pmf), 34  
 conditional, 137, 170  
 construction, 37  
 joint, 169  
 bivariate, 129  
 marginal, 132, 169  
 properties, 36  
 sample, 202
- Quartile, 230
- Random factor, 529–530
- Random number generation, 54, 194, 239
- Random variable, 26  
 absolutely continuous, 37  
 censored, 193  
 continuous, 33  
 discrete, 33  
 identically distributed, 33  
 moments, 65
- Random vector, 128  
 continuous, 133  
 discrete, 129
- Randomized complete block design (RCB), 528  
 ANOVA table, 540  
 assumptions, 530–531  
 blocking factor, 528  
 complete, 528, 542  
 confidence interval, 537  
 contrasts, 532, 535, 537  
 $F$  test, 538–539  
 hierarchical model, 530  
 identifiability, 530  
 incomplete, 542  
 intraclass correlation, 531–532  
 overparameterized model, 528  
 randomized, 529  
 relation to ANOVA, 541  
 relation to EIV, 590  
 residuals, 533, 535  
 Scheffé's method, 539
- Randomized complete block design (RCB) *(continued)*  
 sum of squares, 539  
 $t$  test, 537  
 testing blocks, 542  
 union-intersection test, 538  
 variance estimate, 534
- Randomized response, 334
- Randomized test (*see* Hypothesis test, randomized)
- Randomly stopped sum, 195
- Rao–Blackwell Theorem, 316  
 generalization, 483
- Regression, 554  
 ANOVA table, 572–573  
 bivariate normal, 566  
 BLUE of intercept, 564  
 BLUE of slope, 563  
 coefficient of determination, 574  
 conditional expectation, 192, 555  
 conditional normal, 565  
 confidence band, 578, 603  
 confidence interval, 574–575  
 data fitting, 556, 560  
 dependent variable, 555  
 design, 564
- EIV (*see* Errors in variables regression)  
 extrapolation, 580  
 independent variable, 555  
 inference, 556  
 least absolute deviation, 595  
 least squares (*see* Least squares estimate)  
 linear, 555, 564  
 linear in  $x$ , 555–556  
 maximum probability estimator, 598
- MLE, 567  
 biased, 567  
 sampling distribution, 569
- population, 554  
 prediction interval, 576, 577  
 predictor variable, 555  
 residuals, 567  
 response variable, 555  
 $r^2$ , 574  
 sum of squares, 557, 573  
 $t$  statistic, 571  
 $t$  test, 572, 574  
 toward the mean, 555  
 zero intercept, 334
- Risk function, 462  
 comparison, 472, 479
- Sample, 201  
 exponential family, 212  
 finite population, 203  
 infinite population, 203  
 mean, 205–206  
 Cauchy, 210
- Sample (*continued*)  
 expected value, 208  
 location-scale, 211  
 mgf, 209  
 normal, 209, 220–221  
 pdf, 237  
 recursion relation, 238  
 trimmed, 242  
 median, 229  
 midrange, 235  
 pdf, pmf, 202  
 range, 229  
 simple random, 204  
 standard deviation, 206  
 variance, 206  
 bias, 244  
 expected value, 208  
 identity, 237–238  
 normal, 220–221  
 recursion relation, 238  
 with replacement, 203  
 without replacement, 203
- Sample range  
 uniform, 235
- Sample size  
 hypothesis test—normal, 360–361
- Sample space, 1, 461
- Sampling distribution, 205
- Satterthwaite approximation, 287, 396
- Scale family (*see* Location-scale family)  
 pivotal quantity, 413
- Scheffé's method, 523–524  
 ANOVA, 545  
 RCB, 539
- Selection and ranking procedures, 552
- Set estimator  
 relation to confidence interval, 404–405
- Set operations  
 associativity, 3  
 commutativity, 3  
 complementation, 3  
 DeMorgan's Laws, 3  
 Distributive Law, 3  
 identities, 37  
 intersection, 3  
 countable, 4  
 uncountable, 5  
 union, 2  
 countable, 4  
 uncountable, 5
- Sigma algebra (*see* Borel field)
- Skewness, 80
- Slutsky's Theorem, 220, 384
- Standard deviation, 58
- Statistic, 205
- Stein estimation, 495, 506–507  
 ANOVA, 552

- Stein's Lemma, 187, 497  
 Stirling's formula, 39, 239  
 Stochastic order  
   decreasing, 126  
   greater, 43, 78  
   increasing, 126, 417  
     chi squared, 241  
      $F$ , 241  
   location-scale family, 391  
   MLR, 390–391  
   power, 401  
 Structural inference, 413  
 Sufficiency  
   relation to invariance, 278  
 Sufficiency Principle, 247  
   formal, 268  
   interpretations, 272  
 Sufficient statistic, 247  
   ANOVA, 543  
   Bernoulli, 249  
   discrete uniform, 252  
   exponential family, 254  
 Factorization Theorem, 250  
   minimal, 254–255  
     characterization, 255  
     necessary, 283  
     normal, 256  
     uniform, 256, 280  
   normal, 249, 251, 253, 255  
   relation to confidence set, 458  
 Sum of squares  
   residual, 557  
 Support, 48  
*t* test (*see also* Likelihood ratio test (LRT), *t* test)  
   ANOVA, 516–517
- t* test (*continued*)  
   approximate, 396–397  
   invariant, 354, 363  
   LRT, 351  
     paired, 395  
     two-sided, 357  
   paired, 395  
   RCB, 537  
   regression, 574  
   two-sample, 396  
   two-sided, 397  
   UMP  
     invariant, 376  
     unbiased, 376, 395  
   union-intersection, 363
- Taylor's Theorem, 328  
   use in CLT, 217
- Test (*see* Hypothesis test)
- Test function, 351, 366, 373, 400
- Test statistic, 346
- Threshold parameter, 118
- Total least squares, 584
- Transformations  
   monotone, 48
- Transitivity, 438
- Trimmed mean, 242
- Type I error (*see* Hypothesis test, type I error)
- Type II error (*see* Hypothesis test, type II error)
- Unbiased estimator, 208, 303  
   best, 307  
     binomial, 321  
     characterization, 318  
     completeness, 320  
     Poisson, 312
- Unbiased estimator (*continued*)  
   sufficiency, 317  
   uniform, 320  
   unique, 317  
   linear, 338  
   best (*see* Regression, BLUE)  
   location, 336  
   normal, 304  
     standard deviation, 339  
   Poisson, 308  
   relation to Bayes estimator, 343  
   uniform, 312–313, 319  
   variance bound, 314  
   zero, 318
- Unbiased test (*see* Hypothesis test, unbiased)
- Uniformly minimum variance  
   unbiased estimator (*see* Unbiased estimator, best)
- Uniformly most accurate (UMA)  
   (*see* Confidence set, UMA)
- Uniformly most powerful test (*see* Hypothesis test, UMP)
- Union-intersection test  
   ANOVA, 514  
   RCB, 538
- Utility function, 462
- Variance, 58  
   conditional, 140, 158  
   properties, 59  
   quadratic function of mean, 97  
   sum, 162, 207, 237  
   Taylor approximation, 329
- Venn diagram, 4
- Warden problem, 20, 42

