Osmania University Library

Call No. 519.9 Accession No. 6760

Author H71I

Title

This book should be returned on or before the date last marked below.

Robert V. Hogg
Allen T. Craig
THE UNIVERSITY OF IOWA

Introduction to Mathematical Statistics *3rd Edition*

Macmillan Publishing Co., Inc.
NEW YORK
Collier Macmillan Publishers
LONDON

© Copyright, Macmillan Publishing Co., Inc., 1970

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without permission in writing from the Publisher.

PRINTING 12 YEAR 67890

Earlier editions copyright 1958 and 1959 and © copyright 1965 by Macmillan Publishing Co., Inc.

Library of Congress catalog card number: 74–77968

MACMILLAN PUBLISHING CO., INC.
866 THIRD AVENUE, NEW YORK, NEW YORK 10022

COLLIER-MACMILLAN CANADA, LTD., TORONTO, ONTARIO

Printed in the United States of America

Preface

We are much indebted to our colleagues throughout the country who have so generously provided us with suggestions on both the order of presentation and the kind of material to be included in this edition of our *Introduction to Mathematical Statistics*. Thus you will find, among other things, that there is a greatly expanded discussion of random variables, that the distribution of the order statistics has been incorporated with the work on transformations which are not one-to-one, that an early proof of the independence of $\overline{X}$ and S^2 (when sampling from a normal distribution) has been introduced, and that limiting distributions are now found in Chapter 5. Thus essentially all of the distribution theory which is needed for the remainder of the book is found in the first five chapters. At the request of a number of teachers we have added some new exercises and a chapter on nonparametric methods. Although a number of other worthwhile suggestions were made, we have had to forego their inclusion in order to keep the book in reasonable bounds.

We hope, and believe, that this edition is written at the same mathematical level as were its predecessors, so that a good grounding in the calculus should be adequate mathematical preparation. This remark does not apply to the last chapter, which does require an elementary knowledge of matrix algebra.

We are obligated to Catherine M. Thompson and Maxine Merrington and to Professor E. S. Pearson for permission to include Tables II and V, which are abridgments and adaptations of tables published in *Biometrika*. We wish to thank Oliver & Boyd, Ltd., Edinburgh, for permission to include Table IV, which is an abridgment and adaptation of Table III from the book *Statistical Tables for Biological, Agricultural, and Medical Research* by the late Professor Sir Ronald A. Fisher, Cambridge, and Dr. Frank Yates, Rothamsted.

R. V. H.

Iowa City, Iowa

A. T. C.

Contents

Chapter 1
Distributions of Random Variables **1**

1.1 Introduction *1*
1.2 Algebra of Sets *3*
1.3 Set Functions *7*
1.4 The Probability Set Function *11*
1.5 Random Variables *14*
1.6 The Probability Density Function *21*
1.7 The Distribution Function *29*
1.8 Certain Probability Models *36*
1.9 Mathematical Expectation *41*
1.10 Some Special Mathematical Expectations *46*
1.11 Chebyshev's Inequality *54*

Chapter 2
Conditional Probability and Stochastic Independence **57**

2.1 Conditional Probability *57*
2.2 Marginal and Conditional Distributions
2.3 The Correlation Coefficient *69*
2.4 Stochastic Independence *76*

Chapter 3
Some Special Distributions **86**

3.1 The Binomial Distribution *86*
3.2 The Poisson Distribution *94*
3.3 The Gamma and Chi-Square Distributions *99*
3.4 The Normal Distribution *104*
3.5 The Bivariate Normal Distribution *111*

Chapter 4
Distributions of Functions of Random Variables **116**

4.1 Sampling Theory *116*
4.2 Transformations of Variables of the Discrete Type *121*
4.3 Transformations of Variables of the Continuous Type *125*
4.4 The t and F Distributions *135*
4.5 Extensions of the Change of Variable Technique *139*
4.6 Distributions of Order Statistics *145*
4.7 The Moment-Generating Function Technique *154*
4.8 The Distributions of $\bar{X}$ and nS^2/σ^2 *163*
4.9 Expectations of Functions of Random Variables *166*

Chapter 5
Limiting Distributions **171**

5.1 Limiting Distributions *171*
5.2 Stochastic Convergence *175*
5.3 Limiting Moment-Generating Functions *178*
5.4 The Central Limit Theorem *182*
5.5 Some Theorems on Limiting Distributions *186*

Chapter 6
Interval Estimation **190**

6.1 Random Intervals *190*
6.2 Confidence Intervals for Means *193*
6.3 Confidence Intervals for Differences of Means *199*
6.4 Confidence Intervals for Variances *202*
6.5 Bayesian Interval Estimates *207*

Chapter 7
Point Estimation and Sufficient Statistics **212**

7.1 The Problem of Point Estimation *212*
7.2 A Sufficient Statistic for a Parameter *215*
7.3 The Rao-Blackwell Theorem *223*
7.4 Completeness *226*
7.5 Uniqueness *229*
7.6 The Exponential Class of Probability Density Functions *231*
7.7 Functions of a Parameter *234*
7.8 The Case of Several Parameters *237*
7.9 Sufficiency, Completeness, and Stochastic Independence *243*

Chapter 8
Further Topics in Point Estimation **248**

8.1 The Rao-Cramér Inequality *248*
8.2 Maximum Likelihood Estimation of Parameters *254*
8.3 Decision Functions *258*
8.4 Bayesian Procedures *261*

Chapter 9
Statistical Hypotheses **265**

9.1 Some Examples and Definitions *265*
9.2 Certain Best Tests *272*
9.3 Uniformly Most Powerful Tests *280*
9.4 The Sequential Probability Ratio Test *285*
9.5 Minimax and Bayesian Tests *290*

Chapter 10
Other Statistical Tests **297**

10.1 Likelihood Ratio Tests *297*
10.2 Chi-Square Tests *308*
10.3 The Distributions of Certain Quadratic Forms *316*
10.4 A Test of the Equality of Several Means *321*
10.5 Noncentral χ^2 and Noncentral F *325*
10.6 The Analysis of Variance *327*
10.7 A Regression Problem *335*
10.8 A Test of Stochastic Independence *339*
10.9 Multiple Comparisons *343*

Chapter 11
Nonparametric Methods **348**

11.1 Confidence Intervals for Distribution Quantiles *348*
11.2 Tolerance Limits for Distributions *353*
11.3 The Sign Test *357*
11.4 A Test of Wilcoxon *359*
11.5 The Equality of Two Distributions *365*
11.6 The Mann-Whitney-Wilcoxon Test *371*
11.7 Distributions under Alternative Hypotheses *376*

Chapter 12
Further Normal Distribution Theory 379

12.1 The Multivariate Normal Distribution *379*
12.2 The Distributions of Certain Quadratic Forms *384*
12.3 The Independence of Certain Quadratic Forms *388*

Appendix A
References 395

Appendix B
Tables 398

Appendix C
Answers to Selected Exercises 404

Index 411

Introduction to Mathematical Statistics

Chapter I

Distributions of Random Variables

1.1 Introduction

Many kinds of investigations may be characterized in part by the fact that repeated experimentation, under essentially the same conditions, is more or less standard procedure. For instance, in medical research, interest may center on the effect of a drug that is to be administered; or an economist may be concerned with the prices of three specified commodities at various time intervals; or the agronomist may wish to study the effect that a chemical fertilizer has on the yield of a cereal grain. The only way in which an investigator can elicit information about any such phenomenon is to perform his experiment. Each experiment terminates with an *outcome*. But it is characteristic of these experiments that the outcome cannot be predicted with certainty prior to the performance of the experiment.

Suppose we have such an experiment, the outcome of which cannot be predicted with certainty, but the experiment is of such a nature that the collection of every possible outcome can be described prior to its performance. If this kind of experiment can be repeated under the same conditions, it is called a *random experiment*, and the collection of every possible outcome is called the experimental space or the *sample space*.

Example 1. In the toss of a coin, let the outcome tails be denoted by T and let the outcome heads be denoted by H. If we assume that the coin may be repeatedly tossed under the same conditions, then the toss of this coin is an example of a random experiment in which the outcome is one of

the two symbols T and H; that is, the sample space is the collection of these two symbols.

Example 2. In the cast of one red die and one white die, let the outcome be the ordered pair (number of spots up on the red die, number of spots up on the white die). If we assume that these two dice may be repeatedly cast under the same conditions, then the cast of this pair of dice is a random experiment and the sample space consists of the 36 ordered pairs $(1, 1), \ldots, (1, 6), (2, 1), \ldots, (2, 6), \ldots, (6, 6)$.

Let $\mathscr{C}$ denote a sample space, and let C represent a part of $\mathscr{C}$. If, upon the performance of the experiment, the outcome is in C, we shall say that the *event* C has occurred. Now conceive of our having made N repeated performances of the random experiment. Then we can count the number f of times (the frequency) that the event C actually occurred throughout the N performances. The ratio f/N is called the *relative frequency* of the event C in these N experiments. A relative frequency is usually quite erratic for small values of N, as you can discover by tossing a coin. But as N increases, experience indicates that relative frequencies tend to stabilize. This suggests that we associate with the event C a number, say p, that is equal or approximately equal to that number about which the relative frequency seems to stabilize. If we do this, then the number p can be interpreted as that number which, in future performances of the experiment, the relative frequency of the event C will either equal or approximate. Thus, although we *cannot* predict the outcome of a random experiment, we *can*, for a large value of N, predict approximately the relative frequency with which the outcome will be in C. The number p associated with the event C is given various names. Sometimes it is called the *probability* that the outcome of the random experiment is in C; sometimes it is called the *probability* of the event C; and sometimes it is called the *probability measure* of C. The context usually suggests an appropriate choice of terminology.

Example 3. Let $\mathscr{C}$ denote the sample space of Example 2 and let C be the collection of every ordered pair of $\mathscr{C}$ for which the sum of the pair is equal to seven. Thus, C is the collection (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), and (6, 1). Suppose the dice are cast $N = 400$ times and let f, the frequency of a sum of seven, be $f = 60$. Then the relative frequency with which the outcome was in C is $f/N = \frac{60}{400} = 0.15$. Thus we might associate with C a number p which is close to 0.15 and p would be called the probability of the event C.

The primary purpose of having a mathematical theory of statistics is to provide mathematical models for random experiments. Once a

model for such an experiment has been provided and the theory worked out in detail, the statistician may, within this framework, make inferences (that is, draw conclusions) about the random experiment. The construction of such a model requires a theory of probability. One of the more logically satisfying theories of probability is that based on the concepts of sets and functions of sets. These concepts are introduced in Sections 1.2 and 1.3.

EXERCISE

1.1. In each of the following random experiments describe the sample space $\mathscr{C}$. Use any experience you may have had (or use your intuition) to assign a value to the probability p of the event C in each of the following instances:

(a) The toss of an unbiased coin where the event C is tails.

(b) The cast of an honest die where the event C is a five or a six.

(c) The draw of a card from an ordinary deck of playing cards where the event C occurs if the card is a spade.

(d) The choice of a number on the interval zero to one where the event C occurs if the number is less than $\frac{1}{3}$.

(e) The choice of a point from a square with opposite vertices $(-1, -1)$ and $(1, 1)$ where the event C occurs if the sum of the coordinates of the point is less than $\frac{1}{2}$.

1.2 Algebra of Sets

The concept of a *set* or a *collection* of objects is usually left undefined. However, a particular set can be described so that there is no misunderstanding as to what collection of objects is under consideration. For example, the set of the first 10 positive integers is sufficiently well described to make clear that the numbers $\frac{3}{4}$ and 14 are not in the set, while the number 3 is in the set. If an object belongs to a set, it is said to be an *element* of the set. For example, if A denotes the set of real numbers x for which $0 \leq x \leq 1$, then $\frac{3}{4}$ is an element of the set A. The fact that $\frac{3}{4}$ is an element of the set A is indicated by writing $\frac{3}{4} \in A$. More generally, $a \in A$ means that a is an element of the set A.

The sets that concern us will frequently be *sets of numbers*. However, the language of sets of *points* proves somewhat more convenient than that of sets of numbers. Accordingly, we briefly indicate how we use this terminology. In analytic geometry considerable emphasis is placed on the fact that to each point on a line (on which an origin and a unit point have been selected) there corresponds one and only one number,

say x; and that to each number x there corresponds one and only one point on the line. This one-to-one correspondence between the numbers and points on a line enables us to speak, without misunderstanding, of the "point x" instead of the "number x." Furthermore, with a plane rectangular coordinate system and with x and y numbers, to each symbol (x, y) there corresponds one and only one point in the plane; and to each point in the plane there corresponds but one such symbol. Here again we may speak of the "point (x, y)," meaning the "ordered number pair x and y." This convenient language can be used when we have a rectangular coordinate system in a space of three or more dimensions. Thus, the "point $(x_1, x_2, \ldots, x_n)$" means the numbers $x_1, x_2, \ldots, x_n$ in the order stated. Accordingly, in describing our sets, we frequently speak of a set of points (a set whose elements are points), being careful, of course, to describe the set so as to avoid any ambiguity. The notation $A = \{x; 0 \leq x \leq 1\}$ is read "A is the one-dimensional set of points x for which $0 \leq x \leq 1$." Similarly, $A = \{(x, y); 0 \leq x \leq 1, 0 \leq y \leq 1\}$ can be read "A is the two-dimensional set of points (x, y) which are interior to, or on the boundary of, a square with opposite vertices at $(0, 0)$ and $(1, 1)$." We now give some definitions (together with illustrative examples) which lead to an elementary algebra of sets adequate for our purposes.

Definition 1. If each element of a set A_1 is also an element of set A_2, the set A_1 is called a *subset* of the set A_2. This is indicated by writing $A_1 \subset A_2$. If $A_1 \subset A_2$ and also $A_2 \subset A_1$, the two sets have the same elements, and this is indicated by writing $A_1 = A_2$.

Example 1. Let $A_1 = \{x; 0 \leq x \leq 1\}$ and $A_2 = \{x; -1 \leq x \leq 2\}$. Here the one-dimensional set A_1 is seen to be a subset of the one-dimensional set A_2; that is, $A_1 \subset A_2$. Subsequently, when the dimensionality of the set is clear, we shall not make specific reference to it.

Example 2. Let $A_1 = \{(x, y); 0 \leq x = y \leq 1\}$ and $A_2 = \{(x, y); 0 \leq x \leq 1, 0 \leq y \leq 1\}$. Since the elements of A_1 are the points on one diagonal of the square, then $A_1 \subset A_2$.

Definition 2. If a set A has no elements, A is called the *null set*. This is indicated by writing $A = 0$.

Definition 3. The set of all elements which belong to at least one of the sets A_1 and A_2 is called the *union* of A_1 and A_2. The union of A_1 and A_2 is indicated by writing $A_1 \cup A_2$. The union of several sets $A_1, A_2, A_3, \ldots$ is the set of all elements that belong to at least one of the several sets. This union is denoted by $A_1 \cup A_2 \cup A_3 \cup \cdots$ or by $A_1 \cup A_2 \cup \cdots \cup A_k$ if a finite number k of sets is involved.

Example 3. Let $A_1 = \{x; x = 0, 1, \ldots, 10\}$ and $A_2 = \{x; x = 8, 9, 10, 11, \text{ or } 11 < x \leq 12\}$. Then $A_1 \cup A_2 = \{x; x = 0, 1, \ldots, 8, 9, 10, 11, \text{ or } 11 < x \leq 12\} = \{x; x = 0, 1, \ldots, 8, 9, 10, \text{ or } 11 \leq x \leq 12\}$.

Example 4. Let A_1 and A_2 be defined as in Example 1. Then $A_1 \cup A_2 = A_2$.

Example 5. Let $A_2 = 0$. Then $A_1 \cup A_2 = A_1$ for every set A_1.

Example 6. For every set A, $A \cup A = A$.

Example 7. Let $A_k = \{x; 1/(k+1) \leq x \leq 1\}$, $k = 1, 2, 3, \ldots$. Then $A_1 \cup A_2 \cup A_3 \cup \cdots = \{x; 0 < x \leq 1\}$. Note that the number zero is not in this set since it is not in one of the sets $A_1, A_2, A_3, \ldots$.

Definition 4. The set of all elements which belong to each of the sets A_1 and A_2 is called the *intersection* of A_1 and A_2. The intersection of A_1 and A_2 is indicated by writing $A_1 \cap A_2$. The intersection of several sets $A_1, A_2, A_3, \ldots$ is the set of all elements that belong to each of the sets $A_1, A_2, A_3, \ldots$. This intersection is denoted by $A_1 \cap A_2 \cap A_3 \cap \cdots$ or by $A_1 \cap A_2 \cap \cdots \cap A_k$ if a finite number k of sets is involved.

Example 8. Let $A_1 = \{(x, y); (x, y) = (0, 0), (0, 1), (1, 1)\}$ and $A_2 = \{(x, y); (x, y) = (1, 1), (1, 2), (2, 1)\}$. Then $A_1 \cap A_2 = \{(x, y); (x, y) = (1, 1)\}$.

Example 9. Let $A_1 = \{(x, y); 0 \leq x + y \leq 1\}$ and $A_2 = \{(x, y); 1 < x + y\}$. Then A_1 and A_2 have no points in common and $A_1 \cap A_2 = 0$.

Example 10. For every set A, $A \cap A = A$ and $A \cap 0 = 0$.

Example 11. Let $A_k = \{x; 0 < x < 1/k\}$, $k = 1, 2, 3, \ldots$. Then $A_1 \cap A_2 \cap A_3 \cdots$ is the null set since there is no point which belongs to each of the sets $A_1, A_2, A_3, \ldots$.

Example 12. Let A_1 and A_2 represent the sets of points enclosed, respectively, by two intersecting circles. Then the sets $A_1 \cup A_2$ and $A_1 \cap A_2$ are represented, respectively, by the shaded regions in the *Venn diagrams* in Figure 1.1.

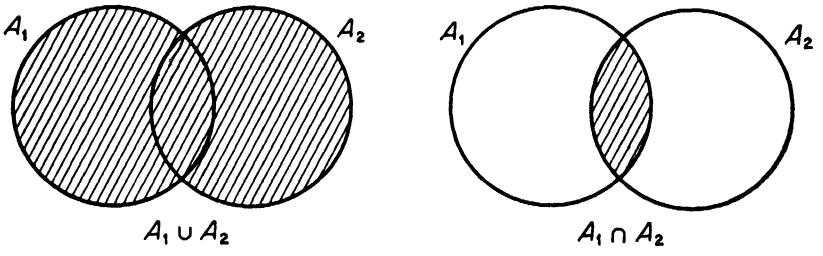


FIGURE 1.1

Example 13. Let A_1, A_2, and A_3 represent the sets of points enclosed, respectively, by three intersecting circles. Then the sets $(A_1 \cup A_2) \cap A_3$ and $(A_1 \cap A_2) \cup A_3$ are depicted in Figure 1.2.

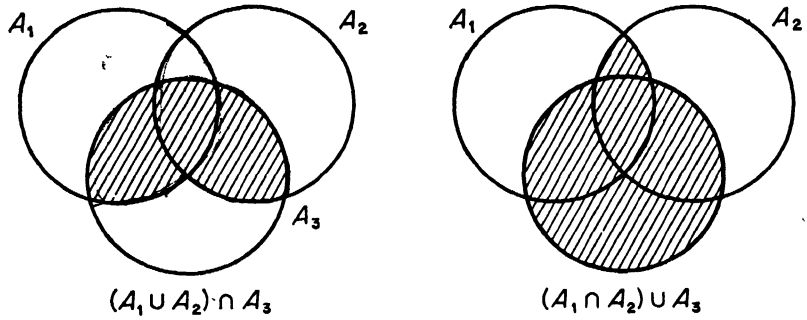


FIGURE 1.2

Definition 5. In certain discussions or considerations the totality of all elements that pertain to the discussion can be described. This set of all elements under consideration is given a special name. It is called the *space.*

We shall often denote spaces by capital script letters such as $\mathscr{A}$, $\mathscr{B}$, and $\mathscr{C}$.

Example 14. Let the number of heads, in tossing a coin four times, be denoted by x. Of necessity, the number of heads will be one of the numbers 0, 1, 2, 3, 4. Here, then, the space is the set $\mathscr{A} = \{x; x = 0, 1, 2, 3, 4\}$.

Example 15. Consider all nondegenerate rectangles of base x and height y. To be meaningful, both x and y must be positive. Thus the space is the set $\mathscr{A} = \{(x, y); x > 0, y > 0\}$.

Definition 6. Let $\mathscr{A}$ denote a space and let A be a subset of the set $\mathscr{A}$. The set which consists of all elements of $\mathscr{A}$ which are not elements of A is called the *complement* of A (actually, with respect to $\mathscr{A}$). The complement of A is denoted by A^*. In particular, $\mathscr{A}^* = 0$.

Example 16. Let $\mathscr{A}$ be defined as in Example 14, and let the set $A = \{x; x = 0, 1\}$. The complement of A (with respect to $\mathscr{A}$) is $A^* = \{x; x = 2, 3, 4\}$.

Example 17. Given $A \subset \mathscr{A}$. Then $A \cup A^* = \mathscr{A}$, $A \cap A^* = 0$, $A \cup \mathscr{A} = \mathscr{A}$, $A \cap \mathscr{A} = A$, and $(A^*)^* = A$.

EXERCISES

1.2. Find the union $A_1 \cup A_2$ and the intersection $A_1 \cap A_2$ of the two sets A_1 and A_2, where:

(a) $A_1 = \{x; x = 0, 1, 2\}$, $A_2 = \{x; x = 2, 3, 4\}$;
(b) $A_1 = \{x; 0 < x < 2\}$, $A_2 = \{x; 1 \leq x < 3\}$;
(c) $A_1 = \{(x, y); 0 < x < 2, 0 < y < 2\}$, $A_2 = \{(x, y); 1 < x < 3, 1 < y < 3\}$.

1.3. Find the complement A^* of the set A with respect to the space $\mathscr{A}$ if :
(a) $\mathscr{A} = \{x; 0 < x < 1\}$, $A = \{x; \frac{1}{8} \leq x < 1\}$;
(b) $\mathscr{A} = \{(x, y, z); x^2 + y^2 + z^2 \leq 1\}$, $A = \{(x, y, z); x^2 + y^2 + z^2 = 1\}$;
(c) $\mathscr{A} = \{(x, y); |x| + |y| \leq 2\}$, $A = \{(x, y); x^2 + y^2 < 2\}$.

1.4. List all possible arrangements of the four letters m, a, r, and y. Let A_1 be the collection of the arrangements in which y is in the last position. Let A_2 be the collection of the arrangements in which m is in the first position. Find the union and the intersection of A_1 and A_2.

1.5. By use of Venn diagrams compare the following sets:
(a) $A_1 \cap (A_2 \cup A_3)$ and $(A_1 \cap A_2) \cup (A_1 \cap A_3)$;
(b) $A_1 \cup (A_2 \cap A_3)$ and $(A_1 \cup A_2) \cap (A_1 \cup A_3)$;
(c) $(A_1 \cup A_2)^*$ and $A_1^* \cap A_2^*$;
(d) $(A_1 \cap A_2)^*$ and $A_1^* \cup A_2^*$.

1.6. If $A_1, A_2, A_3, \ldots$ are sets such that $A_k \subset A_{k+1}$, $k = 1, 2, 3, \ldots$, the $\lim_{k\to\infty} A_k$ is defined as the union $A_1 \cup A_2 \cup A_3 \cup \cdots$. Find the $\lim_{k\to\infty} A_k$ if:
(a) $A_k = \{x; 1/k \leq x \leq 3 - 1/k\}$, $k = 1, 2, 3, \ldots$;
(b) $A_k = \{(x, y); 1/k \leq x^2 + y^2 \leq 4 - 1/k\}$, $k = 1, 2, 3, \ldots$.

1.7. If $A_1, A_2, A_3, \ldots$ are sets such that $A_k \supset A_{k+1}$, $k = 1, 2, 3, \ldots$, the $\lim_{k\to\infty} A_k$ is defined as the intersection $A_1 \cap A_2 \cap A_3 \cap \cdots$. Find the $\lim_{k\to\infty} A_k$ if:
(a) $A_k = \{x; 2 - 1/k < x \leq 2\}$, $k = 1, 2, 3, \ldots$;
(b) $A_k = \{x; 2 < x \leq 2 + 1/k\}$, $k = 1, 2, 3, \ldots$;
(c) $A_k = \{(x, y); 0 \leq x^2 + y^2 \leq 1/k\}$, $k = 1, 2, 3, \ldots$.

1.3 Set Functions

In the calculus, functions like

$$f(x) = 2x, \qquad -\infty < x < \infty,$$

or

$$\begin{aligned} g(x, y) &= e^{-x-y}, \qquad 0 < x < \infty, \quad 0 < y < \infty, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

or possibly

$$\begin{aligned} h(x_1, x_2, \ldots, x_n) &= 3x_1x_2 \cdots x_n, \qquad 0 \leq x_i \leq 1, \quad i = 1, 2, \ldots, n, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

were of common occurrence. The value of $f(x)$ at the "point $x = 1$" is $f(1) = 2$; the value of $g(x, y)$ at the "point $(-1, 3)$" is $g(-1, 3) = 0$; the value of $h(x_1, x_2, \ldots, x_n)$ at the "point $(1, 1, \ldots, 1)$" is 3. Functions like these are called functions of a point or, more simply, *point functions* because they are evaluated (if they have a value) at a point in a space of indicated dimension.

There is no reason why, if they prove useful, we should not have functions that can be evaluated, not necessarily at a point, but for an entire set of points. Such functions are naturally called functions of a set or, more simply, *set functions*. We shall give some examples of set functions and shall evaluate them for certain simple sets.

Example 1. Let A be a set in one-dimensional space and let $Q(A)$ be equal to the number of points in A which correspond to positive integers. Then $Q(A)$ is a function of the set A. Thus, if $A = \{x; 0 < x < 5\}$, then $Q(A) = 4$; if $A = \{x; x = -2, -1\}$, then $Q(A) = 0$; if $A = \{x; -\infty < x < 6\}$, then $Q(A) = 5$.

Example 2. Let A be a set in two-dimensional space and let $Q(A)$ be the area of A, if A has a finite area; otherwise, let $Q(A)$ be undefined. Thus, if $A = \{(x, y); x^2 + y^2 \leq 1\}$, then $Q(A) = \pi$; if $A = \{(x, y); (x, y) = (0, 0), (1, 1), (0, 1)\}$, then $Q(A) = 0$; if $A = \{(x, y); 0 \leq x, 0 \leq y, x + y \leq 1\}$, then $Q(A) = \frac{1}{2}$.

Example 3. Let A be a set in three-dimensional space and let $Q(A)$ be the volume of A, if A has a finite volume; otherwise, let $Q(A)$ be undefined. Thus, if $A = \{(x, y, z); 0 \leq x \leq 2, 0 \leq y \leq 1, 0 \leq z \leq 3\}$, then $Q(A) = 6$; if $A = \{(x, y, z); x^2 + y^2 + z^2 \geq 1\}$, then $Q(A)$ is undefined.

At this point we introduce the following notations. The symbol

$$\int_A f(x)\, dx$$

will mean the ordinary (Riemann) integral of $f(x)$ over a prescribed one-dimensional set A; the symbol

$$\int_A \int g(x, y)\, dx\, dy$$

will mean the Riemann integral of $g(x, y)$ over a prescribed two-dimensional set A; and so on. To be sure, unless these sets A and these functions $f(x)$ and $g(x, y)$ are chosen with care, the integrals will frequently fail to exist. Similarly, the symbol

$$\sum_A f(x)$$

will mean the sum extended over all $x \in A$; the symbol

$$\sum_A \sum g(x, y)$$

will mean the sum extended over all $(x, y) \in A$; and so on.

Example 4. Let A be a set in one-dimensional space and let $Q(A) = \sum_A f(x)$, where

$$f(x) = (\tfrac{1}{2})^x, \qquad x = 1, 2, 3, \ldots,$$
$$= 0 \text{ elsewhere.}$$

If $A = \{x; 0 \leq x \leq 3\}$, then

$$Q(A) = \tfrac{1}{2} + (\tfrac{1}{2})^2 + (\tfrac{1}{2})^3 = \tfrac{7}{8}.$$

Example 5. Let $Q(A) = \sum_A f(x)$, where

$$f(x) = p^x(1 - p)^{1-x}, \qquad x = 0, 1,$$
$$= 0 \text{ elsewhere.}$$

If $A = \{x; x = 0\}$, then

$$Q(A) = \sum_{x=0}^{x=0} p^x(1 - p)^{1-x} = 1 - p;$$

if $A = \{x; 1 \leq x \leq 2\}$, then $Q(A) = f(1) = p$.

Example 6. Let A be a one-dimensional set and let

$$Q(A) = \int_A e^{-x}\, dx.$$

Thus, if $A = \{x; 0 \leq x < \infty\}$, then

$$Q(A) = \int_0^\infty e^{-x}\, dx = 1;$$

if $A = \{x; 1 \leq x \leq 2\}$, then

$$Q(A) = \int_1^2 e^{-x}\, dx = e^{-1} - e^{-2};$$

if $A_1 = \{x; 0 \leq x \leq 1\}$ and $A_2 = \{x; 1 < x \leq 3\}$, then

$$\begin{aligned} Q(A_1 \cup A_2) &= \int_0^3 e^{-x}\, dx \\ &= \int_0^1 e^{-x}\, dx + \int_1^3 e^{-x}\, dx \\ &= Q(A_1) + Q(A_2); \end{aligned}$$

if $A = A_1 \cup A_2$, where $A_1 = \{x; 0 \leq x \leq 2\}$ and $A_2 = \{x; 1 \leq x \leq 3\}$, then

$$\begin{aligned} Q(A) = Q(A_1 \cup A_2) &= \int_0^3 e^{-x}\, dx \\ &= \int_0^2 e^{-x}\, dx + \int_1^3 e^{-x}\, dx - \int_1^2 e^{-x}\, dx \\ &= Q(A_1) + Q(A_2) - Q(A_1 \cap A_2). \end{aligned}$$

Example 7. Let A be a set in n-dimensional space and let

$$Q(A) = \int\cdots\int_A dx_1\, dx_2 \cdots dx_n.$$

If $A = \{(x_1, x_2, \ldots, x_n);\ 0 \leq x_1 \leq x_2 \leq \cdots \leq x_n \leq 1\}$, then

$$Q(A) = \int_0^1 \int_0^{x_n} \cdots \int_0^{x_3} \int_0^{x_2} dx_1\, dx_2 \cdots dx_{n-1}\, dx_n$$

$$= \frac{1}{n!}, \qquad \text{where } n! = n(n-1)\cdots 3 \cdot 2 \cdot 1.$$

EXERCISES

1.8. For every one-dimensional set A, let $Q(A) = \sum_A f(x)$, where $f(x) = (\frac{2}{3})(\frac{1}{3})^x$, $x = 0, 1, 2, \ldots$, zero elsewhere. If $A_1 = \{x;\ x = 0, 1, 2, 3\}$ and $A_2 = \{x;\ x = 0, 1, 2, \ldots\}$, find $Q(A_1)$ and $Q(A_2)$.

1.9. For every one-dimensional set A for which the integral exists, let $Q(A) = \int_A f(x)\, dx$, where $f(x) = 6x(1 - x)$, $0 < x < 1$, zero elsewhere; otherwise, let $Q(A)$ be undefined. If $A_1 = \{x;\ \frac{1}{4} < x < \frac{3}{4}\}$, $A_2 = \{x;\ x = \frac{1}{2}\}$, and $A_3 = \{x;\ 0 < x < 10\}$, find $Q(A_1)$, $Q(A_2)$, and $Q(A_3)$.

1.10. For every one-dimensional set A, let $Q(A)$ be equal to the number of points in A that correspond to positive integers. If $A_1 = \{x;\ x$ a multiple of 3, less than or equal to $50\}$ and $A_2 = \{x;\ x$ a multiple of 7, less than or equal to $50\}$, find $Q(A_1)$, $Q(A_2)$, $Q(A_1 \cup A_2)$, and $Q(A_1 \cap A_2)$. Show that $Q(A_1 \cup A_2) = Q(A_1) + Q(A_2) - Q(A_1 \cap A_2)$.

1.11. For every two-dimensional set A, let $Q(A)$ be equal to the number of points (x, y) in A for which both x and y are positive integers. Find $Q(A_1)$ and $Q(A_2)$, where $A_1 = \{(x, y);\ x^2 + y^2 \leq 4\}$ and $A_2 = \{(x, y);\ x^2 + y^2 \leq 9\}$. Note that $A_1 \subset A_2$ and that $Q(A_1) \leq Q(A_2)$.

1.12. Let $Q(A) = \int\int_A (x^2 + y^2)\, dx\, dy$ for every two-dimensional set A for which the integral exists; otherwise, let $Q(A)$ be undefined. If $A_1 = \{(x, y);\ -1 \leq x \leq 1,\ -1 \leq y \leq 1\}$, $A_2 = \{(x, y);\ -1 \leq x = y \leq 1\}$, and $A_3 = \{(x, y);\ x^2 + y^2 \leq 1\}$, find $Q(A_1)$, $Q(A_2)$, and $Q(A_3)$.

1.13. Let $Q(A) = \int\int\int_A z\, dx\, dy\, dz$ for every three-dimensional set A for which the integral exists; otherwise, let $Q(A)$ be undefined. Evaluate the set function for the sets $A_1 = \{(x, y, z);\ (x, y, z) = (0, 0, 0)\}$ and $A_2 = \{(x, y, z);\ x^2 + y^2 \leq 4,\ 0 \leq z \leq 1\}$.

1.14. To join a certain club, a person must be either a statistician or a mathematician, or both. Of the 25 members in this club, 19 are statisticians and 16 are mathematicians. How many persons in the club are both a statistician and a mathematician?

1.15. After a hard-fought football game, it was reported that, of the 11 starting players, 8 hurt a hip, 6 hurt an arm, 5 hurt a knee, 3 hurt both a hip and an arm, 2 hurt both a hip and a knee, 1 hurt both an arm and a knee, and no one hurt all three. Comment on the accuracy of the report.

1.16. Let $Q(A) = \int_A x^2\,dx$ for every one-dimensional set A for which the integral exists; otherwise, let $Q(A)$ be undefined. (a) If $A_k = \{x; 0 \leq x \leq 2 - 1/k\}$, $k = 1, 2, 3, \ldots$, show that $Q(\lim_{k\to\infty} A_k) = \lim_{k\to\infty} Q(A_k)$. (b) If $A_k = \{x; |x| \leq 1 + 1/k\}$, $k = 1, 2, 3, \ldots$, show that $Q(\lim_{k\to\infty} A_k) = \lim_{k\to\infty} Q(A_k)$.

1.4 The Probability Set Function

Let $\mathscr{C}$ denote the set of every possible outcome of a random experiment; that is, $\mathscr{C}$ is the sample space. It is our purpose to define a set function $P(C)$ such that if C is a subset of $\mathscr{C}$, then $P(C)$ is the probability that the outcome of the random experiment is an element of C. Henceforth it will be tacitly assumed that the structure of each set C is sufficiently simple to allow the computation. We have already seen that advantages accrue if we take $P(C)$ to be that number about which the relative frequency f/N of the event C tends to stabilize after a long series of experiments. This important fact suggests some of the properties that we would surely want the set function $P(C)$ to possess. For example, no relative frequency is ever negative; accordingly, we would want $P(C)$ to be a nonnegative set function. Again, the relative frequency of the whole sample space $\mathscr{C}$ is always one. Thus we would want $P(\mathscr{C}) = 1$. Finally, if $C_1, C_2, C_3, \ldots$ are subsets of $\mathscr{C}$ such that no two of these subsets have a point in common, the relative frequency of the union of these sets is the sum of the relative frequencies of the sets, and we would want the set function $P(C)$ to reflect this additive property. We now formally define a probability set function.

Definition 7. If $P(C)$ is defined for a type of subset of the space $\mathscr{C}$, and if

(a) $P(C) \geq 0$,

(b) $P(C_1 \cup C_2 \cup C_3 \cup \cdots) = P(C_1) + P(C_2) + P(C_3) + \cdots$, where the sets C_i, $i = 1, 2, 3, \ldots$, are such that no two have a point in common,

(c) $P(\mathscr{C}) = 1$,

then $P(C)$ is called the *probability set function* of the outcome of the random experiment. For each subset C of $\mathscr{C}$, the number $P(C)$ is called the probability that the outcome of the random experiment is an

element of the set C, or the probability of the event C, or the probability measure of the set C.

A probability set function tells us how the probability is distributed over various subsets C of a sample space $\mathscr{C}$. In this sense we speak of a distribution of probability.

Remark. In the definition, the phrase "a type of subset of the space $\mathscr{C}$" would be explained more fully in a more advanced course. Nevertheless, a few observations can be made about the collection of subsets that are of the type. From condition (c) of the definition, we see that the space $\mathscr{C}$ must be in the collection. Condition (b) implies that if the sets $C_1, C_2, C_3, \ldots$ are in the collection, their union is also one of that type. Finally, we observe from the following theorems and their proofs that if the set C is in the collection, its complement must be one of those subsets. In particular, the null set, which is the complement of $\mathscr{C}$, must be in the collection.

The following theorems give us some other properties of a probability set function. In the statement of each of these theorems, $P(C)$ is taken, tacitly, to be a probability set function defined for a certain type of subset of the sample space $\mathscr{C}$.

Theorem 1. *For each* $C \subset \mathscr{C}$, $P(C) = 1 - P(C^*)$.

Proof. We have $\mathscr{C} = C \cup C^*$ and $C \cap C^* = 0$. Thus, from (c) and (b) of Definition 7, it follows that

$$1 = P(C) + P(C^*),$$

which is the desired result.

Theorem 2. *The probability of the null set is zero; that is,* $P(0) = 0$.

Proof. In Theorem 1, take $C = 0$ so that $C^* = \mathscr{C}$. Accordingly, we have

$$P(0) = 1 - P(\mathscr{C}) = 1 - 1 = 0,$$

and the theorem is proved.

Theorem 3. *If* C_1 *and* C_2 *are subsets of* $\mathscr{C}$ *such that* $C_1 \subset C_2$, *then* $P(C_1) \leq P(C_2)$.

Proof. Now $C_2 = C_1 \cup (C_1^* \cap C_2)$ and $C_1 \cap (C_1^* \cap C_2) = 0$. Hence, from (b) of Definition 7,

$$P(C_2) = P(C_1) + P(C_1^* \cap C_2).$$

However, from (a) of Definition 7, $P(C_1^* \cap C_2) \geq 0$; accordingly, $P(C_2) \geq P(C_1)$.

Theorem 4. *For each* $C \subset \mathscr{C}$, $0 \leq P(C) \leq 1$.

Proof. Since $0 \subset C \subset \mathscr{C}$, we have by Theorem 3 that

$$P(0) \leq P(C) \leq P(\mathscr{C}) \qquad \text{or} \qquad 0 \leq P(C) \leq 1,$$

the desired result.

Theorem 5. *If* C_1 *and* C_2 *are subsets of* $\mathscr{C}$, *then*

$$P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2).$$

Proof. Each of the sets $C_1 \cup C_2$ and C_2 can be represented, respectively, as a union of nonintersecting sets as follows:

$$C_1 \cup C_2 = C_1 \cup (C_1^* \cap C_2) \qquad \text{and} \qquad C_2 = (C_1 \cap C_2) \cup (C_1^* \cap C_2).$$

Thus, from (b) of Definition 7,

$$P(C_1 \cup C_2) = P(C_1) + P(C_1^* \cap C_2)$$

and

$$P(C_2) = P(C_1 \cap C_2) + P(C_1^* \cap C_2).$$

If the second of these equations is solved for $P(C_1^* \cap C_2)$ and this result substituted in the first equation, we obtain

$$P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2).$$

This completes the proof.

Example 1. Let $\mathscr{C}$ denote the sample space of Example 2 of Section 1.1. Let the probability set function assign a probability of $\frac{1}{36}$ to each of the 36 points in $\mathscr{C}$. If $C_1 = \{c; c = (1, 1), (2, 1), (3, 1), (4, 1), (5, 1)\}$ and $C_2 = \{c; c = (1, 2), (2, 2), (3, 2)\}$, then $P(C_1) = \frac{5}{36}$, $P(C_2) = \frac{3}{36}$, $P(C_1 \cup C_2) = \frac{8}{36}$, and $P(C_1 \cap C_2) = 0$.

Example 2. Two coins are to be tossed and the outcome is the ordered pair (face on the first coin, face on the second coin). Thus the sample space may be represented as $\mathscr{C} = \{c; c = (H, H), (H, T), (T, H), (T, T)\}$. Let the probability set function assign a probability of $\frac{1}{4}$ to each element of $\mathscr{C}$. Let $C_1 = \{c; c = (H, H), (H, T)\}$ and $C_2 = \{c; c = (H, H), (T, H)\}$. Then $P(C_1) = P(C_2) = \frac{1}{2}$, $P(C_1 \cap C_2) = \frac{1}{4}$, and, in accordance with Theorem 5, $P(C_1 \cup C_2) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$.

Let $\mathscr{C}$ denote a sample space and let $C_1, C_2, C_3, \ldots$ denote subsets of $\mathscr{C}$. If these subsets are such that no two have an element in common,

they are called mutually disjoint sets and the corresponding events $C_1, C_2, C_3, \ldots$ are said to be *mutually exclusive events*. Then, for example, $P(C_1 \cup C_2 \cup C_3 \cup \cdots) = P(C_1) + P(C_2) + P(C_3) + \cdots$ in accordance with (b) of Definition 7. Moreover, if $\mathscr{C} = C_1 \cup C_2 \cup C_3 \cup \cdots$, the mutually exclusive events are further characterized as being *exhaustive* and the probability of their union is obviously equal to one.

EXERCISES

1.17. A positive integer from one to six is to be chosen by casting a die. Thus the elements c of the sample space $\mathscr{C}$ are 1, 2, 3, 4, 5, 6. Let $C_1 = \{c; c = 1, 2, 3, 4\}$, $C_2 = \{c; c = 3, 4, 5, 6\}$. If the probability set function P assigns a probability of $\frac{1}{6}$ to each of the elements of $\mathscr{C}$, compute $P(C_1)$, $P(C_2)$, $P(C_1 \cap C_2)$, and $P(C_1 \cup C_2)$.

1.18. A random experiment consists in drawing a card from an ordinary deck of 52 playing cards. Let the probability set function P assign a probability of $\frac{1}{52}$ to each of the 52 possible outcomes. Let C_1 denote the collection of the 13 hearts and let C_2 denote the collection of the 4 kings. Compute $P(C_1)$, $P(C_2)$, $P(C_1 \cap C_2)$, and $P(C_1 \cup C_2)$.

1.19. A coin is to be tossed as many times as is necessary to turn up one head. Thus the elements c of the sample space $\mathscr{C}$ are H, TH, TTH, $TTTH$, and so forth. Let the probability set function P assign to these elements the respective probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$, and so forth. Show that $P(\mathscr{C}) = 1$. Let $C_1 = \{c; c \text{ is } H, TH, TTH, TTTH, \text{ or } TTTTH\}$. Compute $P(C_1)$. Let $C_2 = \{c; c \text{ is } TTTTH \text{ or } TTTTTH\}$. Compute $P(C_2)$, $P(C_1 \cap C_2)$, and $P(C_1 \cup C_2)$.

1.20. If the sample space is $\mathscr{C} = C_1 \cup C_2$ and if $P(C_1) = 0.8$ and $P(C_2) = 0.5$, find $P(C_1 \cap C_2)$.

1.21. If C_1 and C_2 are subsets of the sample space $\mathscr{C}$, show that

$$P(C_1 \cap C_2) \leq P(C_1) \leq P(C_1 \cup C_2) \leq P(C_1) + P(C_2).$$

1.22. Let C_1, C_2, and C_3 be three mutually disjoint subsets of the sample space $\mathscr{C}$. Find $P[(C_1 \cup C_2) \cap C_3]$ and $P(C_1^* \cup C_2^*)$.

1.23. If C_1, C_2, and C_3 are subsets of $\mathscr{C}$, show that

$$P(C_1 \cup C_2 \cup C_3) = P(C_1) + P(C_2) + P(C_3) - P(C_1 \cap C_2) - P(C_1 \cap C_3) - P(C_2 \cap C_3) + P(C_1 \cap C_2 \cap C_3).$$

What is the generalization of this result to four or more subsets of $\mathscr{C}$? *Hint.* Write $P(C_1 \cup C_2 \cup C_3) = P[C_1 \cup (C_2 \cup C_3)]$ and use Theorem 5.

1.5 Random Variables

The student will perceive that a sample space $\mathscr{C}$ may be tedious to describe if the elements of $\mathscr{C}$ are not numbers. We shall now discuss

how we may formulate a rule, or a set of rules, by which the elements c of $\mathscr{C}$ may be represented by numbers x or ordered pairs of numbers (x_1, x_2) or, more generally, ordered n-tuplets of numbers $(x_1, \ldots, x_n)$. We begin the discussion with a very simple example. Let the random experiment be the toss of a coin and let the sample space associated with the experiment be $\mathscr{C} = \{c;$ where c is T or c is $H\}$ and T and H represent, respectively, tails and heads. Let X be a function such that $X(c) = 0$ if c is T and let $X(c) = 1$ if c is H. Thus X is a single-valued, real-valued function defined on the sample space $\mathscr{C}$ which takes us from the sample space $\mathscr{C}$ to a space of real numbers $\mathscr{A} = \{x; x = 0, 1\}$. We call X a random variable and, in this example, the space associated with X is $\mathscr{A} = \{x; x = 0, 1\}$. We now formulate the definition of a random variable.

Definition 8. Given a random experiment with a sample space $\mathscr{C}$. A function X, which assigns to each element $c \in \mathscr{C}$ one and only one real number $X(c) = x$, is called a *random variable*. The space of X is the set of real numbers $\mathscr{A} = \{x; x = X(c), c \in \mathscr{C}\}$.

It may be that the set $\mathscr{C}$ has elements which are themselves real numbers. In such an instance we could write $X(c) = c$ so that $\mathscr{A} = \mathscr{C}$.

Let X be a random variable which is defined on a sample space $\mathscr{C}$ and let $\mathscr{A}$ be the space of X. Further, let A be a subset of $\mathscr{A}$. Just as we used the terminology "the event C," with $C \subset \mathscr{C}$, we shall now speak of "the event A." The probability $P(C)$ of the event C has been defined. We wish now to define the probability of the event A. This probability will be denoted by $\Pr[X \in A]$, where Pr is an abbreviation of "the probability that." With A a subset of $\mathscr{A}$, let C be that subset of $\mathscr{C}$ such that $C = \{c; c \in \mathscr{C}$ and $X(c) \in A\}$. Thus C has as its elements all outcomes in $\mathscr{C}$ for which the random variable X has a value which is in A. This prompts us to define, as we now do, $\Pr(X \in A)$ to be equal to $P(C)$, where $C = \{c; c \in \mathscr{C}$ and $X(c) \in A\}$. Thus $\Pr(X \in A)$ is an assignment of probability to a set A, which is a subset of the space $\mathscr{A}$ associated with the random variable X. This assignment is determined by the probability set function P and the random variable X and is sometimes denoted by $P_X(A)$. That is,

$$\Pr(X \in A) = P_X(A) = P(C),$$

where $C = \{c; c \in \mathscr{C}$ and $X(c) \in A\}$. Thus a random variable X is a function which carries the probability from a sample space $\mathscr{C}$ to a space $\mathscr{A}$ of real numbers. In this sense, with $A \subset \mathscr{A}$, the probability $P_X(A)$ is often called an *induced probability*.

The function $P_X(A)$ satisfies the conditions (a), (b), and (c) of the definition of a probability set function (p. 11). That is, $P_X(A)$ is also a probability set function. Conditions (a) and (c) are easily verified by observing, for an appropriate C, that

$$P_X(A) = P(C) \geq 0,$$

and that $\mathscr{C} = \{c; c \in \mathscr{C} \text{ and } X(c) \in \mathscr{A}\}$ requires

$$P_X(\mathscr{A}) = P(\mathscr{C}) = 1.$$

In discussing condition (b), let us restrict our attention to two mutually exclusive events A_1 and A_2. Here $P_X(A_1 \cup A_2) = P(C)$, where $C = \{c; c \in \mathscr{C} \text{ and } X(c) \in A_1 \cup A_2\}$. However,

$$C = \{c; c \in \mathscr{C} \text{ and } X(c) \in A_1\} \cup \{c; c \in \mathscr{C} \text{ and } X(c) \in A_2\},$$

or, for brevity, $C = C_1 \cup C_2$. But C_1 and C_2 are disjoint sets. This must be so, for if some c were common, say c_i, then $X(c_i) \in A_1$ and $X(c_i) \in A_2$. That is, the same number $X(c_i)$ belongs to both A_1 and A_2. This is a contradiction because A_1 and A_2 are disjoint sets. Accordingly,

$$P(C) = P(C_1) + P(C_2).$$

However, by definition, $P(C_1)$ is $P_X(A_1)$ and $P(C_2)$ is $P_X(A_2)$ and thus

$$P_X(A_1 \cup A_2) = P_X(A_1) + P_X(A_2).$$

This is condition (b) for two disjoint sets.

Thus each of $P_X(A)$ and $P(C)$ is a probability set function. But the student should fully recognize that the probability set function P is defined for subsets C of $\mathscr{C}$, whereas P_X is defined for subsets A of $\mathscr{A}$, and, in general, they are not the same set function. Nevertheless, they are closely related and some authors even drop the index X and write $P(A)$ for $P_X(A)$. They think it is quite clear that $P(A)$ means the probability of A, a subset of $\mathscr{A}$, and $P(C)$ means the probability of C, a subset of $\mathscr{C}$. From this point on, we shall adopt this convention and simply write $P(A)$.

Perhaps an additional example will be helpful. Let a coin be tossed twice and let our interest be in the *number* of heads to be observed. Thus the sample space is $\mathscr{C} = \{c; \text{where } c \text{ is } TT \text{ or } TH \text{ or } HT \text{ or } HH\}$. Let $X(c) = 0$ if c is TT; let $X(c) = 1$ if c is either TH or HT; and let $X(c) = 2$ if c is HH. Thus the space of the random variable X is $\mathscr{A} = \{x; x = 0, 1, 2\}$. Consider the subset A of the space $\mathscr{A}$, where $A = \{x; x = 1\}$. How is the probability of the event A defined? We take the subset C of $\mathscr{C}$ to have as its elements all outcomes in $\mathscr{C}$ for which the random variable X has a value which is an element of A.

Because $X(c) = 1$ if c is either TH or HT, then $C = \{c;$ where c is TH or $HT\}$. Thus $P(A) = \Pr[X \in A] = P(C)$. Since $A = \{x; x = 1\}$, then $P(A) = \Pr[X \in A]$ can be written more simply as $\Pr[X = 1]$. Let $C_1 = \{c; c$ is $TT\}$, $C_2 = \{c;\ c$ is $TH\}$, $C_3 = \{c;\ c$ is $HT\}$, and $C_4 = \{c;\ c$ is $HH\}$ denote subsets of $\mathscr{C}$. Suppose that our probability set function $P(C)$ assigns a probability of $\frac{1}{4}$ to each of the sets C_i, $i = 1, 2, 3, 4$. Then $P(C_1) = \frac{1}{4}$, $P(C_2 \cup C_3) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$, and $P(C_4) = \frac{1}{4}$. Let us now point out how much simpler it is to couch these statements in a language which involves the random variable X. Because X is the number of heads to be observed in tossing a coin two times, we have

$$\Pr(X = 0) = \tfrac{1}{4}, \text{ since } P(C_1) = \tfrac{1}{4};$$
$$\Pr(X = 1) = \tfrac{1}{2}, \text{ since } P(C_2 \cup C_3) = \tfrac{1}{2};$$

and

$$\Pr(X = 2) = \tfrac{1}{4}, \text{ since } P(C_4) = \tfrac{1}{4}.$$

This may be further condensed in the following table:

x	0	1	2
$\Pr(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

This table depicts the distribution of probability over the elements of $\mathscr{A}$, the space of the random variable X.

We shall now discuss two random variables. Again, we start with an example. A coin is to be tossed three times and our interest is in the ordered number pair (number of H's on first two tosses, number H's on all three tosses). Thus, the sample space is $\mathscr{C} = \{c; c = c_i, i = 1, 2, \ldots, 8\}$, where c_1 is TTT, c_2 is TTH, c_3 is THT, c_4 is HTT, c_5 is THH, c_6 is HTH, c_7 is HHT, and c_8 is HHH. Let X_1 and X_2 be two functions such that $X_1(c_1) = X_1(c_2) = 0$, $X_1(c_3) = X_1(c_4) = X_1(c_5) = X_1(c_6) = 1$, $X_1(c_7) = X_1(c_8) = 2$; and $X_2(c_1) = 0$, $X_2(c_2) = X_2(c_3) = X_2(c_4) = 1$, $X_2(c_5) = X_2(c_6) = X_2(c_7) = 2$, $X_2(c_8) = 3$. Thus X_1 and X_2 are single-valued, real-valued functions defined on the sample space $\mathscr{C}$ which takes us from that sample space to the space of ordered number pairs

$$\mathscr{A} = \{(x_1, x_2); (x_1, x_2) = (0, 0), (0, 1), (1, 1), (1, 2), (2, 2), (2, 3)\}.$$

Thus X_1 and X_2 are two random variables defined on the space $\mathscr{C}$, and, in this example, the space of these random variables is the two-dimensional set $\mathscr{A}$ given immediately above. We now formulate the definition of the space of two random variables.

Definition 9. Given a random experiment with a sample space $\mathscr{C}$. Consider two random variables X_1 and X_2, which assign to each element

c of $\mathscr{C}$ one and only one ordered pair of numbers $X_1(c) = x_1$, $X_2(c) = x_2$. The space of X_1 and X_2 is the set of ordered pairs $\mathscr{A} = \{(x_1, x_2); x_1 = X_1(c), x_2 = X_2(c), c \in \mathscr{C}\}$.

Let $\mathscr{A}$ be the space associated with the two random variables X_1 and X_2 and let A be a subset of $\mathscr{A}$. As in the case of one random variable, we shall speak of the event A. We wish to define the probability of the event A, which we denote by $\Pr[(X_1, X_2) \in A]$. Take $C = \{c; c \in \mathscr{C} \text{ and } [X_1(c), X_2(c)] \in A\}$, where $\mathscr{C}$ is the sample space. We then define $\Pr[(X_1, X_2) \in A] = P(C)$, where P is the probability set function defined for subsets C of $\mathscr{C}$. Here again we could denote $\Pr[(X_1, X_2) \in A]$ by the probability set function $P_{X_1, X_2}(A)$; but, with our previous convention, we simply write

$$P(A) = \Pr[(X_1, X_2) \in A].$$

Again it is important to observe that this function is a probability set function defined for subsets A of the space $\mathscr{A}$.

Let us return to the example in our discussion of two random variables. Consider the subset A of $\mathscr{A}$, where $A = \{(x_1, x_2); (x_1, x_2) = (1, 1), (1, 2)\}$. To compute $\Pr[(X_1, X_2) \in A] = P(A)$ we must include as elements of C all outcomes in $\mathscr{C}$ for which the random variables X_1 and X_2 take values (x_1, x_2) which are elements of A. Now $X_1(c_3) = 1$, $X_2(c_3) = 1$, $X_1(c_4) = 1$, and $X_2(c_4) = 1$. Also $X_1(c_5) = 1$, $X_2(c_5) = 2$, $X_1(c_6) = 1$, and $X_2(c_6) = 2$. Thus $P(A) = \Pr[(X_1, X_2) \in A] = P(C)$, where $C = \{c; c = c_3, c_4, c_5, \text{ or } c_6\}$. Suppose that our probability set function $P(C)$ assigns a probability of $\frac{1}{8}$ to each of the eight elements of $\mathscr{C}$. Then $P(A)$, which can be written as $\Pr(X_1 = 1, X_2 = 1 \text{ or } 2)$, is equal to $\frac{4}{8} = \frac{1}{2}$. It is left for the student to show that we can tabulate the probability which is then assigned to each of the elements of $\mathscr{A}$, with the following result:

(x_1, x_2)	(0, 0)	(0, 1)	(1, 1)	(1, 2)	(2, 2)	(2, 3)
$\Pr[(X_1, X_2) = (x_1, x_2)]$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

This table depicts the distribution of probability over the elements of $\mathscr{A}$, the space of the random variables X_1 and X_2.

The preceding notions about one and two random variables can be immediately extended to n random variables. We make the following definition of the space of n random variables.

Definition 10. Given a random experiment with the sample space $\mathscr{C}$. Let the random variable X_i assign to each element $c \in \mathscr{C}$ one and

only one real number $X_i(c) = x_i$, $i = 1, 2, \ldots, n$. The space of these random variables is the set of ordered n-tuplets $\mathscr{A} = \{(x_1, x_2, \ldots, x_n); x_1 = X_1(c), \ldots, x_n = X_n(c), c \in \mathscr{C}\}$. Further, let A be a subset of $\mathscr{A}$. Then $\Pr[(X_1, \ldots, X_n) \in A] = P(C)$, where $C = \{c; c \in \mathscr{C}$ and $[X_1(c), X_2(c), \ldots, X_n(c)] \in A\}$.

Again we should make the comment that $\Pr[(X_1, \ldots, X_n) \in A]$ could be denoted by the probability set function $P_{X_1,\ldots,X_n}(A)$. But, if there is no chance of misunderstanding, it will be written simply as $P(A)$.

Up to this point, our illustrative examples have dealt with a sample space $\mathscr{C}$ which contains a finite number of elements. We now give an example of a sample space $\mathscr{C}$ which is an interval.

Example 1. Let the outcome of a random experiment be a point on the interval $(0, 1)$. Thus, $\mathscr{C} = \{c; 0 < c < 1\}$. Let the probability set function be given by

$$P(C) = \int_C dz.$$

For instance, if $C = \{c; \frac{1}{4} < c < \frac{1}{2}\}$, then

$$P(C) = \int_{1/4}^{1/2} dz = \frac{1}{4}.$$

Define the random variable X to be $X = X(c) = c + 2$. Accordingly, the space of X is $\mathscr{A} = \{x; 2 < x < 3\}$. We wish to determine the probability set function of X, namely, $P(A)$, $A \subset \mathscr{A}$. At this time, let A be the set $\{x; 2 < x < b\}$, where $2 < b < 3$. Now X is between 2 and b when and only when $c \in C = \{c; 0 < c < b - 2\}$. Hence

$$P_X(A) = P(A) = P(C) = \int_0^{b-2} dz.$$

In the integral, make the change of variable $x = z + 2$ and obtain

$$P_X(A) = P(A) = \int_2^b dx = \int_A dx$$

since $A = \{x; 2 < x < b\}$. This kind of argument holds for every set $A \subset \mathscr{A}$ for which the integral

$$P(A) = \int_A dx$$

exists. Thus the probability set function of X is this integral.

In statistics we are usually more interested in the probability set function of the random variable X than we are in the sample space $\mathscr{C}$ and the probability set function $P(C)$. Therefore, in most instances, we

begin with an assumed distribution of probability for the random variable X. Moreover, we do this same kind of thing with two or more random variables. Two illustrative examples follow.

Example 2. Let the probability set function $P(A)$ of a random variable X be

$$P(A) = \int_A f(x)\,dx, \qquad \text{where } f(x) = \frac{3x^2}{8}, \quad x \in \mathscr{A} = \{x; 0 < x < 2\}.$$

Let $A_1 = \{x; 0 < x < \frac{1}{2}\}$ and $A_2 = \{x; 1 < x < 2\}$ be two subsets of $\mathscr{A}$. Then

$$P(A_1) = \Pr(X \in A_1) = \int_{A_1} f(x)\,dx = \int_0^{1/2} \frac{3x^2}{8}\,dx = \frac{1}{64}$$

and

$$P(A_2) = \Pr(X \in A_2) = \int_{A_2} f(x)\,dx = \int_1^2 \frac{3x^2}{8}\,dx = \frac{7}{8}.$$

To compute $P(A_1 \cup A_2)$, we note that $A_1 \cap A_2 = 0$; then we have $P(A_1 \cup A_2) = P(A_1) + P(A_2) = \frac{57}{64}$.

Example 3. Let $\mathscr{A} = \{(x, y); 0 < x < y < 1\}$ be the space of two random variables X and Y. Let the probability set function be

$$P(A) = \int_A \int 2\,dx\,dy.$$

If A is taken to be $A_1 = \{(x, y); \frac{1}{2} < x < y < 1\}$, then

$$P(A_1) = \Pr[(X, Y) \in A_1] = \int_{1/2}^{1} \int_{1/2}^{y} 2\,dx\,dy = \frac{1}{4}.$$

If A is taken to be $A_2 = \{(x, y); x < y < 1, 0 < x \leq \frac{1}{2}\}$, then $A_2 = A_1^*$, and

$$P(A_2) = \Pr[(X, Y) \in A_2] = P(A_1^*) = 1 - P(A_1) = \frac{3}{4}.$$

EXERCISES

1.24. Let a card be selected from an ordinary deck of playing cards. The outcome c is one of these 52 cards. Let $X(c) = 4$ if c is an ace, let $X(c) = 3$ if c is a king, let $X(c) = 2$ if c is a queen, let $X(c) = 1$ if c is a jack, and let $X(c) = 0$ otherwise. Suppose $P(C)$ assigns a probability of $\frac{1}{52}$ to each outcome c. Describe the induced probability $P_X(A)$ on the space $\mathscr{A} = \{x; x = 0, 1, 2, 3, 4\}$ of the random variable X.

1.25. Let a point be selected from the sample space $\mathscr{C} = \{c; 0 < c < 10\}$. Let $C \subset \mathscr{C}$ and let the probability set function be $P(C) = \int_C \frac{1}{10}\,dz$. Define the random variable X to be $X = X(c) = 2c - 10$. Find the probability set function of X.

1.26. Let the probability set function $P(A)$ of two random variables X and Y be $P(A) = \sum_A \sum f(x, y)$, where $f(x, y) = \frac{1}{52}$, $(x, y) \in \mathscr{A} = \{(x, y); (x, y) = (0, 1), (0, 2), \ldots, (0, 13), (1, 1), \ldots, (1, 13), \ldots, (3, 13)\}$. Compute $P(A) = \Pr[(X, Y) \in A]$: (a) when $A = \{(x, y); (x, y) = (0, 4), (1, 3), (2, 2)\}$; (b) when $A = \{(x, y); x + y = 4, (x, y) \in \mathscr{A}\}$.

1.27. Let the probability set function $P(A)$ of the random variable X be $P(A) = \int_A f(x)\,dx$, where $f(x) = 2x/9$, $x \in \mathscr{A} = \{x; 0 < x < 3\}$. Let $A_1 = \{x; 0 < x < 1\}$, $A_2 = \{x; 2 < x < 3\}$. Compute $P(A_1) = \Pr[X \in A_1]$, $P(A_2) = \Pr(X \in A_2)$, and $P(A_1 \cup A_2) = \Pr(X \in A_1 \cup A_2)$.

1.28. Let the space of the random variable X be $\mathscr{A} = \{x; 0 < x < 1\}$. If $A_1 = \{x; 0 < x < \frac{1}{2}\}$ and $A_2 = \{x; \frac{1}{2} \leq x < 1\}$, find $P(A_2)$ if $P(A_1) = \frac{1}{4}$.

1.29. Let the space of the random variable X be $\mathscr{A} = \{x; 0 < x < 10\}$ and let $P(A_1) = \frac{3}{8}$, where $A_1 = \{x; 1 < x < 5\}$. Show that $P(A_2) \leq \frac{5}{8}$, where $A_2 = \{x; 5 \leq x < 10\}$.

1.30. Let the subsets $A_1 = \{x; \frac{1}{4} < x < \frac{1}{2}\}$ and $A_2 = \{x; \frac{1}{2} \leq x < 1\}$ of the space $\mathscr{A} = \{x; 0 < x < 1\}$ of the random variable X be such that $P(A_1) = \frac{1}{8}$ and $P(A_2) = \frac{1}{2}$. Find $P(A_1 \cup A_2)$, $P(A_1^*)$, and $P(A_1^* \cap A_2^*)$.

1.31. Let $A_1 = \{(x, y); x \leq 2, y \leq 4\}$, $A_2 = \{(x, y); x \leq 2, y \leq 1\}$, $A_3 = \{(x, y); x \leq 0, y \leq 4\}$, and $A_4 = \{(x, y); x \leq 0, y \leq 1\}$ be subsets of the space $\mathscr{A}$ of two random variables X and Y, which is the entire two-dimensional plane. If $P(A_1) = \frac{7}{8}$, $P(A_2) = \frac{4}{8}$, $P(A_3) = \frac{3}{8}$, and $P(A_4) = \frac{2}{8}$, find $P(A_5)$, where $A_5 = \{(x, y); 0 < x \leq 2, 1 < y \leq 4\}$.

1.32. Let the probability set function of the random variable X be

$$P(A) = \int_A e^{-x}\,dx, \qquad \text{where } \mathscr{A} = \{x; 0 < x < \infty\}.$$

Let $A_k = \{x; 2 - 1/k < x \leq 3\}$, $k = 1, 2, 3, \ldots$. Find $\lim_{k \to \infty} A_k$ and $P\left(\lim_{k \to \infty} A_k\right)$. Find $P(A_k)$ and $\lim_{k \to \infty} P(A_k)$. Note that $\lim_{k \to \infty} P(A_k) = P\left(\lim_{k \to \infty} A_k\right)$.

1.6 The Probability Density Function

Let X denote a random variable with space $\mathscr{A}$ and let A be a subset of $\mathscr{A}$. If we know how to compute $P(C)$, $C \subset \mathscr{C}$, then for each A under consideration we can compute $P(A) = \Pr[X \in A]$; that is, we know how the probability is distributed over the various subsets of $\mathscr{A}$. In this sense, we speak of the distribution of the random variable X, meaning, of course, the distribution of probability. Moreover, we can use this convenient terminology when more than one random variable is involved and, in the sequel, we shall do this.

In this section, we shall investigate some random variables whose distributions can be described very simply by what will be called the *probability density function*. The two types of distributions that we shall consider are called, respectively, the *discrete type* and the *continuous type*. For simplicity of presentation, we first consider a distribution of one random variable.

(a) *The discrete type of random variable.* Let X denote a random variable with one-dimensional space $\mathscr{A}$. Suppose the space $\mathscr{A}$ is a set of points such that there is at most a finite number of points of $\mathscr{A}$ in every finite interval. Such a set $\mathscr{A}$ will be called a set of discrete points. Let a function $f(x)$ be such that $f(x) > 0$, $x \in \mathscr{A}$, and that

$$\sum_{\mathscr{A}} f(x) = 1.$$

Whenever a probability set function $P(A)$, $A \subset \mathscr{A}$, can be expressed in terms of such an $f(x)$ by

$$P(A) = \Pr(X \in A) = \sum_{A} f(x),$$

then X is called a random variable of the *discrete type*, and X is said to have a distribution of the discrete type.

Example 1. Let X be a random variable of the discrete type with space $\mathscr{A} = \{x; x = 0, 1, 2, 3, 4\}$. Let

$$P(A) = \sum_{A} f(x),$$

where

$$f(x) = \frac{4!}{x!\,(4-x)!}\left(\frac{1}{2}\right)^4, \qquad x \in \mathscr{A},$$

and, as usual, $0! = 1$. Then if $A = \{x; x = 0, 1\}$, we have

$$\Pr(X \in A) = \frac{4!}{0!\,4!}\left(\frac{1}{2}\right)^4 + \frac{4!}{1!\,3!}\left(\frac{1}{2}\right)^4 = \frac{5}{16}.$$

Example 2. Let X be a random variable of the discrete type with space $= \{x; x = 1, 2, 3, \ldots\}$, and let

$$f(x) = (\tfrac{1}{2})^x, \qquad x \in \mathscr{A}.$$

Then

$$\Pr(X \in A) = \sum_{A} f(x).$$

If $A = \{x; x = 1, 3, 5, 7, \ldots\}$, we have

$$\Pr(X \in A) = (\tfrac{1}{2}) + (\tfrac{1}{2})^3 + (\tfrac{1}{2})^5 + \cdots = \tfrac{2}{3}.$$

(b) *The continuous type of random variable.* Let the one-dimensional set $\mathscr{A}$ be such that the Riemann integral

$$\int_{\mathscr{A}} f(x)\,dx = 1,$$

where (i) $f(x) > 0$, $x \in \mathscr{A}$, and (ii) $f(x)$ has at most a finite number of discontinuities in every finite interval that is a subset of $\mathscr{A}$. If $\mathscr{A}$ is the space of the random variable X and if the probability set function $P(A)$, $A \subset \mathscr{A}$, can be expressed in terms of such an $f(x)$ by

$$P(A) = \Pr(X \in A) = \int_A f(x)\,dx,$$

then X is said to be a random variable of the *continuous type* and to have a distribution of that type.

Example 3. Let the space $\mathscr{A} = \{x; 0 < x < \infty\}$, and let

$$f(x) = e^{-x}, \qquad x \in \mathscr{A}.$$

If X is a random variable of the continuous type so that

$$\Pr(X \in A) = \int_A e^{-x}\,dx,$$

we have, with $A = \{x; 0 < x < 1\}$,

$$\Pr(X \in A) = \int_0^1 e^{-x}\,dx = 1 - e^{-1}.$$

Example 4. Let X be a random variable of the continuous type with space $\mathscr{A} = \{x; 0 < x < 1\}$. Let the probability set function be

$$P(A) = \int_A f(x)\,dx,$$

where

$$f(x) = cx^2, \qquad x \in \mathscr{A}.$$

Since $P(A)$ is a probability set function, $P(\mathscr{A}) = 1$. Hence the constant c is determined by

$$\int_0^1 cx^2\,dx = 1,$$

or $c = 3$.

It is seen that whether the random variable X is of the discrete type or of the continuous type, the probability $\Pr(X \in A)$ is completely determined by a function $f(x)$. In either case $f(x)$ is called the *probability density function* (hereafter abbreviated p.d.f.) of the random variable X. If we restrict ourselves to random variables of either the discrete type

or the continuous type, we may work exclusively with the p.d.f. $f(x)$. This affords an enormous simplification; but it should be recognized that this simplification is obtained at considerable cost from a mathematical point of view. Not only shall we exclude from consideration many random variables that do not have these types of distributions but we shall also exclude many interesting subsets of the space. In this book, however, we shall in general restrict ourselves to these simple types of random variables.

Remarks. Let X denote the number of spots that show when a die is cast. We can assume that X is a random variable with $\mathscr{A} = \{x; x = 1, 2, \ldots, 6\}$ and with a p.d.f. $f(x) = \frac{1}{6}$, $x \in \mathscr{A}$. Other assumptions can be made to provide different mathematical models for this experiment. Experimental evidence can be used to help one decide which model is the more realistic. Next, let X denote the point at which a balanced pointer comes to rest. If the circumference is graduated $0 \leq x < 1$, a reasonable mathematical model for this experiment is to take X to be a random variable with $\mathscr{A} = \{x; 0 \leq x < 1\}$ and with a p.d.f. $f(x) = 1$, $x \in \mathscr{A}$.

The notion of the p.d.f. of one random variable X can be extended to the notion of the p.d.f. of two or more random variables. Under certain restrictions on the space $\mathscr{A}$ and the function $f > 0$ on $\mathscr{A}$ (restrictions which will not be enumerated here), we say that the two random variables X and Y are of the discrete type or of the continuous type, and have a distribution of that type, according as the probability set function $P(A)$, $A \subset \mathscr{A}$, can be expressed as

$$P(A) = \Pr[(X, Y) \in A] = \sum_{A}\sum f(x, y),$$

or as

$$P(A) = \Pr[(X, Y) \in A] = \int_{A}\int f(x, y)\, dx\, dy.$$

In either case f is called the p.d.f. of the two random variables X and Y. Of necessity, $P(\mathscr{A}) = 1$ in each case. More generally, we say that the n random variables $X_1, X_2, \ldots, X_n$ are of the discrete type or of the continuous type, and have a distribution of that type, according as the probability set function $P(A)$, $A \subset \mathscr{A}$, can be expressed as

$$P(A) = \Pr[(X_1, \ldots, X_n) \in A] = \sum_{A}\cdots\sum f(x_1, \ldots, x_n),$$

or as

$$P(A) = \Pr[(X_1, \ldots, X_n) \in A] = \int_{A}\cdots\int f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n.$$

The idea to be emphasized is that a function f, whether in one or more variables, essentially satisfies the conditions of being a p.d.f. if $f > 0$ on a space $\mathscr{A}$ and if its integral [for the continuous type of random variable(s)] or its sum [for the discrete type of random variable(s)] over $\mathscr{A}$ is one.

Our notation can be considerably simplified when we restrict ourselves to random variables of the continuous or discrete types. Suppose the space of a continuous type of random variable X is $\mathscr{A} = \{x; 0 < x < \infty\}$ and that the p.d.f. of X is e^{-x}, $x \in \mathscr{A}$. We shall in no manner alter the distribution of X [that is, alter any $P(A)$, $A \subset \mathscr{A}$] if we extend the definition of the p.d.f. of X by writing

$$\begin{aligned} f(x) &= e^{-x}, \qquad 0 < x < \infty, \\ &= 0 \text{ elsewhere}, \end{aligned}$$

and then refer to $f(x)$ as the p.d.f. of X. We have

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{\infty} e^{-x}\,dx = 1.$$

Thus we may treat the entire axis of reals as though it were the space of X. Accordingly, we now replace

$$\int_{\mathscr{A}} f(x)\,dx \qquad \text{by} \qquad \int_{-\infty}^{\infty} f(x)\,dx.$$

Similarly, we may extend the definition of a p.d.f. $f(x, y)$ over the entire xy-plane; or a p.d.f. $f(x, y, z)$ throughout three-dimensional space, and so on. We shall do this consistently so that tedious, repetitious references to the space $\mathscr{A}$ can be avoided. Once this is done, we replace

$$\int_{\mathscr{A}}\int f(x, y)\,dx\,dy \qquad \text{by} \qquad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy,$$

and so on. Similarly, after extending the definition of a p.d.f. of the discrete type, we replace, for one random variable,

$$\sum_{\mathscr{A}} f(x) \qquad \text{by} \qquad \sum_{x} f(x),$$

and, for two random variables,

$$\sum\sum_{\mathscr{A}} f(x, y) \qquad \text{by} \qquad \sum_{y}\sum_{x} f(x, y),$$

and so on.

In accordance with this convention (of extending the definition of a p.d.f.), it is seen that a point function f, whether in one or more variables, essentially satisfies the conditions of being a p.d.f. if (a) f is defined and is not negative for all real values of its argument(s) and if (b) its

integral [for the continuous type of random variable(s)], or its sum [for the discrete type of random variable(s)] over all real values of its argument(s) is one.

If $f(x)$ is the p.d.f. of a continuous type of random variable X and if A is the set $\{x; a < x < b\}$, then $P(A) = \Pr(X \in A)$ can be written as

$$\Pr(a < X < b) = \int_a^b f(x)\,dx.$$

Moreover, if $A = \{x; x = a\}$, then

$$P(A) = \Pr(X \in A) = \Pr(X = a) = \int_a^a f(x)\,dx = 0,$$

since the integral $\int_a^a f(x)\,dx$ is defined in calculus to be zero. That is, if X is a random variable of the continuous type, the probability of every set consisting of a single point is zero. This fact enables us to write, say,

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

More important, this fact allows us to change the value of the p.d.f. of a continuous type of random variable X at a single point without altering the distribution of X. For instance, the p.d.f.

$$\begin{aligned} f(x) &= e^{-x}, \qquad 0 < x < \infty, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

can be written as

$$\begin{aligned} f(x) &= e^{-x}, \qquad 0 \leq x < \infty, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

without changing any $P(A)$. We observe that these two functions differ only at $x = 0$ and $\Pr(X = 0) = 0$. More generally, if two probability density functions of random variables of the continuous type differ only on a set having probability zero, the two corresponding probability set functions are exactly the same. Unlike the continuous type, the p.d.f. of a discrete type of random variable may not be changed at any point, since a change in such a p.d.f. alters the distribution of probability.

Finally, if a p.d.f. in one or more variables is explicitly defined, we can see by inspection whether the random variables are of the continuous or discrete type. For example, it seems obvious that the p.d.f.

$$\begin{aligned} f(x, y) &= \frac{9}{4^{x+y}}, \qquad x = 1, 2, 3, \ldots, y = 1, 2, 3, \ldots, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

is a p.d.f. of two discrete-type random variables X and Y, whereas the p.d.f.

$$\begin{aligned} f(x, y) &= 4xye^{-x^2-y^2}, \qquad 0 < x < \infty, 0 < y < \infty, \\ &= 0 \text{ elsewhere}, \end{aligned}$$

is clearly a p.d.f. of two continuous-type random variables X and Y. In such cases it seems unnecessary to specify which of the two simpler types of random variables is under consideration.

Example 5. Let the random variable X have the p.d.f.

$$\begin{aligned} f(x) &= 2x, \qquad 0 < x < 1, \\ &= 0 \text{ elsewhere}. \end{aligned}$$

Find $\Pr(\frac{1}{2} < X < \frac{3}{4})$ and $\Pr(-\frac{1}{2} < X < \frac{1}{2})$. First,

$$\Pr(\tfrac{1}{2} < X < \tfrac{3}{4}) = \int_{1/2}^{3/4} f(x)\,dx = \int_{1/2}^{3/4} 2x\,dx = \frac{5}{16}.$$

Next,

$$\begin{aligned} \Pr(-\tfrac{1}{2} < X < \tfrac{1}{2}) &= \int_{-1/2}^{1/2} f(x)\,dx \\ &= \int_{-1/2}^{0} 0\,dx + \int_{0}^{1/2} 2x\,dx \\ &= 0 + \tfrac{1}{4} = \tfrac{1}{4}. \end{aligned}$$

Example 6. Let

$$\begin{aligned} f(x, y) &= 6x^2y, \qquad 0 < x < 1, 0 < y < 1, \\ &= 0 \text{ elsewhere}, \end{aligned}$$

be the p.d.f. of two random variables X and Y. We have, for instance,

$$\begin{aligned} \Pr(0 < X < \tfrac{3}{4}, \tfrac{1}{3} < Y < 2) &= \int_{1/3}^{2} \int_{0}^{3/4} f(x, y)\,dx\,dy \\ &= \int_{1/3}^{1} \int_{0}^{3/4} 6x^2y\,dx\,dy + \int_{1}^{2} \int_{0}^{3/4} 0\,dx\,dy \\ &= \tfrac{3}{8} + 0 = \tfrac{3}{8}. \end{aligned}$$

EXERCISES

1.33. For each of the following, find the constant c so that $f(x)$ satisfies the conditions of being a p.d.f. of one random variable X.

(a) $f(x) = c(\frac{2}{3})^x$, $x = 1, 2, 3, \ldots$, zero elsewhere.

(b) $f(x) = cxe^{-x}$, $0 < x < \infty$, zero elsewhere.

1.34. Let $f(x) = x/15$, $x = 1, 2, 3, 4, 5$, zero elsewhere, be the p.d.f. of X. Find $\Pr[X = 1 \text{ or } 2]$, $\Pr[\frac{1}{2} < X < \frac{5}{2}]$, and $\Pr[1 \leq X \leq 2]$.

1.35. Let $f(x) = 1/x^2$, $1 < x < \infty$, zero elsewhere, be the p.d.f. of X. If $A_1 = \{x; 1 < x < 2\}$ and $A_2 = \{x; 4 < x < 5\}$, find $P(A_1 \cup A_2)$ and $P(A_1 \cap A_2)$.

1.36. Let $f(x_1, x_2) = 4x_1x_2$, $0 < x_1 < 1$, $0 < x_2 < 1$, zero elsewhere, be the p.d.f. of X_1 and X_2. Find Pr $(0 < X_1 < \frac{1}{2}, \frac{1}{4} < X_2 < 1)$, Pr $(X_1 = X_2)$, Pr $(X_1 < X_2)$, and Pr $(X_1 \leq X_2)$.

1.37. Let $f(x_1, x_2, x_3) = \exp[-(x_1 + x_2 + x_3)]$, $0 < x_1 < \infty$, $0 < x_2 < \infty$, $0 < x_3 < \infty$, zero elsewhere, be the p.d.f. of X_1, X_2, X_3. Compute Pr $(X_1 < X_2 < X_3)$ and Pr $(X_1 = X_2 < X_3)$. The symbol exp (w) means e^w.

1.38. A *mode* of a distribution of one random variable X of the continuous or discrete type is a value of x that maximizes the p.d.f. $f(x)$. If there is only one such x, it is called the *mode of the distribution*. Find the mode of each of the following distributions:

(a) $f(x) = (\frac{1}{2})^x$, $x = 1, 2, 3, \ldots$, zero elsewhere.

(b) $f(x) = 12x^2(1 - x)$, $0 < x < 1$, zero elsewhere.

(c) $f(x) = (\frac{1}{2})x^2e^{-x}$, $0 < x < \infty$, zero elsewhere.

1.39. A *median* of a distribution of one random variable X of the discrete or continuous type is a value of x such that Pr $(X < x) \leq \frac{1}{2}$ and Pr $(X \leq x) \geq \frac{1}{2}$. If there is only one such x, it is called the *median of the distribution*. Find the median of each of the following distributions:

(a) $f(x) = \dfrac{4!}{x!\,(4 - x)!}\left(\dfrac{1}{4}\right)^x\left(\dfrac{3}{4}\right)^{4-x}$, $x = 0, 1, 2, 3, 4$, zero elsewhere.

(b) $f(x) = 3x^2$, $0 < x < 1$, zero elsewhere.

(c) $f(x) = \dfrac{1}{\pi(1 + x^2)}$, $-\infty < x < \infty$.

1.40. Let $0 < p < 1$. A $(100p)$th *percentile* (*quantile* of order p) of the distribution of a random variable X is a value ξ_p such that Pr $(X < \xi_p) \leq p$ and Pr $(X \leq \xi_p) \geq p$. Find the twentieth percentile of the distribution which has p.d.f. $f(x) = 4x^3$, $0 < x < 1$, zero elsewhere.

1.41. Let $f(x) = 12x(1 - x)^2$, $0 < x < 1$, zero elsewhere, be the p.d.f. of X. Find the mode of the distribution. Show that the shortest interval (a, b) such that $P(a < X < b) = p$, where p is an assigned number, $0 < p < 1$, contains the mode. *Hint.* From $\int_a^b f(x)\,dx = p$ and $d(b - a)/da = 0$, first show that $f(a) = f(b)$.

1.42. Given that the nonnegative function $g(x)$ has the property that

$$\int_0^\infty g(x)\,dx = 1.$$

Show that $f(x_1, x_2) = [2g(\sqrt{x_1^2 + x_2^2})]/(\pi\sqrt{x_1^2 + x_2^2})$, $0 < x_1 < \infty$, $0 < x_2 < \infty$, zero elsewhere, satisfies the conditions of being a p.d.f. of two continuous-type random variables X_1 and X_2.

1.43. Let $f(x) = x/10$, $x = 1, 2, 3, 4$, zero elsewhere, be the p.d.f. of X. Show, for $0 < h$, that:

(a) $\lim_{h \to 0} \Pr(2 - h < X \leq 2) = P(\lim_{h \to 0} \{x; 2 - h < x \leq 2\})$.

(b) $\lim_{h \to 0} \Pr(3 < X \leq 3 + h) = P(\lim_{h \to 0} \{x; 3 < x \leq 3 + h\})$.

1.7 The Distribution Function

Let the random variable X have the probability set function $P(A)$, where A is a one-dimensional set. Take x to be a real number and consider the set A which is an unbounded set from $-\infty$ to x, including the point x itself. For all such sets A we have $P(A) = \Pr(X \in A) = \Pr(X \leq x)$. This probability depends on the point x; that is, this probability is a function of the point x. This point function is denoted by the symbol $F(x) = \Pr(X \leq x)$. The function $F(x)$ is called the *distribution function* (sometimes, *cumulative distribution function*) of the random variable X. Since $F(x) = \Pr(X \leq x)$, then, with $f(x)$ the p.d.f., we have

$$F(x) = \sum_{w \leq x} f(w),$$

for the discrete type of random variable, and

$$F(x) = \int_{-\infty}^{x} f(w)\, dw,$$

for the continuous type of random variable. We speak of a distribution function $F(x)$ as being of the continuous or discrete type depending on whether the random variable is of the continuous or discrete type.

Remark. If X is a random variable of the continuous type, the p.d.f. $f(x)$ has at most a finite number of discontinuities in every finite interval. This means (i) that the distribution function $F(x)$ is everywhere continuous and (ii) that the derivative of $F(x)$ with respect to x exists and is equal to $f(x)$ at each point of continuity of $f(x)$. That is, $F'(x) = f(x)$ at each point of continuity of $f(x)$. If the random variable X is of the discrete type, most surely the p.d.f. $f(x)$ is *not* the derivative of $F(x)$ with respect to x (that is, with respect to Lebesgue measure); but $f(x)$ *is* the (Radon-Nikodym) derivative of $F(x)$ with respect to a counting measure. A derivative is often called a density. Accordingly, we call these derivatives probability density functions.

Example 1. Let the random variable X of the discrete type have the p.d.f. $f(x) = x/6$, $x = 1, 2, 3$, zero elsewhere. The distribution function of X is

$$\begin{aligned} F(x) &= 0, & x < 1, \\ &= \tfrac{1}{6}, & 1 \leq x < 2, \\ &= \tfrac{3}{6}, & 2 \leq x < 3, \\ &= 1, & 3 \leq x. \end{aligned}$$

Here, as depicted in Figure 1.3, $F(x)$ is a step function which is constant in every interval not containing 1, 2, or 3, but has steps of heights $\frac{1}{6}$, $\frac{2}{6}$, and $\frac{3}{6}$ at those respective points. It is also seen that $F(x)$ is everywhere continuous to the right.

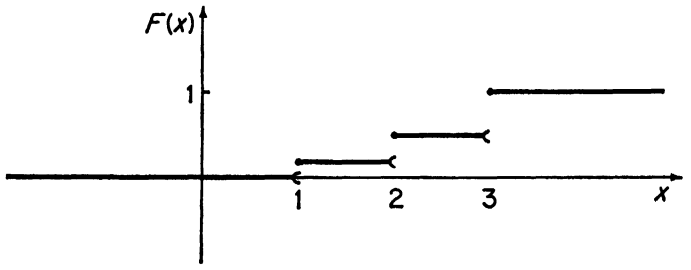


FIGURE 1.3

Example 2. Let the random variable X of the continuous type have the p.d.f. $f(x) = 2/x^3$, $1 < x < \infty$, zero elsewhere. The distribution function of X is

$$F(x) = \int_{-\infty}^{x} 0\, dw = 0, \qquad x < 1,$$

$$= \int_{1}^{x} \frac{2}{w^3}\, dw = 1 - \frac{1}{x^2}, \qquad 1 \leq x.$$

The graph of this distribution function is depicted in Figure 1.4. Here $F(x)$

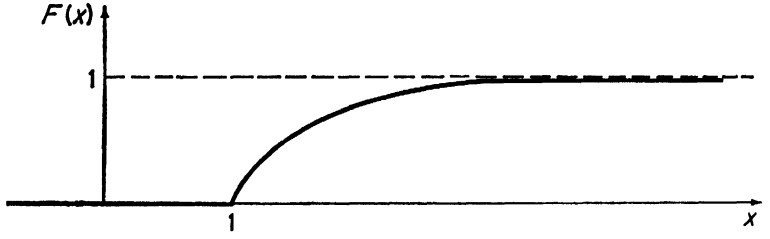


FIGURE 1.4

is a continuous function for all real numbers x; in particular, $F(x)$ is everywhere continuous to the right. Moreover, the derivative of $F(x)$ with respect to x exists at all points except at $x = 1$. Thus the p.d.f. of X is defined by this derivative except at $x = 1$. Since the set $A = \{x; x = 1\}$ is a set of probability measure zero [that is, $P(A) = 0$], we are free to define the p.d.f. at $x = 1$ in any manner we please. One way to do this is to write $f(x) = 2/x^3$, $1 < x < \infty$, zero elsewhere.

There are several properties of a distribution function $F(x)$ that can be listed as a consequence of the properties of the probability set function. Some of these are the following. In listing these properties, we shall not restrict X to be a random variable of the discrete or continuous type. We shall use the symbols $F(\infty)$ and $F(-\infty)$ to mean $\lim_{x \to \infty} F(x)$ and $\lim_{x \to -\infty} F(x)$ respectively. In like manner, the symbols $\{x; x \leq \infty\}$ and $\{x; x \leq -\infty\}$ represent, respectively, the limits of the sets $\{x; x \leq b\}$ and $\{x; x \leq -b\}$ as $b \to \infty$.

(a) $0 \leq F(x) \leq 1$ because $0 \leq \Pr(X \leq x) \leq 1$.

(b) $F(x)$ is a nondecreasing function of x. For, if $x' < x''$, then

$$\{x; x \leq x''\} = \{x; x \leq x'\} \cup \{x; x' < x \leq x''\}$$

and

$$\Pr(X \leq x'') = \Pr(X \leq x') + \Pr(x' < X \leq x'').$$

That is,

$$F(x'') - F(x') = \Pr(x' < X \leq x'') \geq 0.$$

(c) $F(\infty) = 1$ and $F(-\infty) = 0$ because the set $\{x; x \leq \infty\}$ is the entire one-dimensional space and the set $\{x; x \leq -\infty\}$ is the null set.

From the proof of (b), it is observed that, if $a < b$, then

$$\Pr(a < X \leq b) = F(b) - F(a).$$

Suppose we want to use $F(x)$ to compute the probability $\Pr(X = b)$. To do this, consider, with $h > 0$,

$$\lim_{h \to 0} \Pr(b - h < X \leq b) = \lim_{h \to 0} [F(b) - F(b - h)].$$

Intuitively, it seems that $\lim_{h \to 0} \Pr(b - h < X \leq b)$ should exist and be equal to $\Pr(X = b)$ because, as h tends to zero, the limit of the set $\{x; b - h < x \leq b\}$ is the set that contains the single point $x = b$. The fact that this limit is $\Pr(X = b)$ is a theorem that we accept without proof. Accordingly, we have

$$\Pr(X = b) = F(b) - F(b-),$$

where $F(b-)$ is the left-hand limit of $F(x)$ at $x = b$. That is, the probability that $X = b$ is the height of the step that $F(x)$ has at $x = b$. Hence, if the distribution function $F(x)$ is continuous at $x = b$, then $\Pr(X = b) = 0$.

There is a fourth property of $F(x)$ that is now listed.

(d) $F(x)$ is continuous to the right at each point x.

To prove this property, consider, with $h > 0$,

$$\lim_{h \to 0} \Pr(a < X \leq a + h) = \lim_{h \to 0} [F(a + h) - F(a)].$$

We accept without proof a theorem which states, with $h > 0$, that

$$\lim_{h \to 0} \Pr(a < X \leq a + h) = P(0) = 0.$$

Here also, the theorem is intuitively appealing because, as h tends to zero, the limit of the set $\{x; a < x \leq a + h\}$ is the null set. Accordingly, we write

$$0 = F(a+) - F(a),$$

where $F(a+)$ is the right-hand limit of $F(x)$ at $x = a$. Hence $F(x)$ is continuous to the right at every point $x = a$.

The preceding discussion may be summarized in the following manner: A distribution function $F(x)$ is a nondecreasing function of x, which is everywhere continuous to the right and has $F(-\infty) = 0$, $F(\infty) = 1$. The probability $\Pr(a < X \leq b)$ is equal to the difference $F(b) - F(a)$. If x is a discontinuity point of $F(x)$, then the probability $\Pr(X = x)$ is equal to the jump which the distribution function has at the point x. If x is a continuity point of $F(x)$, then $\Pr(X = x) = 0$.

Let X be a random variable of the continuous type that has p.d.f. $f(x)$, and let A be a set of probability measure zero; that is, $P(A) = \Pr(X \in A) = 0$. It has been observed that we may change the definition of $f(x)$ at any point in A without in any way altering the distribution of probability. The freedom to do this with the p.d.f. $f(x)$, of a continuous type of random variable does not extend to the distribution function $F(x)$; for, if $F(x)$ is changed at so much as one point x, the probability $\Pr(X \leq x) = F(x)$ is changed, and we have a different distribution of probability. That is, the distribution function $F(x)$, not the p.d.f. $f(x)$, is really the fundamental concept.

Remark. The definition of the distribution function makes it clear that the probability set function P determines the distribution function F. It is true, although not so obvious, that a probability set function P can be found from a distribution function F. That is, P and F give the same information about the distribution of probability and it is a matter of convenience as to which function is used.

We now give an illustrative example.

Example 3. Let a distribution function be given by

$$\begin{aligned} F(x) &= 0, && x < 0, \\ &= \frac{x+1}{2}, && 0 \le x < 1, \\ &= 1, && 1 \le x. \end{aligned}$$

Then, for instance,

$$\Pr(-3 < X \le \tfrac{1}{2}) = F(\tfrac{1}{2}) - F(-3) = \tfrac{3}{4} - 0 = \tfrac{3}{4}$$

and

$$\Pr(X = 0) = F(0) - F(0-) = \tfrac{1}{2} - 0 = \tfrac{1}{2}.$$

The graph of $F(x)$ is shown in Figure 1.5. We see that $F(x)$ is not always continuous nor is it a step function. Accordingly, the corresponding distribution is neither of the continuous type nor of the discrete type. It may be described as a mixture of those types.

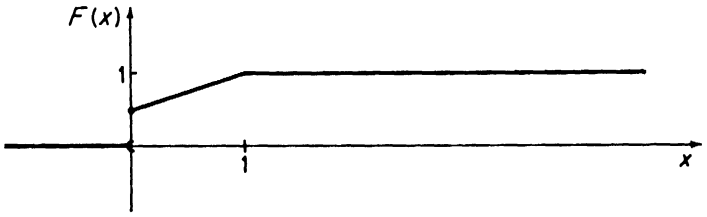


FIGURE 1.5

We shall now point out an important fact about a function of a random variable. Let X denote a random variable with space $\mathscr{A}$. Consider the function $Y = u(X)$ of the random variable X. Since X is a function defined on a sample space $\mathscr{C}$, then $Y = u(X)$ is a composite function defined on $\mathscr{C}$. That is, $Y = u(X)$ is itself a random variable which has its own space $\mathscr{B} = \{y; y = u(x), x \in \mathscr{A}\}$ and its own probability set function. If $y \in \mathscr{B}$, the event $Y = u(X) \le y$ occurs when, and only when, the event $X \in A \subset \mathscr{A}$ occurs, where $A = \{x; u(x) \le y\}$. That is, the distribution function of Y is

$$G(y) = \Pr(Y \le y) = \Pr[u(X) \le y] = P(A).$$

The following example illustrates a method of finding the distribution function and the p.d.f. of a function of a random variable.

Example 4. Let $f(x) = \frac{1}{2}$, $-1 < x < 1$, zero elsewhere, be the p.d.f. of the random variable X. Define the random variable Y by $Y = X^2$. We wish to find the p.d.f. of Y. If $y \ge 0$, the probability $\Pr(Y \le y)$ is equivalent to

$$\Pr(X^2 \le y) = \Pr(-\sqrt{y} \le X \le \sqrt{y}).$$

Accordingly, the distribution function of Y, $G(y) = \Pr(Y \leq y)$, is given by

$$\begin{aligned} G(y) &= 0, \qquad y < 0, \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \tfrac{1}{2}\, dx = \sqrt{y}, \qquad 0 \leq y < 1, \\ &= 1, \qquad 1 \leq y. \end{aligned}$$

Since Y is a random variable of the continuous type, the p.d.f. of Y is $g(y) = G'(y)$ at all points of continuity of $g(y)$. Thus we may write

$$\begin{aligned} g(y) &= \frac{1}{2\sqrt{y}}, \qquad 0 < y < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Let the random variables X and Y have the probability set function $P(A)$, where A is a two-dimensional set. If A is the unbounded set $\{(u, v); u \leq x, v \leq y\}$, where x and y are real numbers, we have

$$P(A) = \Pr[(X, Y) \in A] = \Pr(X \leq x, Y \leq y).$$

This function of the point (x, y) is called the *distribution function* of X and Y and is denoted by

$$F(x, y) = \Pr(X \leq x, Y \leq y).$$

If X and Y are random variables of the continuous type that have p.d.f. $f(x, y)$, then

$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v)\, du\, dv.$$

Accordingly, at points of continuity of $f(x, y)$, we have

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

It is left as an exercise to show, in every case, that

$$\Pr(a < X \leq b, c < Y \leq d) = F(b, d) - F(b, c) - F(a, d) + F(a, c),$$

for all real constants $a < b$, $c < d$.

The distribution function of the n random variables $X_1, X_2, \ldots, X_n$ is the point function

$$F(x_1, x_2, \ldots, x_n) = \Pr(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n).$$

An illustrative example follows.

Example 5. Let $f(x, y, z) = e^{-(x+y+z)}$, $0 < x, y, z < \infty$, zero elsewhere, be the p.d.f. of the random variables X, Y, and Z. Then the distribution function of X, Y, and Z is given by

$$\begin{aligned} F(x, y, z) &= \Pr(X \leq x, Y \leq y, Z \leq z) \\ &= \int_0^z \int_0^y \int_0^x e^{-u-v-w}\, du\, dv\, dw \\ &= (1 - e^{-x})(1 - e^{-y})(1 - e^{-z}), \qquad 0 \leq x, y, z < \infty, \end{aligned}$$

and is equal to zero elsewhere. Incidentally, except for a set of probability measure zero, we have

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

EXERCISES

1.44. Let $f(x)$ be the p.d.f. of a random variable X. Find the distribution function $F(x)$ of X and sketch its graph if:

(a) $f(x) = 1$, $x = 0$, zero elsewhere.
(b) $f(x) = \frac{1}{3}$, $x = -1, 0, 1$, zero elsewhere.
(c) $f(x) = x/15$, $x = 1, 2, 3, 4, 5$, zero elsewhere.
(d) $f(x) = 3(1 - x)^2$, $0 < x < 1$, zero elsewhere.
(e) $f(x) = 1/x^2$, $1 < x < \infty$, zero elsewhere.
(f) $f(x) = \frac{1}{3}$, $0 < x < 1$ or $2 < x < 4$, zero elsewhere.

1.45. Find the median of each of the distributions in Exercise 1.44.

1.46. Given the distribution function

$$\begin{aligned} F(x) &= 0, && x < -1, \\ &= \frac{x + 2}{4}, && -1 \leq x < 1, \\ &= 1, && 1 \leq x. \end{aligned}$$

Sketch the graph of $F(x)$ and then compute: (a) $\Pr(-\frac{1}{2} < X \leq \frac{1}{2})$; (b) $\Pr(X = 0)$; (c) $\Pr(X = 1)$; and (d) $\Pr(2 < X \leq 3)$.

1.47. Let $F(x, y)$ be the distribution function of X and Y. Show that $\Pr(a < X \leq b, c < Y \leq d) = F(b, d) - F(b, c) - F(a, d) + F(a, c)$, for all real constants $a < b$, $c < d$.

1.48. Let $f(x) = 1$, $0 < x < 1$, zero elsewhere, be the p.d.f. of X. Find the distribution function and the p.d.f. of $Y = \sqrt{X}$.

1.49. Let $f(x) = x/6$, $x = 1, 2, 3$, zero elsewhere, be the p.d.f. of X. Find the distribution function and the p.d.f. of $Y = X^2$. *Hint.* Note that X is a random variable of the discrete type.

1.50. Explain why, with $h > 0$, the two limits $\lim_{h \to 0} \Pr(b - h < X \leq b)$ and $\lim_{h \to 0} F(b - h)$ exist. *Hint.* Note that $\Pr(b - h < X \leq b)$ is bounded below by zero and $F(b - h)$ is bounded above by both $F(b)$ and one.

1.51. Show that the function $F(x, y)$ that is equal to one, provided $x + 2y \geq 1$, and that is equal to zero, provided $x + 2y < 1$, cannot be a distribution function of two random variables. *Hint.* Find four numbers $a < b, c < d$, so that $F(b, d) - F(a, d) - F(b, c) + F(a, c)$ is less than zero.

1.52. Let $F(x)$ be the distribution function of the random variable X. If m is a number such that $F(m) = \frac{1}{2}$, show that m is a median of the distribution.

1.53. Let $f(x) = \frac{1}{3}$, $-1 < x < 2$, zero elsewhere, be the p.d.f. of X. Find the distribution function and the p.d.f. of $Y = X^2$.

1.8 Certain Probability Models

Consider an experiment in which one chooses at random a point from the closed interval $[a, b]$ that is on the real line. Thus the sample space $\mathscr{C}$ is $[a, b]$. Let the random variable X be the identity function defined on $\mathscr{C}$. Thus the space $\mathscr{A}$ of X is $\mathscr{A} = \mathscr{C}$. Suppose it is reasonable to *assume*, from the nature of the experiment, that, if an interval A is a subset of $\mathscr{A}$, the probability of the event A is proportional to the length of A. Hence if A is the interval $[a, x]$, $x \leq b$, then

$$P(A) = \Pr(X \in A) = \Pr(a \leq X \leq x) = c(x - a),$$

where c is the constant of proportionality.

In the above expression, if we take $x = b$, we have

$$1 = \Pr(a \leq X \leq b) = c(b - a),$$

so $c = 1/(b - a)$. Thus we will have an appropriate probability model if we take the distribution function of X, $F(x) = \Pr(X \leq x)$, to be

$$\begin{aligned} F(x) &= 0, && x < a, \\ &= \frac{x - a}{b - a}, && a \leq x \leq b, \\ &= 1, && b < x. \end{aligned}$$

Accordingly, the p.d.f. of X, $f(x) = F'(x)$, may be written

$$\begin{aligned} f(x) &= \frac{1}{b - a}, && a \leq x \leq b, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

The derivative of $F(x)$ does not exist at $x = a$ nor at $x = b$; but the set $\{x; x = a, b\}$ is a set of probability measure zero, and we elect to define $f(x)$ to be equal to $1/(b - a)$ at those two points, just as a matter of convenience. We observe that this p.d.f. is a constant on $\mathscr{A}$. If the p.d.f. of one or more variables of the continuous type or of the discrete type is a constant on the space $\mathscr{A}$, we say that the probability is distributed *uniformly* over $\mathscr{A}$. Thus, in the above example, we say that X has a *uniform distribution* over the interval $[a, b]$.

Consider next an experiment in which one chooses at random a point (X, Y) from the unit square $\mathscr{C} = \mathscr{A} = \{(x, y);\ 0 < x < 1, 0 < y < 1\}$. Suppose that our interest is not in X nor in Y but in $Z = X + Y$. Once a suitable probability model has been adopted, we shall see how to find the p.d.f. of Z. To be specific, let the nature of the random experiment be such that it is reasonable to *assume* that the distribution of probability over the unit square is uniform. Then the p.d.f. of X and Y may be written

$$\begin{aligned} f(x, y) &= 1, \qquad 0 < x < 1, 0 < y < 1, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

and this describes the probability model. Now let the distribution function of Z be denoted by $G(z) = \Pr(X + Y \leq z)$. Then

$$\begin{aligned} G(z) &= 0, \qquad z < 0, \\ &= \int_0^z \int_0^{z-x} dy\, dx = \frac{z^2}{2}, \qquad 0 \leq z < 1, \\ &= 1 - \int_{z-1}^1 \int_{z-x}^1 dy\, dx = 1 - \frac{(2 - z)^2}{2}, \qquad 1 \leq z < 2, \\ &= 1, \qquad 2 \leq z. \end{aligned}$$

Since $G'(z)$ exists for all values of z, the p.d.f. of Z may then be written

$$\begin{aligned} g(z) &= z, \qquad 0 < z < 1, \\ &= 2 - z, \qquad 1 \leq z < 2, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

It is clear that a different choice of the p.d.f. $f(x, y)$ that describes the probability model will, in general, lead to a different p.d.f. of Z.

We wish presently to extend and generalize some of the notions expressed in the next three sentences. Let the discrete type of random variable X have a uniform distribution of probability over the k points of the space $\mathscr{A} = \{x; x = 1, 2, \ldots, k\}$. The p.d.f. of X is then $f(x) = 1/k, x \in \mathscr{A}$, zero elsewhere. This type of p.d.f. is used to describe the

probability model when each of the k points has the same probability, namely, $1/k$.

The probability model described in the preceding paragraph will now be adapted to a more general situation. Let a probability set function $P(C)$ be defined on a sample space $\mathscr{C}$. Here $\mathscr{C}$ may be a set in one, or two, or more dimensions. Let $\mathscr{C}$ be partitioned into k mutually disjoint subsets $C_1, C_2, \ldots, C_k$ in such a way that the union of these k mutually disjoint subsets is the sample space $\mathscr{C}$. Thus the events $C_1, C_2, \ldots, C_k$ are mutually exclusive and exhaustive. Suppose that the random experiment is of such a character that it may be *assumed* that each of the mutually exclusive and exhaustive events $C_i, i = 1, 2, \ldots, k$, has the same probability. Necessarily then, $P(C_i) = 1/k, i = 1, 2, \ldots, k$. Let the event E be the union of r of these mutually exclusive events, say

$$E = C_1 \cup C_2 \cup \cdots \cup C_r, \qquad r \leq k.$$

Then

$$P(E) = P(C_1) + P(C_2) + \cdots + P(C_r) = \frac{r}{k}.$$

Frequently, the integer k is called the total number of ways (for this particular partition of $\mathscr{C}$) in which the random experiment can terminate and the integer r is called the number of ways that are favorable to the event E. So, in this terminology, $P(E)$ is equal to the number of ways favorable to the event E divided by the total number of ways in which the experiment can terminate. It should be emphasized that in order to assign, *in this manner*, the probability r/k to the event E, we must assume that each of the mutually exclusive and exhaustive events $C_1, C_2, \ldots, C_k$ has the same probability $1/k$. This assumption then becomes *part of* our probability model. Obviously, if this assumption is not realistic in an application, the probability of the event E cannot be computed in this way.

We next present two examples that are illustrative of this model.

Example 1. Let a card be drawn at random from an ordinary deck of 52 playing cards. The sample space $\mathscr{C}$ is the union of $k = 52$ outcomes, and it is reasonable to assume that each of these outcomes has the same probability $\frac{1}{52}$. Accordingly, if E_1 is the set of outcomes that are spades, $P(E_1) = \frac{13}{52} = \frac{1}{4}$ because there are $r_1 = 13$ spades in the deck; that is, $\frac{1}{4}$ is the probability of drawing a card that is a spade. If E_2 is the set of outcomes that are kings, $P(E_2) = \frac{4}{52} = \frac{1}{13}$ because there are $r_2 = 4$ kings in the deck; that is, $\frac{1}{13}$ is the probability of drawing a card that is a king. These computations are very

easy because there are no difficulties in the determination of the appropriate values of r and k. However, instead of drawing only one card, suppose five cards are taken, at random and without replacement, from this deck. We can think of each five-card hand as being an outcome in a sample space. It is reasonable to assume that each of these outcomes has the same probability. Now if E_1 is the set of outcomes in which each card of the hand is a spade, $P(E_1)$ is equal to the number r_1 of all spade hands divided by the total number, say k, of five-card hands. It is shown in many books on algebra that

$$r_1 = \binom{13}{5} = \frac{13!}{5!\,8!} \qquad \text{and} \qquad k = \binom{52}{5} = \frac{52!}{5!\,47!}.$$

In general, if n is a positive integer and if x is a nonnegative integer with $x \leq n$, then the binomial coefficient

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

is equal to the number of combinations of n things taken x at a time. Thus, here,

$$P(E_1) = \frac{\binom{13}{5}}{\binom{52}{5}} = \frac{(13)(12)(11)(10)(9)}{(52)(51)(50)(49)(48)} = 0.0005,$$

approximately. Next, let E_2 be the set of outcomes in which at least one card is a spade. Then E_2^* is the set of outcomes in which no card is a spade. There are $r_2^* = \binom{39}{5}$ such outcomes. Hence

$$P(E_2^*) = \frac{\binom{39}{5}}{\binom{52}{5}} \qquad \text{and} \qquad P(E_2) = 1 - P(E_2^*).$$

Now suppose E_3 is the set of outcomes in which exactly three cards are kings and exactly two cards are queens. We can select the three kings in any one of $\binom{4}{3}$ ways and the two queens in any one of $\binom{4}{2}$ ways. By a well-known counting principle, the number of outcomes in E_3 is $r_3 = \binom{4}{3}\binom{4}{2}$. Thus $P(E_3) = \binom{4}{3}\binom{4}{2}\Big/\binom{52}{5}$. Finally, let E_4 be the set of outcomes in which there are exactly two kings, two queens, and one jack. Then

$$P(E_4) = \frac{\binom{4}{2}\binom{4}{2}\binom{4}{1}}{\binom{52}{5}},$$

because the numerator of this fraction is the number of outcomes in E_4.

Example 2. A lot, consisting of 100 fuses, is inspected by the following procedure. Five of these fuses are chosen at random and tested; if all 5 "blow" at the correct amperage, the lot is accepted. If, in fact, there are 20 defective fuses in the lot, the probability of accepting the lot is, under appropriate assumptions,

$$\frac{\binom{80}{5}}{\binom{100}{5}} = 0.32,$$

approximately. More generally, let the random variable X be the number of defective fuses among the 5 that are inspected. The space of X is $\mathscr{A} = \{x; x = 0, 1, 2, 3, 4, 5\}$ and the p.d.f. of X is given by

$$f(x) = \Pr(X = x) = \frac{\binom{20}{x}\binom{80}{5-x}}{\binom{100}{5}}, \qquad x = 0, 1, 2, 3, 4, 5,$$
$$= 0 \text{ elsewhere.}$$

This is an example of a discrete type of distribution which is called a *hypergeometric distribution.*

EXERCISES

(In order to solve some of these exercises, the student must make certain assumptions.)

1.54. A bowl contains 16 chips, of which 6 are red, 7 are white, and 3 are blue. If 4 chips are taken at random and without replacement, find the probability: (a) that each of the 4 chips is red; (b) that none of the 4 chips is red; and (c) that there is at least 1 chip of each color.

1.55. A person has purchased 10 of 1000 tickets sold in a certain raffle. To determine the five prize winners, 5 tickets are to be drawn at random and without replacement. Compute the probability that this person will win at least one prize.

1.56. Compute the probability of being dealt at random and without replacement a 13-card bridge hand consisting of: (a) 6 spades, 4 hearts, 2 diamonds, and 1 club; (b) 13 cards of the same suit.

1.57. Three distinct integers are chosen at random from the first 20 positive integers. Compute the probability that: (a) their sum is even; (b) their product is even.

1.58. There are five red chips and three blue chips in a bowl. The red chips are numbered 1, 2, 3, 4, 5, respectively, and the blue chips are numbered

1, 2, 3, respectively. If two chips are to be drawn at random and without replacement, find the probability that these chips have either the same number or the same color.

1.59. Let X have the uniform distribution given by the p.d.f. $f(x) = \frac{1}{5}$, $x = -2, -1, 0, 1, 2$, zero elsewhere. Find the p.d.f. of $Y = X^2$. *Hint.* Note that Y has a distribution of the discrete type.

1.60. Let X and Y have the p.d.f. $f(x, y) = 1$, $0 < x < 1$, $0 < y < 1$, zero elsewhere. Find the p.d.f. of the product $Z = XY$.

1.61. Let 13 cards be taken, at random and without replacement, from an ordinary deck of playing cards. If X is the number of spades in these 13 cards, find the p.d.f. of X. If, in addition, Y is the number of hearts in these 13 cards, find the probability $\Pr(X = 2, Y = 5)$. What is the p.d.f. of X and Y?

1.62. Four distinct integers are chosen at random and without replacement from the first 10 positive integers. Let the random variable X be the next to the smallest of these four numbers. Find the p.d.f. of X.

1.63. In a lot of 50 light bulbs, there are 2 bad bulbs. An inspector examines 5 bulbs, which are selected at random and without replacement. (a) Find the probability of at least 1 defective bulb among the 5. (b) How many bulbs should he examine so that the probability of finding at least 1 bad bulb exceeds $\frac{1}{2}$?

1.9 Mathematical Expectation

One of the more useful concepts in problems involving distributions of random variables is that of mathematical expectation. Let X be a random variable having a p.d.f. $f(x)$, and let $u(X)$ be a function of X such that

$$\int_{-\infty}^{\infty} u(x)f(x)\,dx$$

exists, if X is a continuous type of random variable, or such that

$$\sum_{x} u(x)f(x)$$

exists, if X is a discrete type of random variable. The integral, or the sum, as the case may be, is called the *mathematical expectation* (or expected value) of $u(X)$ and is denoted by $E[u(X)]$. That is,

$$E[u(X)] = \int_{-\infty}^{\infty} u(x)f(x)\,dx,$$

if X is a continuous type of random variable, or

$$E[u(X)] = \sum_x u(x)f(x),$$

if X is a discrete type of random variable.

Remarks. The usual definition of $E[u(X)]$ requires that the integral (or sum) converge absolutely. However, in this book, each $u(x)$ is of such a character that if the integral (or sum) exists, the convergence is absolute. Accordingly, we have not burdened the student with this additional provision.

The terminology "mathematical expectation" or "expected value" has its origin in games of chance. This can be illustrated as follows: Three small similar discs, numbered 1, 2, and 2, respectively, are placed in a bowl and are mixed. A player is to be blindfolded and is to draw a disc from the bowl. If he draws the disc numbered 1, he will receive nine dollars; if he draws either disc numbered 2, he will receive three dollars. It seems reasonable to assume that the player has a "$\frac{1}{3}$ claim" on the nine dollars and a "$\frac{2}{3}$ claim" on the three dollars. His "total claim" is $9(\frac{1}{3}) + 3(\frac{2}{3})$, or five dollars. If we take X to be a random variable having the p.d.f. $f(x) = x/3$, $x = 1, 2$, zero elsewhere, and $u(x) = 15 - 6x$, then $E[u(X)] = \sum_x u(x)f(x) = \sum_{x=1}^{2} (15 - 6x)(x/3) = 5$. That is, the mathematical expectation of $u(X)$ is precisely the player's "claim" or expectation.

The student may observe that $u(X)$ is a random variable Y with its own distribution of probability. Suppose the p.d.f. of Y is $g(y)$. Then $E(Y)$ is given by

$$\int_{-\infty}^{\infty} yg(y)\,dy \qquad \text{or} \qquad \sum_y yg(y),$$

according as Y is of the continuous type or of the discrete type. The question is: "Does this have the same value as $E[u(X)]$, which was defined above?" The answer to this question is in the affirmative, as will be shown in Chapter 4.

More generally, let $X_1, X_2, \ldots, X_n$ be random variables having p.d.f. $f(x_1, x_2, \ldots, x_n)$ and let $u(X_1, X_2, \ldots, X_n)$ be a function of these variables such that the n-fold integral

$$(1) \qquad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_1, x_2, \ldots, x_n)f(x_1, x_2, \ldots, x_n)\,dx_1\,dx_2 \cdots dx_n$$

exists, if the random variables are of the continuous type, or such that the n-fold sum

$$(2) \qquad \sum_{x_n} \cdots \sum_{x_1} u(x_1, x_2, \ldots, x_n)f(x_1, x_2, \ldots, x_n)$$

exists if the random variables are of the discrete type. The n-fold integral (or the n-fold sum, as the case may be) is called the *mathematical expectation*, denoted by $E[u(X_1, X_2, \ldots, X_n)]$, of the function $u(X_1, X_2, \ldots, X_n)$.

In this paragraph we shall point out some fairly obvious but useful facts about mathematical expectations when they exist.

(a) If k is a constant, then $E(k) = k$. This follows from expression (1) [or (2)] upon setting $u = k$ and recalling that an integral (or sum) of a constant times a function is the constant times the integral (or sum) of the function. Of course, the integral (or sum) of the function f is one.

(b) If k is a constant and v is a function, then $E(kv) = kE(v)$. This follows from expression (1) [or (2)] upon setting $u = kv$ and rewriting expression (1) [or (2)] as k times the integral (or sum) of the product vf.

(c) If k_1 and k_2 are constants and v_1 and v_2 are functions, then $E(k_1v_1 + k_2v_2) = k_1E(v_1) + k_2E(v_2)$. This, too, follows from expression (1) [or (2)] upon setting $u = k_1v_1 + k_2v_2$ because the integral (or sum) of $(k_1v_1 + k_2v_2)f$ is equal to the integral (or sum) of k_1v_1f plus the integral (or sum) of k_2v_2f. Repeated application of this property shows that if $k_1, k_2, \ldots, k_m$ are constants and $v_1, v_2, \ldots, v_m$ are functions, then

$$E(k_1v_1 + k_2v_2 + \cdots + k_mv_m) = k_1E(v_1) + k_2E(v_2) + \cdots + k_mE(v_m).$$

This property of mathematical expectation leads us to characterize the symbol E as a linear operator.

Example 1. Let X have the p.d.f.

$$\begin{aligned} f(x) &= 2(1 - x), \qquad 0 < x < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Then

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_0^1 (x)2(1 - x)\,dx = \tfrac{1}{3},$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2f(x)\,dx = \int_0^1 (x^2)2(1 - x)\,dx = \tfrac{1}{6},$$

and, of course,

$$E(6X + 3X^2) = 6(\tfrac{1}{3}) + 3(\tfrac{1}{6}) = \tfrac{5}{2}.$$

Example 2. Let X have the p.d.f.

$$\begin{aligned} f(x) &= \frac{x}{6}, \qquad x = 1, 2, 3, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Then

$$\begin{aligned} E(X^3) &= \sum_x x^3f(x) = \sum_{x=1}^{3} x^3\frac{x}{6} \\ &= \tfrac{1}{6} + \tfrac{16}{6} + \tfrac{81}{6} = \tfrac{98}{6}. \end{aligned}$$

Example 3. Let X and Y have the p.d.f.

$$f(x, y) = x + y, \qquad 0 < x < 1, 0 < y < 1,$$
$$= 0 \text{ elsewhere.}$$

Accordingly,

$$E(XY^2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy^2 f(x, y)\, dx\, dy$$
$$= \int_0^1 \int_0^1 xy^2(x + y)\, dx\, dy$$
$$= \tfrac{17}{72}.$$

Example 4. Let us divide, at random, a horizontal line segment of length five into two parts. If X is the length of the left-hand part, it is reasonable to assume that X has the p.d.f.

$$f(x) = \tfrac{1}{5}, \qquad 0 < x < 5,$$
$$= 0 \text{ elsewhere.}$$

The expected value of the length X is $E(X) = \frac{5}{2}$ and the expected value of the length $5 - X$ is $E(5 - X) = \frac{5}{2}$. But the expected value of the product of the two lengths is equal to

$$E[X(5 - X)] = \int_0^5 x(5 - x)(\tfrac{1}{5})\, dx = \tfrac{25}{6} \neq (\tfrac{5}{2})^2.$$

That is, in general, the expected value of a product is not equal to the product of the expected values.

Example 5. A bowl contains five chips which cannot be distinguished by a sense of touch alone. Three of the chips are marked one dollar each and the remaining two are marked four dollars each. A player is blindfolded and draws, at random and without replacement, two chips from the bowl. The player is paid an amount equal to the sum of the values of the two chips that he draws and the game is over. If it costs four dollars and seventy-five cents to play this game, would we care to participate for any protracted period of time? Because we are unable to distinguish the chips by sense of touch, we assume that each of the 10 pairs that can be drawn has the same probability of being drawn. Let the random variable X be the number of chips, of the two to be chosen, that are marked one dollar. Then, under our assumption, X has the hypergeometric p.d.f.

$$f(x) = \frac{\binom{3}{x}\binom{2}{2 - x}}{\binom{5}{2}}, \qquad x = 0, 1, 2,$$
$$= 0 \text{ elsewhere.}$$

If $X = x$, the player receives $u(x) = x + 4(2 - x) = 8 - 3x$ dollars. Hence his mathematical expectation is equal to

$$E[8 - 3X] = \sum_{x=0}^{2} (8 - 3x)f(x) = \tfrac{44}{10},$$

or four dollars and forty cents.

EXERCISES

1.64. Let X have the p.d.f. $f(x) = (x + 2)/18$, $-2 < x < 4$, zero elsewhere. Find $E(X)$, $E[(X + 2)^3]$, and $E[6X - 2(X + 2)^3]$.

1.65. Suppose $f(x) = \frac{1}{5}$, $x = 1, 2, 3, 4, 5$, zero elsewhere, is the p.d.f. of the discrete type of random variable X. Compute $E(X)$ and $E(X^2)$. Use these two results to find $E[(X + 2)^2]$ by writing $(X + 2)^2 = X^2 + 4X + 4$.

1.66. If X and Y have the p.d.f. $f(x, y) = \frac{1}{3}$, $(x, y) = (0, 0), (0, 1), (1, 1)$, zero elsewhere, find $E[(X - \frac{1}{3})(Y - \frac{2}{3})]$.

1.67. Let the p.d.f. of X and Y be $f(x, y) = e^{-x-y}$, $0 < x < \infty$, $0 < y < \infty$, zero elsewhere. Let $u(X, Y) = X$, $v(X, Y) = Y$, and $w(X, Y) = XY$. Show that $E[u(X, Y)] \cdot E[v(X, Y)] = E[w(X, Y)]$.

1.68. Let the p.d.f. of X and Y be $f(x, y) = 2$, $0 < x < y$, $0 < y < 1$, zero elsewhere. Let $u(X, Y) = X$, $v(X, Y) = Y$ and $w(X, Y) = XY$. Show that $E[u(X, Y)] \cdot E[v(X, Y)] \neq E[w(X, Y)]$.

1.69. Let X have a p.d.f. $f(x)$ that is positive at $x = -1, 0, 1$ and is zero elsewhere. (a) If $f(0) = \frac{1}{2}$, find $E(X^2)$. (b) If $f(0) = \frac{1}{2}$ and if $E(X) = \frac{1}{6}$, determine $f(-1)$ and $f(1)$.

1.70. A bowl contains 10 chips, of which 8 are marked two dollars each and 2 are marked five dollars each. Let a person choose, at random and without replacement, 3 chips from this bowl. If the person is to receive the sum of the resulting amounts, find his expectation.

1.71. Let X be a random variable of the continuous type that has p.d.f. $f(x)$. If m is the unique median of the distribution of X and b is a real constant, show that

$$E(|X - b|) = E(|X - m|) + 2\int_{m}^{b} (b - x)f(x)\,dx,$$

provided the expectations exist. For what value of b is $E(|X - b|)$ a minimum?

1.72. Let $f(x) = 2x$, $0 < x < 1$, zero elsewhere, be the p.d.f. of X. (a) Compute $E(\sqrt{X})$. (b) Find the distribution function and the p.d.f. of $Y = \sqrt{X}$. (c) Compute $E(Y)$ and compare this result with the answer which was obtained in part (a).

1.73. Two distinct integers are chosen at random and without replacement from the first six positive integers. Compute the expected value of the absolute value of the difference of these two numbers.

1.10 Some Special Mathematical Expectations

Certain mathematical expectations, if they exist, have special names and symbols to represent them. We shall mention now only those associated with one random variable. First, let $u(X) = X$, where X is a random variable of the discrete type having a p.d.f. $f(x)$. Then

$$E(X) = \sum_x xf(x).$$

If the discrete points of the space of positive probability density are $a_1, a_2, a_3, \ldots$, then

$$E(X) = a_1f(a_1) + a_2f(a_2) + a_3f(a_3) + \cdots.$$

This sum of products is seen to be a "weighted average" of the values $a_1, a_2, a_3, \ldots$, the "weight" associated with each a_i being $f(a_i)$. This suggests that we call $E(X)$ the arithmetic mean of the values of X, or, more simply, the *mean value* of X (or the mean value of the distribution).

The mean value μ of a random variable X is defined, when it exists, to be $\mu = E(X)$, where X is a random variable of the discrete or of the continuous type.

Another special mathematical expectation is obtained by taking $u(X) = (X - \mu)^2$. If, initially, X is a random variable of the discrete type having a p.d.f. $f(x)$, then

$$\begin{aligned} E[(X - \mu)^2] &= \sum_x (x - \mu)^2 f(x) \\ &= (a_1 - \mu)^2 f(a_1) + (a_2 - \mu)^2 f(a_2) + \cdots, \end{aligned}$$

if $a_1, a_2, \ldots$ are the discrete points of the space of positive probability density. This sum of products may be interpreted as a "weighted average" of the squares of the deviations of the numbers $a_1, a_2, \ldots$ from the mean value μ of those numbers where the "weight" associated with each $(a_i - \mu)^2$ is $f(a_i)$. This mean value of the square of the deviation of X from its mean value μ is called the *variance* of X (or the variance of the distribution).

The variance of X will be denoted by σ^2, and we define σ^2, if it exists, by $\sigma^2 = E[(X - \mu)^2]$, whether X is a discrete or a continuous type of random variable.

It is worthwhile to observe that

$$\sigma^2 = E[(X - \mu)^2] = E(X^2 - 2\mu X + \mu^2);$$

and since E is a linear operator,

$$\begin{aligned}\sigma^2 &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2.\end{aligned}$$

This frequently affords an easier way of computing the variance of X.

It is customary to call σ (the positive square root of the variance) the *standard deviation* of X (or the standard deviation of the distribution). The number σ is sometimes interpreted as a measure of the dispersion of the points of the space relative to the mean value μ. We note that if the space contains only one point x for which $f(x) > 0$, then $\sigma = 0$.

Remark. Let the random variable X of the continuous type have the p.d.f. $f(x) = 1/(2a)$, $-a < x < a$, zero elsewhere, so that $\sigma = a/\sqrt{3}$ is the standard deviation of the distribution of X. Next, let the random variable Y of the continuous type have the p.d.f. $g(y) = 1/(4a)$, $-2a < y < 2a$, zero elsewhere, so that $\sigma = 2a/\sqrt{3}$ is the standard deviation of the distribution of Y. Here the standard deviation of Y is greater than that of X; this reflects the fact that the probability for Y is more widely distributed (relative to the mean zero) than is the probability for X.

We next define a third special mathematical expectation called the *moment-generating function* of a random variable X. Suppose there is a positive number h such that for $-h < t < h$ the mathematical expectation $E(e^{tX})$ exists. Thus

$$E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx,$$

if X is a continuous type of random variable, or

$$E(e^{tX}) = \sum_{x} e^{tx} f(x),$$

if X is a discrete type of random variable. This expectation is called the moment-generating function of X (or of the distribution) and is denoted by $M(t)$. That is,

$$M(t) = E(e^{tX}).$$

It is evident that if we set $t = 0$, we have $M(0) = 1$. As will be seen by example, not every distribution has a moment-generating function.

But it is difficult to overemphasize the importance of a moment-generating function when it does exist. This importance stems from the fact that the moment-generating function is unique and completely determines the distribution of the random variable; thus, if two random variables have the same moment-generating function, they have the same distribution. This property of a moment-generating function will be very useful in subsequent chapters. Proof of the uniqueness of the moment-generating function is based on the theory of transforms in analysis, and therefore we merely assert this uniqueness.

Although the fact that a moment-generating function (when it exists) completely determines a distribution of one random variable will not be proved, it does seem desirable to try to make the assertion plausible. This can be done if the random variable is of the discrete type. For example, let it be given that

$$M(t) = \tfrac{1}{10}e^t + \tfrac{2}{10}e^{2t} + \tfrac{3}{10}e^{3t} + \tfrac{4}{10}e^{4t}$$

is, for all real values of t, the moment-generating function of a random variable X of the discrete type. If we let $f(x)$ be the p.d.f. of X and let $a, b, c, d, \ldots$ be the discrete points in the space of X at which $f(x) > 0$, then

$$M(t) = \sum_x e^{tx} f(x),$$

or

$$\tfrac{1}{10}e^t + \tfrac{2}{10}e^{2t} + \tfrac{3}{10}e^{3t} + \tfrac{4}{10}e^{4t} = f(a)e^{at} + f(b)e^{bt} + \cdots.$$

Because this is an identity for all real values of t, it seems that the right-hand member should consist of but four terms and that each of the four should equal, respectively, one of those in the left-hand member; hence we may take $a = 1$, $f(a) = \frac{1}{10}$; $b = 2$, $f(b) = \frac{2}{10}$; $c = 3$, $f(c) = \frac{3}{10}$; $d = 4$, $f(d) = \frac{4}{10}$. Or, more simply, the p.d.f. of X is

$$f(x) = \frac{x}{10}, \qquad x = 1, 2, 3, 4,$$
$$= 0 \text{ elsewhere.}$$

On the other hand, let X be a random variable of the continuous type and let it be given that

$$M(t) = \frac{1}{(1-t)^2}, \qquad t < 1,$$

is the moment-generating function of X. That is, we are given

$$\frac{1}{(1-t)^2} = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx, \qquad t < 1.$$

It is not at all obvious how $f(x)$ is found. However, it is easy to see that a distribution with p.d.f.

$$\begin{aligned} f(x) &= xe^{-x}, \qquad 0 < x < \infty, \\ &= 0 \text{ elsewhere} \end{aligned}$$

has the moment-generating function $M(t) = (1 - t)^{-2}$, $t < 1$. Thus the random variable X has a distribution with this p.d.f. in accordance with the assertion of the uniqueness of the moment-generating function.

Since a distribution that has a moment-generating function $M(t)$ is completely determined by $M(t)$, it would not be surprising if we could obtain some properties of the distribution directly from $M(t)$. For example, the existence of $M(t)$ for $-h < t < h$ implies that derivatives of all order exist at $t = 0$. Thus

$$\frac{dM(t)}{dt} = M'(t) = \int_{-\infty}^{\infty} xe^{tx}f(x)\,dx,$$

if X is of the continuous type, or

$$\frac{dM(t)}{dt} = M'(t) = \sum_{x} xe^{tx}f(x),$$

if X is of the discrete type. Upon setting $t = 0$, we have in either case

$$M'(0) = E(X) = \mu.$$

The second derivative of $M(t)$ is

$$M''(t) = \int_{-\infty}^{\infty} x^2e^{tx}f(x)\,dx, \qquad \text{or} \qquad \sum_{x} x^2e^{tx}f(x),$$

so that $M''(0) = E(X^2)$. Accordingly,

$$\sigma^2 = E(X^2) - \mu^2 = M''(0) - [M'(0)]^2.$$

For example, if $M(t) = (1 - t)^{-2}$, $t < 1$, as in the illustration above, then

$$M'(t) = 2(1 - t)^{-3}$$

and

$$M''(t) = 6(1 - t)^{-4}.$$

Hence

$$\mu = M'(0) = 2$$

and

$$\sigma^2 = M''(0) - \mu^2 = 6 - 4 = 2.$$

Of course we could have computed μ and σ^2 from the p.d.f. by

$$\mu = \int_{-\infty}^{\infty} xf(x)\,dx \qquad \text{and} \qquad \sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2,$$

respectively. Sometimes one way is easier than the other.

In general, if m is a positive integer and if $M^{(m)}(t)$ means the mth derivative of $M(t)$, we have, by repeated differentiation with respect to t,

$$M^{(m)}(0) = E(X^m).$$

Now

$$E(X^m) = \int_{-\infty}^{\infty} x^m f(x)\,dx, \qquad \text{or} \qquad \sum_x x^m f(x),$$

and integrals (or sums) of this sort are, in mechanics, called *moments*. Since $M(t)$ generates the values of $E(X^m)$, $m = 1, 2, 3, \ldots$, it is called the moment-generating function. In fact, we shall sometimes call $E(X^m)$ the mth moment of the distribution, or the mth moment of X.

Example 1. Let X have the p.d.f.

$$\begin{aligned} f(x) &= \tfrac{1}{2}(x + 1), \qquad -1 < x < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Then the mean value of X is

$$\mu = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{-1}^{1} x\frac{x+1}{2}\,dx = \tfrac{1}{3}$$

while the variance of X is

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2 = \int_{-1}^{1} x^2\frac{x+1}{2}\,dx - (\tfrac{1}{3})^2 = \tfrac{2}{9}.$$

Example 2. If X has the p.d.f.

$$\begin{aligned} f(x) &= \frac{1}{x^2}, \qquad 1 < x < \infty, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

then the mean value of X does not exist, since

$$\begin{aligned} \int_1^{\infty} x\frac{1}{x^2}\,dx &= \lim_{b\to\infty} \int_1^b \frac{1}{x}\,dx \\ &= \lim_{b\to\infty} [\ln b - \ln 1] \end{aligned}$$

does not exist.

Example 3. Given that the series

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \cdots$$

converges to $\pi^2/6$. Then

$$f(x) = \frac{6}{\pi^2 x^2}, \qquad x = 1, 2, 3, \ldots,$$
$$= 0 \text{ elsewhere},$$

is the p.d.f. of a discrete type of random variable X. The moment-generating function of this distribution, if it exists, is given by

$$M(t) = E(e^{tX}) = \sum_x e^{tx} f(x)$$
$$= \sum_{x=1}^{\infty} \frac{6e^{tx}}{\pi^2 x^2}.$$

The ratio test may be used to show that this series diverges if $t > 0$. Thus there does not exist a positive number h such that $M(t)$ exists for $-h < t < h$. Accordingly, the distribution having the p.d.f. $f(x)$ of this example does not have a moment-generating function.

Example 4. Let X have the moment-generating function $M(t) = e^{t^2/2}$, $-\infty < t < \infty$. We can differentiate $M(t)$ any number of times to find the moments of X. However, it is instructive to consider this alternative method. The function $M(t)$ is represented by the following MacLaurin's series.

$$e^{t^2/2} = 1 + \frac{1}{1!}\left(\frac{t^2}{2}\right) + \frac{1}{2!}\left(\frac{t^2}{2}\right)^2 + \cdots + \frac{1}{k!}\left(\frac{t^2}{2}\right)^k + \cdots$$
$$= 1 + \frac{1}{2!}t^2 + \frac{(3)(1)}{4!}t^4 + \cdots + \frac{(2k-1)\cdots(3)(1)}{(2k)!}t^{2k} + \cdots.$$

In general, the MacLaurin's series for $M(t)$ is

$$M(t) = M(0) + \frac{M'(0)}{1!}t + \frac{M''(0)}{2!}t^2 + \cdots + \frac{M^{(m)}(0)}{m!}t^m + \cdots$$
$$= 1 + \frac{E(X)}{1!}t + \frac{E(X^2)}{2!}t^2 + \cdots + \frac{E(X^m)}{m!}t^m + \cdots.$$

Thus the coefficient of $(t^m/m!)$ in the MacLaurin's series representation of $M(t)$ is $E(X^m)$. So, for our particular $M(t)$, we have

$$E(X^{2k}) = (2k-1)(2k-3)\cdots(3)(1) = \frac{(2k)!}{2^k k!},$$

$k = 1, 2, 3, \ldots$, and $E(X^{2k-1}) = 0$, $k = 1, 2, 3, \ldots$.

Remark. In a more advanced course, we would not work with the moment-generating function because so many distributions do not have

moment-generating functions. Instead, we would let i denote the imaginary unit, t an arbitrary real, and we would define $E(e^{itX})$. This expectation exists for *every* distribution and it is called the *characteristic function* of the distribution. Every distribution has a unique characteristic function; and to each characteristic function there corresponds a unique distribution of probability. If X has a distribution with characteristic function $\varphi(t)$, then, for instance, if $E(X)$ and $E(X^2)$ exist, they are given, respectively, by $iE(X) = \varphi'(0)$ and $i^2E(X^2) = \varphi''(0)$. Students who are familiar with complex-valued functions may write $\varphi(t) = M(it)$ and, throughout this book, may prove certain theorems in complete generality.

EXERCISES

1.74. Find the mean and variance, if they exist, of each of the following distributions.

(a) $f(x) = \dfrac{3!}{x!\,(3-x)!}\left(\dfrac{1}{2}\right)^3$, $x = 0, 1, 2, 3$, zero elsewhere.

(b) $f(x) = 6x(1-x)$, $0 < x < 1$, zero elsewhere.

(c) $f(x) = 2/x^3$, $1 < x < \infty$, zero elsewhere.

1.75. Let $f(x) = (\frac{1}{2})^x$, $x = 1, 2, 3, \ldots$, zero elsewhere, be the p.d.f. of the random variable X. Find the moment-generating function, the mean, and the variance of X.

1.76. For each of the following probability density functions, compute $\Pr(\mu - 2\sigma < X < \mu + 2\sigma)$.

(a) $f(x) = 6x(1-x)$, $0 < x < 1$, zero elsewhere.

(b) $f(x) = (\frac{1}{2})^x$, $x = 1, 2, 3, \ldots$, zero elsewhere.

1.77. If the variance of the random variable X exists, show that $E(X^2) \geq [E(X)]^2$.

1.78. Let a random variable X of the continuous type have a p.d.f. $f(x)$ whose graph is symmetric with respect to $x = c$. If the mean value of X exists, show that $E(X) = c$.

1.79. Let the random variable X have mean μ, standard deviation σ, and moment-generating function $M(t)$, $-h < t < h$. Show that

$$E\left(\frac{X-\mu}{\sigma}\right) = 0, \qquad E\left[\left(\frac{X-\mu}{\sigma}\right)^2\right] = 1,$$

and

$$E\left\{\exp\left[t\left(\frac{X-\mu}{\sigma}\right)\right]\right\} = e^{-\mu t/\sigma}\,M\left(\frac{t}{\sigma}\right), \qquad -h\sigma < t < h\sigma.$$

1.80. Show that the moment-generating function of the random variable X having the p.d.f. $f(x) = \frac{1}{3}$, $-1 < x < 2$, zero elsewhere, is

$$M(t) = \frac{e^{2t} - e^{-t}}{3t}, \qquad t \neq 0,$$
$$= 1, \qquad t = 0.$$

1.81. Let X be a random variable such that $E[(X - b)^2]$ exists for all real b. Show that $E[(X - b)^2]$ is a minimum when $b = E(X)$.

1.82. Let $f(x_1, x_2) = 2x_1$, $0 < x_1 < 1$, $0 < x_2 < 1$, zero elsewhere, be the p.d.f. of X_1 and X_2. Compute $E(X_1 + X_2)$ and $E\{[X_1 + X_2 - E(X_1 + X_2)]^2\}$.

1.83. Let $f(x, y, z) = e^{-x-y-z}$, $0 < x < \infty$, $0 < y < \infty$, $0 < z < \infty$, zero elsewhere, be the p.d.f. of the random variables X, Y, and Z. Find $E(e^{t(X+Y+Z)})$, $t < 1$.

1.84. Let X be a random variable of the discrete type. Prove that the p.d.f. of X is of the form $f(x) = 1$, $x = b$, zero elsewhere, if and only if $E[(X - b)^2] = 0$. A discrete distribution of this extreme form is called a *degenerate* distribution.

1.85. Let X be a random variable such that $K(t) = E(t^X)$ exists for all real values of t in a certain open interval that includes the point $t = 1$. Show that $K^{(m)}(1)$ is equal to the mth *factorial moment* $E[X(X - 1) \cdots (X - m + 1)]$.

1.86. Let X be a random variable. If m is a positive integer, the expectation $E[(X - b)^m]$, if it exists, is called the mth moment of the distribution about the point b. Let the first, second, and third moments of the distribution about the point 7 be 3, 11, and 15, respectively. Determine the mean μ of X, and then find the first, second, and third moments of the distribution about the point μ.

1.87. Let X be a random variable such that $R(t) = E(e^{t(X-b)})$ exists for $-h < t < h$. If m is a positive integer, show that $R^{(m)}(0)$ is equal to the mth moment of the distribution about the point b.

1.88. Let $\psi(t) = \ln M(t)$, where $M(t)$ is the moment-generating function of a distribution. Prove that $\psi'(0) = \mu$ and $\psi''(0) = \sigma^2$.

1.89. Find the mean and the variance of the distribution that has the distribution function

$$F(x) = 0, \qquad x < 0,$$
$$= \frac{x}{8}, \qquad 0 \leq x < 2,$$
$$= \frac{x^2}{16}, \qquad 2 \leq x < 4,$$
$$= 1, \qquad 4 \leq x.$$

1.90. Find the moments of the distribution that has moment-generating function $M(t) = (1 - t)^{-3}$, $t < 1$. *Hint.* Differentiate twice the series

$$(1 - t)^{-1} = 1 + t + t^2 + t^3 + \cdots, \qquad -1 < t < 1.$$

1.91. Let X be a random variable of the continuous type with p.d.f. $f(x)$ which is positive provided $0 < x < b < \infty$, and is equal to zero elsewhere. Show that

$$E(X) = \int_0^b [1 - F(x)]\, dx,$$

where $F(x)$ is the distribution function of X.

1.11 Chebyshev's Inequality

In this section we shall prove a theorem that enables us to find upper (or lower) bounds for certain probabilities. These bounds, however, are not necessarily close to the exact probabilities and, accordingly, we ordinarily do not use the theorem to approximate a probability. The principal uses of the theorem and a special case of it are in theoretical discussions.

Theorem 6. *Let $u(X)$ be a nonnegative function of the random variable X. If $E[u(X)]$ exists, then, for every positive constant c,*

$$\Pr [u(X) \geq c] \leq \frac{E[u(X)]}{c}.$$

Proof. The proof is given when the random variable X is of the continuous type; but the proof can be adapted to the discrete case if we replace integrals by sums. Let $A = \{x; u(x) \geq c\}$ and let $f(x)$ denote the p.d.f. of X. Then

$$E[u(X)] = \int_{-\infty}^{\infty} u(x)f(x)\, dx = \int_A u(x)f(x)\, dx + \int_{A^*} u(x)f(x)\, dx.$$

Since each of the integrals in the extreme right-hand member of the preceding equation is nonnegative, the left-hand member is greater than or equal to either of them. In particular,

$$E[u(X)] \geq \int_A u(x)f(x)\, dx.$$

However, if $x \in A$, then $u(x) \geq c$; accordingly, the right-hand member of the preceding inequality is not increased if we replace $u(x)$ by c. Thus,

$$E[u(X)] \geq c \int_A f(x)\, dx.$$

Since

$$\int_A f(x)\,dx = \Pr(X \in A) = \Pr[u(X) \geq c],$$

it follows that

$$E[u(X)] \geq c \Pr[u(X) \geq c],$$

which is the desired result.

The preceding theorem is a generalization of an inequality which is often called Chebyshev's inequality. This inequality will now be established.

Theorem 7. Chebyshev's Inequality. *Let the random variable X have a distribution of probability about which we assume only that there is a finite variance σ^2. This, of course, implies that there is a mean μ. Then for every $k > 0$,*

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof. In Theorem 6, take $u(X) = (X - \mu)^2$ and $c = k^2\sigma^2$. Then we have

$$\Pr[(X - \mu)^2 \geq k^2\sigma^2] \leq \frac{E[(X - \mu)^2]}{k^2\sigma^2}.$$

Since the numerator of the right-hand member of the preceding inequality is σ^2, the inequality may be written

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2},$$

which is the desired result. Naturally, we would take the positive number k to be greater than one to have an inequality of interest.

It is seen that the number $1/k^2$ is an upper bound for the probability $\Pr(|X - \mu| \geq k\sigma)$. In the following example, this upper bound and the exact value of the probability are compared in special instances.

Example 1. Let X have the p.d.f.

$$f(x) = \frac{1}{2\sqrt{3}}, \qquad -\sqrt{3} < x < \sqrt{3},$$
$$= 0 \text{ elsewhere.}$$

Here $\mu = 0$ and $\sigma^2 = 1$. If $k = \frac{3}{2}$, we have the exact probability

$$\Pr(|X - \mu| \geq k\sigma) = \Pr\left[|X| \geq \frac{3}{2}\right] = 1 - \int_{-3/2}^{3/2} \frac{1}{2\sqrt{3}}\,dx = 1 - \frac{\sqrt{3}}{2}.$$

By Chebyshev's inequality, the preceding probability has the upper bound $1/k^2 = \frac{4}{9}$. Since $1 - \sqrt{3}/2 = 0.134$, approximately, the exact probability in this case is considerably less than the upper bound $\frac{4}{9}$. If we take $k = 2$, we have the exact probability $\Pr(|X - \mu| \geq 2\sigma) = \Pr(|X| \geq 2) = 0$. This again is considerably less than the upper bound $1/k^2 = \frac{1}{4}$ provided by Chebyshev's inequality.

In each instance in the preceding example, the probability $\Pr(|X - \mu| \geq k\sigma)$ and its upper bound $1/k^2$ differ considerably. This suggests that this inequality might be made sharper. However, if we want an inequality that holds for every $k > 0$ and holds for all random variables having finite variance, such an improvement is impossible, as is shown by the following example.

Example 2. Let the random variable X of the discrete type have probabilities $\frac{1}{8}, \frac{6}{8}, \frac{1}{8}$ at the points $x = -1, 0, 1$, respectively. Here $\mu = 0$ and $\sigma^2 = \frac{1}{4}$. If $k = 2$, then $1/k^2 = \frac{1}{4}$ and $\Pr(|X - \mu| \geq k\sigma) = \Pr(|X| \geq 1) = \frac{1}{4}$. That is, the probability $\Pr(|X - \mu| \geq k\sigma)$ here attains the upper bound $1/k^2 = \frac{1}{4}$. Hence the inequality cannot be improved without further assumptions about the distribution of X.

EXERCISES

1.92. Let X be a random variable with mean μ and let $E[(X - \mu)^{2k}]$ exist. Show, with $c > 0$, that $\Pr(|X - \mu| \geq c) \leq E[(X - \mu)^{2k}]/c^{2k}$.

1.93. Let X be a random variable such that $\Pr(X \leq 0) = 0$ and let $\mu = E(X)$ exist. Show that $\Pr(X \geq 2\mu) \leq \frac{1}{2}$.

1.94. If X is a random variable such that $E(X) = 3$ and $E(X^2) = 13$, use Chebyshev's inequality to determine a lower bound for the probability $\Pr(-2 < X < 8)$.

1.95. Let X be a random variable with moment-generating function $M(t)$, $-h < t < h$. Prove that

$$\Pr(X \geq a) \leq e^{-at} M(t), \qquad 0 < t < h,$$

and that

$$\Pr(X \leq a) \leq e^{-at} M(t), \qquad -h < t < 0.$$

1.96. The moment-generating function of X exists for all real values of t and is given by

$$M(t) = \frac{e^t - e^{-t}}{2t}, \qquad t \neq 0, \quad M(0) = 1.$$

Use the results of the preceding exercise to show that $\Pr(X \geq 1) = 0$ and $\Pr(X \leq -1) = 0$. Note that here h is infinite.

Chapter 2

Conditional Probability and Stochastic Independence

2.1 Conditional Probability

In some random experiments, we are interested only in those outcomes that are elements of a subset C_1 of the sample space $\mathscr{C}$. This means, for our purposes, that the sample space is effectively the subset C_1. We are now confronted with the problem of defining a probability set function with C_1 as the "new" sample space.

Let the probability set function $P(C)$ be defined on the sample space $\mathscr{C}$ and let C_1 be a subset of $\mathscr{C}$ such that $P(C_1) > 0$. We agree to consider only those outcomes of the random experiment that are elements of C_1; in essence, then, we take C_1 to be a sample space. Let C_2 be another subset of $\mathscr{C}$. How, relative to the new sample space C_1, do we want to define the probability of the event C_2? Once defined, this probability is called the *conditional probability* of the event C_2, relative to the hypothesis of the event C_1; or, more briefly, the conditional probability of C_2, given C_1. Such a conditional probability is denoted by the symbol $P(C_2|C_1)$. We now return to the question that was raised about the definition of this symbol. Since C_1 is now the sample space, the only elements of C_2 that concern us are those, if any, that are also elements of C_1, that is, the elements of $C_1 \cap C_2$. It seems desirable, then, to define the symbol $P(C_2|C_1)$ in such a way that

$$P(C_1|C_1) = 1 \qquad \text{and} \qquad P(C_2|C_1) = P(C_1 \cap C_2|C_1).$$

Moreover, from a relative frequency point of view, it would seem logically inconsistent if we did not require that the ratio of the probabilities of the events $C_1 \cap C_2$ and C_1, relative to the space C_1, be the same as the

ratio of the probabilities of these events relative to the space $\mathscr{C}$; that is, we should have

$$\frac{P(C_1 \cap C_2|C_1)}{P(C_1|C_1)} = \frac{P(C_1 \cap C_2)}{P(C_1)}.$$

These three desirable conditions imply that the relation

$$P(C_2|C_1) = \frac{P(C_1 \cap C_2)}{P(C_1)}$$

is a suitable *definition* of the conditional probability of the event C_2, given the event C_1, provided $P(C_1) > 0$. Moreover, we have:

(a) $P(C_2|C_1) \geq 0$.

(b) $P(C_2 \cup C_3 \cup \cdots|C_1) = P(C_2|C_1) + P(C_3|C_1) + \cdots$, provided $C_2, C_3, \ldots$ are mutually disjoint sets.

(c) $P(C_1|C_1) = 1$.

Properties (a) and (c) are evident; proof of property (b) is left as an exercise. But these are precisely the conditions that a probability set function must satisfy. Accordingly, $P(C_2|C_1)$ is a probability set function, defined for subsets of C_1. It may be called the conditional probability set function, relative to the hypothesis C_1; or the conditional probability set function, given C_1. It should be noted that this conditional probability set function, given C_1, is defined at this time only when $P(C_1) > 0$.

We have now defined the concept of conditional probability for subsets C of a sample space $\mathscr{C}$. We wish to do the same kind of thing for subsets A of $\mathscr{A}$, where $\mathscr{A}$ is the space of one or more random variables defined on $\mathscr{C}$. Let P denote the probability set function of the induced probability on $\mathscr{A}$. If A_1 and A_2 are subsets of $\mathscr{A}$, the conditional probability of the event A_2, given the event A_1, is

$$P(A_2|A_1) = \frac{P(A_1 \cap A_2)}{P(A_1)}$$

provided $P(A_1) > 0$. This definition will apply to any space which has a probability set function assigned to it.

Example 1. A hand of 5 cards is to be dealt at random and without replacement from an ordinary deck of 52 playing cards. The conditional probability of an all-spade hand (C_2), relative to the hypothesis that there are at least 4 spades in the hand (C_1), is, since $C_1 \cap C_2 = C_2$,

$$P(C_2|C_1) = \frac{P(C_2)}{P(C_1)} = \frac{\binom{13}{5}\Big/\binom{52}{5}}{\left[\binom{13}{4}\binom{39}{1} + \binom{13}{5}\right]\Big/\binom{52}{5}}.$$

It is worth noting, if we let the random variable X equal the number of spades in a 5-card hand, that a reasonable probability model for X is given by the hypergeometric p.d.f.

$$f(x) = \frac{\binom{13}{x}\binom{39}{5-x}}{\binom{52}{5}}, \qquad x = 0, 1, 2, 3, 4, 5,$$
$$= 0 \text{ elsewhere.}$$

Accordingly, we can write $P(C_2|C_1) = \Pr(X = 5)/[\Pr(X = 4, 5)] = f(5)/[f(4) + f(5)]$.

From the definition of the conditional probability set function, we observe that

$$P(C_1 \cap C_2) = P(C_1)P(C_2|C_1).$$

This relation is frequently called the *multiplication rule* for probabilities. Sometimes, after considering the nature of the random experiment, it is possible to make reasonable assumptions so that both $P(C_1)$ and $P(C_2|C_1)$ can be assigned. Then $P(C_1 \cap C_2)$ can be computed under these assumptions. This will be illustrated in Examples 2 and 3.

Example 2. A bowl contains eight chips. Three of the chips are red and the remaining five are blue. Two chips are to be drawn successively, at random and without replacement. We want to compute the probability that the first draw results in a red chip (C_1) and that the second draw results in a blue chip (C_2). It is reasonable to assign the following probabilities:

$$P(C_1) = \tfrac{3}{8} \qquad \text{and} \qquad P(C_2|C_1) = \tfrac{5}{7}.$$

Thus, under these assignments, we have $P(C_1 \cap C_2) = (\frac{3}{8})(\frac{5}{7}) = \frac{15}{56}$.

Example 3. From an ordinary deck of playing cards, cards are to be drawn successively, at random and without replacement. The probability that the third spade appears on the sixth draw is computed as follows. Let C_1 be the event of two spades in the first five draws and let C_2 be the event of a spade on the sixth draw. Thus the probability that we wish to compute is $P(C_1 \cap C_2)$. It is reasonable to take

$$P(C_1) = \frac{\binom{13}{2}\binom{39}{3}}{\binom{52}{5}} \qquad \text{and} \qquad P(C_2|C_1) = \tfrac{11}{47}.$$

The desired probability $P(C_1 \cap C_2)$ is then the product of these two numbers. More generally, if $X + 3$ is the number of draws necessary to produce

exactly three spades, a reasonable probability model for the random variable X is given by the p.d.f.

$$f(x) = \left[\frac{\binom{13}{2}\binom{39}{x}}{\binom{52}{2+x}}\right]\left(\frac{11}{50-x}\right), \qquad x = 0, 1, 2, \ldots, 39,$$
$$= 0 \text{ elsewhere.}$$

Then the particular probability which we computed is $P(C_1 \cap C_2) = \Pr(X = 3) = f(3)$.

The multiplication rule can be extended to three or more events. In the case of three events, we have, by using the multiplication rule for two events,

$$\begin{aligned} P(C_1 \cap C_2 \cap C_3) &= P[(C_1 \cap C_2) \cap C_3] \\ &= P(C_1 \cap C_2)P(C_3|C_1 \cap C_2). \end{aligned}$$

But $P(C_1 \cap C_2) = P(C_1)P(C_2|C_1)$. Hence

$$P(C_1 \cap C_2 \cap C_3) = P(C_1)P(C_2|C_1)P(C_3|C_1 \cap C_2).$$

This procedure can be used to extend the multiplication rule to four or more events. The general formula for k events can be proved by mathematical induction.

Example 4. Four cards are to be dealt successively, at random and without replacement, from an ordinary deck of playing cards. The probability of receiving a spade, a heart, a diamond, and a club, in that order, is $(\frac{13}{52})(\frac{13}{51})(\frac{13}{50})(\frac{13}{49})$. This follows from the extension of the multiplication rule. In this computation, the assumptions that are involved seem clear.

EXERCISES

(In order to solve certain of these exercises, the student is required to make assumptions.)

2.1. If $P(C_1) > 0$ and if $C_2, C_3, C_4, \ldots$ are mutually disjoint sets, show that $P(C_2 \cup C_3 \cup \cdots | C_1) = P(C_2|C_1) + P(C_3|C_1) + \cdots$.

2.2. Prove that

$$P(C_1 \cap C_2 \cap C_3 \cap C_4) = P(C_1)P(C_2|C_1)P(C_3|C_1 \cap C_2)P(C_4|C_1 \cap C_2 \cap C_3).$$

2.3. A bowl contains eight chips. Three of the chips are red and five are blue. Four chips are to be drawn successively at random and without replacement. (a) Compute the probability that the colors alternate. (b) Compute the probability that the first blue chip appears on the third draw. (c) If

$X + 1$ is the number of draws needed to produce the first blue chip, determine the p.d.f. of X.

2.4. A hand of 13 cards is to be dealt at random and without replacement from an ordinary deck of playing cards. Find the conditional probability that there are at least three kings in the hand relative to the hypothesis that the hand contains at least two kings.

2.5. A drawer contains eight pairs of socks. If six socks are taken at random and without replacement, compute the probability that there is at least one matching pair among these six socks. *Hint.* Compute the probability that there is not a matching pair.

2.6. A bowl contains 10 chips. Four of the chips are red, 5 are white, and 1 is blue. If 3 chips are taken at random and without replacement, compute the conditional probability that there is 1 chip of each color relative to the hypothesis that there is exactly 1 red chip among the 3.

2.7. Let each of the mutually disjoint sets $C_1, \ldots, C_m$ have nonzero probability. If the set C is a subset of the union of $C_1, \ldots, C_m$, show that

$$P(C) = P(C_1)P(C|C_1) + \cdots + P(C_m)P(C|C_m).$$

If $P(C) > 0$, prove *Bayes' formula*:

$$P(C_i|C) = \frac{P(C_i)P(C|C_i)}{P(C_1)P(C|C_1) + \cdots + P(C_m)P(C|C_m)}, \qquad i = 1, \ldots, m.$$

Hint. $P(C)P(C_i|C) = P(C_i)P(C|C_i)$.

2.8. Bowl I contains 3 red chips and 7 blue chips. Bowl II contains 6 red chips and 4 blue chips. A bowl is selected at random and then 1 chip is drawn from this bowl. (a) Compute the probability that this chip is red. (b) Relative to the hypothesis that the chip is red, find the conditional probability that it is drawn from bowl II.

2.9. Bowl I contains 6 red chips and 4 blue chips. Five of these 10 chips are selected at random and without replacement and put in bowl II, which was originally empty. One chip is then drawn at random from bowl II. Relative to the hypothesis that this chip is blue, find the conditional probability that 2 red chips and 3 blue chips are transferred from bowl I to bowl II.

2.2 Marginal and Conditional Distributions

Let $f(x_1, x_2)$ be the p.d.f. of two random variables X_1 and X_2. From this point on, for emphasis and clarity, we shall call a p.d.f. or a distribution function a *joint* p.d.f. or a *joint* distribution function when more than one random variable is involved. Thus $f(x_1, x_2)$ is the joint

p.d.f. of the random variables X_1 and X_2. Consider the event $a < X_1 < b$, $a < b$. This event can occur when and only when the event $a < X_1 < b$, $-\infty < X_2 < \infty$ occurs; that is, the two events are equivalent, so that they have the same probability. But the probability of the latter event has been defined and is given by

$$\Pr(a < X_1 < b, -\infty < X_2 < \infty) = \int_a^b \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2\, dx_1$$

for the continuous case, and by

$$\Pr(a < X_1 < b, -\infty < X_2 < \infty) = \sum_{a < x_1 < b} \sum_{x_2} f(x_1, x_2)$$

for the discrete case. Now each of

$$\int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2 \qquad \text{and} \qquad \sum_{x_2} f(x_1, x_2)$$

is a function of x_1 alone, say $f_1(x_1)$. Thus, for every $a < b$, we have

$$\begin{aligned} \Pr(a < X_1 < b) &= \int_a^b f_1(x_1)\, dx_1 && \text{(continuous case)}, \\ &= \sum_{a < x_1 < b} f_1(x_1) && \text{(discrete case)}, \end{aligned}$$

so that $f_1(x_1)$ is the p.d.f. of X_1 alone. Since $f_1(x_1)$ is found by summing (or integrating) the joint p.d.f. $f(x_1, x_2)$ over all x_2 for a fixed x_1, we can think of recording this sum in the "margin" of the x_1x_2-plane. Accordingly, $f_1(x_1)$ is called the marginal p.d.f. of X_1. In like manner

$$\begin{aligned} f_2(x_2) &= \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1 && \text{(continuous case)}, \\ &= \sum_{x_1} f(x_1, x_2) && \text{(discrete case)}, \end{aligned}$$

is called the marginal p.d.f. of X_2.

Example 1. Let the joint p.d.f. of X_1 and X_2 be

$$\begin{aligned} f(x_1, x_2) &= \frac{x_1 + x_2}{21}, \qquad x_1 = 1, 2, 3, x_2 = 1, 2, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Then, for instance,

$$\Pr(X_1 = 3) = f(3, 1) + f(3, 2) = \tfrac{3}{7}$$

and

$$\Pr(X_2 = 2) = f(1, 2) + f(2, 2) + f(3, 2) = \tfrac{4}{7}.$$

On the other hand, the marginal p.d.f. of X_1 is

$$f_1(x_1) = \sum_{x_2=1}^{2} \frac{x_1 + x_2}{21} = \frac{2x_1 + 3}{21}, \qquad x_1 = 1, 2, 3,$$

zero elsewhere, and the marginal p.d.f. of X_2 is

$$f_2(x_2) = \sum_{x_1=1}^{3} \frac{x_1 + x_2}{21} = \frac{6 + 3x_2}{21}, \qquad x_2 = 1, 2,$$

zero elsewhere. Thus the preceding probabilities may be computed as $\Pr(X_1 = 3) = f_1(3) = \frac{3}{7}$ and $\Pr(X_2 = 2) = f_2(2) = \frac{4}{7}$.

We shall now discuss the notion of a conditional p.d.f. Let X_1 and X_2 denote random variables of the discrete type which have the joint p.d.f. $f(x_1, x_2)$ which is positive on $\mathscr{A}$ and is zero elsewhere. Let $f_1(x_1)$ and $f_2(x_2)$ denote, respectively, the marginal probability density functions of X_1 and X_2. Take A_1 to be the set $A_1 = \{(x_1, x_2); x_1 = x_1', -\infty < x_2 < \infty\}$, where x_1' is such that $P(A_1) = \Pr(X_1 = x_1') = f_1(x_1') > 0$, and take A_2 to be the set $A_2 = \{(x_1, x_2); -\infty < x_1 < \infty, x_2 = x_2'\}$. Then, by definition, the conditional probability of the event A_2, given the event A_1, is

$$P(A_2|A_1) = \frac{P(A_1 \cap A_2)}{P(A_1)} = \frac{\Pr(X_1 = x_1', X_2 = x_2')}{\Pr(X_1 = x_1')} = \frac{f(x_1', x_2')}{f_1(x_1')}.$$

That is, if (x_1, x_2) is any point at which $f_1(x_1) > 0$, the conditional probability that $X_2 = x_2$, given that $X_1 = x_1$, is $f(x_1, x_2)/f_1(x_1)$. With x_1 held fast, and with $f_1(x_1) > 0$, this function of x_2 satisfies the conditions of being a p.d.f. of a discrete type of random variable X_2 because $f(x_1, x_2)/f_1(x_1)$ is not negative and

$$\sum_{x_2} \frac{f(x_1, x_2)}{f_1(x_1)} = \frac{1}{f_1(x_1)} \sum_{x_2} f(x_1, x_2) = \frac{f_1(x_1)}{f_1(x_1)} = 1.$$

We now define the symbol $f(x_2|x_1)$ by the relation

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}, \qquad f_1(x_1) > 0,$$

and we call $f(x_2|x_1)$ the *conditional p.d.f.* of the discrete type of random variable X_2, given that the discrete type of random variable $X_1 = x_1$. In a similar manner we define the symbol $f(x_1|x_2)$ by the relation

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}, \qquad f_2(x_2) > 0,$$

and we call $f(x_1|x_2)$ the conditional p.d.f. of the discrete type of random variable X_1, given that the discrete type of random variable $X_2 = x_2$.

Now let X_1 and X_2 denote random variables of the continuous type that have the joint p.d.f. $f(x_1, x_2)$ and the marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, respectively. We shall use the results of the preceding paragraph to motivate a definition of a conditional p.d.f. of a continuous type of random variable. When $f_1(x_1) > 0$, we define the symbol $f(x_2|x_1)$ by the relation

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}.$$

In this relation, x_1 is to be thought of as having a fixed (but any fixed) value for which $f_1(x_1) > 0$. It is evident that $f(x_2|x_1)$ is nonnegative and that

$$\begin{aligned}\int_{-\infty}^{\infty} f(x_2|x_1)\,dx_2 &= \int_{-\infty}^{\infty} \frac{f(x_1, x_2)}{f_1(x_1)}\,dx_2 \\ &= \frac{1}{f_1(x_1)} \int_{-\infty}^{\infty} f(x_1, x_2)\,dx_2 \\ &= \frac{1}{f_1(x_1)} f_1(x_1) = 1.\end{aligned}$$

That is, $f(x_2|x_1)$ has the properties of a p.d.f. of one continuous type of random variable. It is called the *conditional p.d.f.* of the continuous type of random variable X_2, given that the continuous type of random variable X_1 has the value x_1. When $f_2(x_2) > 0$, the conditional p.d.f. of the continuous type of random variable X_1, given that the continuous type of random variable X_2 has the value x_2, is defined by

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}, \qquad f_2(x_2) > 0.$$

Since each of $f(x_2|x_1)$ and $f(x_1|x_2)$ is a p.d.f. of one random variable (whether of the discrete or the continuous type), each has all the properties of such a p.d.f. Thus, we can compute probabilities and mathematical expectations. If the random variables are of the continuous type, the probability

$$\Pr(a < X_2 < b|X_1 = x_1) = \int_a^b f(x_2|x_1)\,dx_2$$

is called "the conditional probability that $a < X_2 < b$, given that $X_1 = x_1$." If there is no ambiguity, this may be written in the form

$\Pr(a < X_2 < b|x_1)$. Similarly, the conditional probability that $c < X_1 < d$, given $X_2 = x_2$, is

$$\Pr(c < X_1 < d|X_2 = x_2) = \int_c^d f(x_1|x_2)\,dx_1.$$

If $u(X_2)$ is a function of X_2, the expectation

$$E[u(X_2)|x_1] = \int_{-\infty}^{\infty} u(x_2)f(x_2|x_1)\,dx_2$$

is called the conditional expectation of $u(X_2)$, given $X_1 = x_1$. In particular, if they exist, $E(X_2|x_1)$ is the mean and $E\{[X_2 - E(X_2|x_1)]^2|x_1\}$ is the variance of the conditional distribution of X_2, given $X_1 = x_1$. It is convenient to refer to these as the "conditional mean" and the "conditional variance" of X_2, given $X_1 = x_1$. Of course we have

$$E\{[X_2 - E(X_2|x_1)]^2|x_1\} = E(X_2^2|x_1) - [E(X_2|x_1)]^2$$

from an earlier result. In like manner, the conditional expectation of $u(X_1)$, given $X_2 = x_2$, is given by

$$E[u(X_1)|x_2] = \int_{-\infty}^{\infty} u(x_1)f(x_1|x_2)\,dx_1.$$

With random variables of the discrete type, these conditional probabilities and conditional expectations are computed by using summation instead of integration. An illustrative example follows.

Example 2. Let X_1 and X_2 have the joint p.d.f.

$$\begin{aligned} f(x_1, x_2) &= 2, \qquad 0 < x_1 < x_2 < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Then the marginal probability density functions are, respectively,

$$\begin{aligned} f_1(x_1) = \int_{x_1}^1 2\,dx_2 &= 2(1 - x_1), \qquad 0 < x_1 < 1, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

and

$$\begin{aligned} f_2(x_2) = \int_0^{x_2} 2\,dx_1 &= 2x_2, \qquad 0 < x_2 < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

The conditional p.d.f. of X_1, given $X_2 = x_2$, is

$$\begin{aligned} f(x_1|x_2) &= \frac{2}{2x_2} = \frac{1}{x_2}, \qquad 0 < x_1 < x_2,\ 0 < x_2 < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Here the conditional mean and conditional variance of X_1, given $X_2 = x_2$, are, respectively,

$$E(X_1|x_2) = \int_{-\infty}^{\infty} x_1 f(x_1|x_2)\, dx_1$$

$$= \int_0^{x_2} x_1 \left(\frac{1}{x_2}\right) dx_1$$

$$= \frac{x_2}{2}, \qquad 0 < x_2 < 1,$$

and

$$E\{[X_1 - E(X_1|x_2)]^2|x_2\} = \int_0^{x_2} \left(x_1 - \frac{x_2}{2}\right)^2 \left(\frac{1}{x_2}\right) dx_1$$

$$= \frac{x_2^2}{12}, \qquad 0 < x_2 < 1.$$

Finally, we shall compare the values of $\Pr\,(0 < X_1 < \frac{1}{2}|X_2 = \frac{3}{4})$ and $\Pr\,(0 < X_1 < \frac{1}{2})$. We have

$$\Pr\,(0 < X_1 < \tfrac{1}{2}|X_2 = \tfrac{3}{4}) = \int_0^{1/2} f(x_1|\tfrac{3}{4})\, dx_1 = \int_0^{1/2} (\tfrac{4}{3})\, dx_1 = \tfrac{2}{3},$$

but

$$\Pr\,(0 < X_1 < \tfrac{1}{2}) = \int_0^{1/2} f_1(x_1)\, dx_1 = \int_0^{1/2} 2(1 - x_1)\, dx_1 = \tfrac{3}{4}.$$

We shall now discuss the notions of marginal and conditional probability density functions from the point of view of n random variables. All of the preceding definitions can be directly generalized to the case of n variables in the following manner. Let the random variables $X_1, X_2, \ldots, X_n$ have the joint p.d.f. $f(x_1, x_2, \ldots, x_n)$. If the random variables are of the continuous type, then by an argument similar to the two-variable case, we have for every $a < b$,

$$\Pr\,(a < X_1 < b) = \int_a^b f_1(x_1)\, dx_1,$$

where $f_1(x_1)$ is defined by the $(n - 1)$-fold integral

$$f_1(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, dx_2 \cdots dx_n.$$

Accordingly, $f_1(x_1)$ is the p.d.f. of the one random variable X_1 and $f_1(x_1)$ is called the marginal p.d.f. of X_1. The marginal probability density functions $f_2(x_2), \ldots, f_n(x_n)$ of $X_2, \ldots, X_n$, respectively, are similar $(n - 1)$-fold integrals.

Up to this point, each marginal p.d.f. has been a p.d.f. of one random variable. It is convenient to extend this terminology to joint probability density functions. We shall do this now. Let $f(x_1, x_2, \ldots, x_n)$ be the joint p.d.f. of the n random variables $X_1, X_2, \ldots, X_n$, just as before. Now, however, let us take any group of $k < n$ of these random variables and let us find the joint p.d.f. of them. This joint p.d.f. is called the marginal p.d.f. of this particular group of k variables. To fix the ideas, take $n = 6$, $k = 3$, and let us select the group X_2, X_4, X_5. Then the marginal p.d.f. of X_2, X_4, X_5 is the joint p.d.f. of this particular group of three variables, namely,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4, x_5, x_6)\, dx_1\, dx_3\, dx_6,$$

if the random variables are of the continuous type.

We shall next extend the definition of a conditional p.d.f. If $f_1(x_1) > 0$, the symbol $f(x_2, \ldots, x_n|x_1)$ is defined by the relation

$$f(x_2, \ldots, x_n|x_1) = \frac{f(x_1, x_2, \ldots, x_n)}{f_1(x_1)},$$

and $f(x_2, \ldots, x_n|x_1)$ is called the *joint conditional p.d.f.* of $X_2, \ldots, X_n$, given $X_1 = x_1$. The joint conditional p.d.f. of any $n - 1$ random variables, say $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, given $X_i = x_i$, is defined as the joint p.d.f. of $X_1, X_2, \ldots, X_n$ divided by marginal p.d.f. $f_i(x_i)$, provided $f_i(x_i) > 0$. More generally, the joint conditional p.d.f. of $n - k$ of the random variables, for given values of the remaining k variables, is defined as the joint p.d.f. of the n variables divided by the marginal p.d.f. of the particular group of k variables, provided the latter p.d.f. is positive. We remark that there are many other conditional probability density functions; for instance, see Exercise 2.17.

Because a conditional p.d.f. is a p.d.f. of a certain number of random variables, the mathematical expectation of a function of these random variables has been defined. To emphasize the fact that a conditional p.d.f. is under consideration, such expectations are called conditional expectations. For instance, the conditional expectation of $u(X_2, \ldots, X_n)$ given $X_1 = x_1$, is, for random variables of the continuous type, given by

$$E[u(X_2, \ldots, X_n)|x_1] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_2, \ldots, x_n) f(x_2, \ldots, x_n|x_1)\, dx_2 \cdots dx_n,$$

provided $f_1(x_1) > 0$ and the integral converges (absolutely). If the random variables are of the discrete type, conditional mathematical expectations are, of course, computed by using sums instead of integrals.

EXERCISES

2.10. Let X_1 and X_2 have the joint p.d.f. $f(x_1, x_2) = x_1 + x_2, 0 < x_1 < 1, 0 < x_2 < 1$, zero elsewhere. Find the conditional mean and variance of X_2, given $X_1 = x_1, 0 < x_1 < 1$.

2.11. Let $f(x_1|x_2) = c_1 x_1/x_2^2, 0 < x_1 < x_2, 0 < x_2 < 1$, zero elsewhere, and $f_2(x_2) = c_2 x_2^4, 0 < x_2 < 1$, zero elsewhere, denote, respectively, the conditional p.d.f. of X_1, given $X_2 = x_2$, and the marginal p.d.f. of X_2. Determine: (a) the constants c_1 and c_2; (b) the joint p.d.f. of X_1 and X_2; (c) $\Pr(\frac{1}{4} < X_1 < \frac{1}{2}|X_2 = \frac{5}{8})$; and (d) $\Pr(\frac{1}{4} < X_1 < \frac{1}{2})$.

2.12. Let $f(x_1, x_2) = 21x_1^2x_2^3, 0 < x_1 < x_2 < 1$, zero elsewhere, be the joint p.d.f. of X_1 and X_2. Find the conditional mean and variance of X_1, given $X_2 = x_2, 0 < x_2 < 1$.

2.13. If X_1 and X_2 are random variables of the discrete type having p.d.f. $f(x_1, x_2) = (x_1 + 2x_2)/18$, $(x_1, x_2) = (1, 1), (1, 2), (2, 1), (2, 2)$, zero elsewhere, determine the conditional mean and variance of X_2, given $X_1 = x_1$, $x_1 = 1$ or 2.

2.14. Five cards are drawn at random and without replacement from a bridge deck. Let the random variables X_1, X_2, and X_3 denote, respectively, the number of spades, the number of hearts, and the number of diamonds that appear among the five cards. (a) Determine the joint p.d.f. of X_1, X_2, and X_3. (b) Find the marginal probability density functions of X_1, X_2, and X_3. (c) What is the joint conditional p.d.f. of X_2 and X_3, given that $X_1 = 3$?

2.15. Let X_1 and X_2 have the joint p.d.f. $f(x_1, x_2)$ described as follows:

(x_1, x_2)	(0, 0)	(0, 1)	(1, 0)	(1, 1)	(2, 0)	(2, 1)
$f(x_1, x_2)$	$\frac{1}{18}$	$\frac{3}{18}$	$\frac{4}{18}$	$\frac{3}{18}$	$\frac{6}{18}$	$\frac{1}{18}$

and $f(x_1, x_2)$ is equal to zero elsewhere. Find the two marginal probability density functions and the two conditional means.

2.16. Let us choose at random a point from the interval (0, 1) and let the random variable X_1 be equal to the number which corresponds to that point. Then choose a point at random from the interval $(0, x_1)$, where x_1 is the experimental value of X_1; and let the random variable X_2 be equal to the number which corresponds to this point. (a) Make assumptions about the marginal p.d.f. $f_1(x_1)$ and the conditional p.d.f. $f(x_2|x_1)$. (b) Compute $\Pr(X_1 + X_2 \geq 1)$. (c) Find the conditional mean $E(X_1|x_2)$.

2.17. Let $f(x)$ and $F(x)$ denote, respectively, the p.d.f. and the distribution function of the random variable X. The conditional p.d.f. of X, given $X > x_0$, x_0 a fixed number, is defined by $f(x|X > x_0) = f(x)/[1 - F(x_0)]$, $x_0 < x$, zero elsewhere. This kind of conditional p.d.f. finds application in a

problem of time until death, given survival until time x_0. (a) Show that $f(x|X > x_0)$ is a p.d.f. (b) Let $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Compute Pr $(X > 2|\ X > 1)$.

2.3 The Correlation Coefficient

Let X, Y, and Z denote random variables that have joint p.d.f. $f(x, y, z)$. If $u(x, y, z)$ is a function of x, y, and z, then $E[u(X, Y, Z)]$ was defined, subject to its existence, on p. 42. The existence of all mathematical expectations will be assumed in this discussion. The means of X, Y, and Z, say μ_1, μ_2, and μ_3, are obtained by taking $u(x, y, z)$ to be x, y, and z, respectively; and the variances of X, Y, and Z, say σ_1^2, σ_2^2, and σ_3^2, are obtained by setting the function $u(x, y, z)$ equal to $(x - \mu_1)^2$, $(y - \mu_2)^2$, and $(z - \mu_3)^2$, respectively. Consider the mathematical expectation

$$\begin{aligned} E[(X - \mu_1)(Y - \mu_2)] &= E(XY - \mu_2 X - \mu_1 Y + \mu_1\mu_2) \\ &= E(XY) - \mu_2 E(X) - \mu_1 E(Y) + \mu_1\mu_2 \\ &= E(XY) - \mu_1\mu_2. \end{aligned}$$

This number is called the *covariance* of X and Y. The covariance of X and Z is given by $E[(X - \mu_1)(Z - \mu_3)]$, and the covariance of Y and Z is $E[(Y - \mu_2)(Z - \mu_3)]$. If each of σ_1 and σ_2 is positive, the number

$$\rho_{12} = \frac{E[(X - \mu_1)(Y - \mu_2)]}{\sigma_1\sigma_2}$$

is called the *correlation coefficient* of X and Y. If the standard deviations are positive, the correlation coefficient of any two random variables is defined to be the covariance of the two random variables divided by the product of the standard deviations of the two random variables. It should be noted that the expected value of the product of two random variables is equal to the product of their expectations plus their covariance.

Example 1. Let the random variables X and Y have the joint p.d.f.

$$\begin{aligned} f(x, y) &= x + y, \qquad 0 < x < 1, 0 < y < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

We shall compute the correlation coefficient of X and Y. When only two variables are under consideration, we shall denote the correlation coefficient by ρ. Now

$$\mu_1 = E(X) = \int_0^1 \int_0^1 x(x + y)\, dx\, dy = \tfrac{7}{12}$$

and

$$\sigma_1^2 = E(X^2) - \mu_1^2 = \int_0^1 \int_0^1 x^2(x + y)\, dx\, dy - (\tfrac{7}{12})^2 = \tfrac{11}{144}.$$

Similarly,

$$\mu_2 = E(Y) = \tfrac{7}{12} \qquad \text{and} \qquad \sigma_2^2 = E(Y^2) - \mu_2^2 = \tfrac{11}{144}.$$

The covariance of X and Y is

$$E(XY) - \mu_1\mu_2 = \int_0^1 \int_0^1 xy(x + y)\, dx\, dy - (\tfrac{7}{12})^2 = -\tfrac{1}{144}.$$

Accordingly, the correlation coefficient of X and Y is

$$\rho = \frac{-\frac{1}{144}}{\sqrt{(\frac{11}{144})(\frac{11}{144})}} = -\frac{1}{11}.$$

Remark. For certain kinds of distributions of two random variables, say X and Y, the correlation coefficient ρ proves to be a very useful characteristic of the distribution. Unfortunately, the formal definition of ρ does not reveal this fact. At this time we make some observations about ρ, some of which will be explored more fully at a later stage. It will soon be seen that if a joint distribution of two variables has a correlation coefficient (that is, if both of the variances are positive), then ρ satisfies $-1 \leq \rho \leq 1$. If $\rho = 1$, there is a line with equation $y = a + bx$, $b > 0$, the graph of which contains all of the probability for the distribution of X and Y. In this extreme case, we have $\Pr(Y = a + bX) = 1$. If $\rho = -1$, we have the same state of affairs except that $b < 0$. This suggests the following interesting question: When ρ does not have one of its extreme values, is there a line in the xy-plane such that the probability for X and Y tends to be concentrated in a band about this line? Under certain restrictive conditions this is in fact the case, and under those conditions we can look upon ρ as a measure of the intensity of the concentration of the probability for X and Y about that line.

Next, let $f(x, y)$ denote the joint p.d.f. of two random variables X and Y and let $f_1(x)$ denote the marginal p.d.f. of X. The conditional p.d.f. of Y, given $X = x$, is

$$f(y|x) = \frac{f(x, y)}{f_1(x)}$$

at points where $f_1(x) > 0$. Then the conditional mean of Y, given $X = x$, is given by

$$E(Y|x) = \int_{-\infty}^{\infty} yf(y|x)\, dy = \frac{\int_{-\infty}^{\infty} yf(x, y)\, dy}{f_1(x)},$$

when dealing with random variables of the continuous type. This conditional mean of Y, given $X = x$, is, of course, a function of x alone, say $\varphi(x)$. In like vein, the conditional mean of X, given $Y = y$, is a function of y alone, say $\psi(y)$.

In case $\varphi(x)$ is a linear function of x, say $\varphi(x) = a + bx$, we say the conditional mean of Y is linear in x; or that Y has a linear conditional mean. When $\varphi(x) = a + bx$, the constants a and b have simple values which will now be determined.

It will be assumed that neither σ_1^2 nor σ_2^2, the variances of X and Y, is zero. From

$$E(Y|x) = \frac{\int_{-\infty}^{\infty} yf(x, y)\, dy}{f_1(x)} = a + bx,$$

we have

$$\int_{-\infty}^{\infty} yf(x, y)\, dy = (a + bx)f_1(x). \tag{1}$$

If both members of Equation (1) are integrated on x, it is seen that

$$E(Y) = a + bE(X),$$

or

$$\mu_2 = a + b\mu_1, \tag{2}$$

where $\mu_1 = E(X)$ and $\mu_2 = E(Y)$. If both members of Equation (1) are first multiplied by x and then integrated on x, we have

$$E(XY) = aE(X) + bE(X^2),$$

or

$$\rho\sigma_1\sigma_2 + \mu_1\mu_2 = a\mu_1 + b(\sigma_1^2 + \mu_1^2), \tag{3}$$

where $\rho\sigma_1\sigma_2$ is the covariance of X and Y. The simultaneous solution of Equations (2) and (3) yields

$$a = \mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1 \qquad \text{and} \qquad b = \rho\frac{\sigma_2}{\sigma_1}.$$

That is,

$$\varphi(x) = E(Y|x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$$

is the conditional mean of Y, given $X = x$, when the conditional mean of Y is linear in x. If the conditional mean of X, given $Y = y$, is linear in y, then that conditional mean is given by

$$\psi(y) = E(X|y) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2).$$

We shall next investigate the variance of a conditional distribution under the assumption that the conditional mean is linear. The conditional variance of Y is given by

$$(4)\quad E\{[Y - E(Y|x)]^2|x\} = \int_{-\infty}^{\infty} \left[y - \mu_2 - \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)\right]^2 f(y|x)\,dy$$

$$= \frac{\displaystyle\int_{-\infty}^{\infty} \left[(y - \mu_2) - \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)\right]^2 f(x, y)\,dy}{f_1(x)}$$

when the random variables are of the continuous type. This variance is nonnegative and is at most a function of x alone. If then, it is multiplied by $f_1(x)$ and integrated on x, the result obtained will be nonnegative. This result is

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[(y - \mu_2) - \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)\right]^2 f(x, y)\,dy\,dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[(y - \mu_2)^2 - 2\rho\frac{\sigma_2}{\sigma_1}(y - \mu_2)(x - \mu_1) + \rho^2\frac{\sigma_2^2}{\sigma_1^2}(x - \mu_1)^2\right] \times f(x, y)\,dy\,dx$$

$$= E[(Y - \mu_2)^2] - 2\rho\frac{\sigma_2}{\sigma_1}E[(X - \mu_1)(Y - \mu_2)] + \rho^2\frac{\sigma_2^2}{\sigma_1^2}E[(X - \mu_1)^2]$$

$$= \sigma_2^2 - 2\rho\frac{\sigma_2}{\sigma_1}\rho\sigma_1\sigma_2 + \rho^2\frac{\sigma_2^2}{\sigma_1^2}\sigma_1^2$$

$$= \sigma_2^2 - 2\rho^2\sigma_2^2 + \rho^2\sigma_2^2 = \sigma_2^2(1 - \rho^2) \geq 0.$$

That is, if the variance, Equation (4), is denoted by $k(x)$, then $E[k(X)] = \sigma_2^2(1 - \rho^2) \geq 0$. Accordingly, $\rho^2 \leq 1$, or $-1 \leq \rho \leq 1$. It is left as an exercise to prove that $-1 \leq \rho \leq 1$ whether the conditional mean is or is not linear.

Suppose that the variance, Equation (4), is positive but not a function of x; that is, the variance is a constant $k > 0$. Now if k is multiplied by $f_1(x)$ and integrated on x, the result is k, so that $k = \sigma_2^2(1 - \rho^2)$. Thus, in this case, the variance of each conditional distribution of Y, given $X = x$, is $\sigma_2^2(1 - \rho^2)$. If $\rho = 0$, the variance of each conditional distribution of Y, given $X = x$, is σ_2^2, the variance of the marginal distribution of Y. On the other hand, if ρ^2 is near one, the variance of each conditional distribution of Y, given $X = x$, is relatively small, and there is a high concentration of the probability for this conditional distribution near the mean $E(Y|x) = \mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1)$.

It should be pointed out that if the random variables X and Y in the preceding discussion are taken to be of the discrete type, the results just obtained are valid.

Example 2. Let the random variables X and Y have the linear conditional means $E(Y|x) = 4x + 3$ and $E(X|y) = \frac{1}{16}y - 3$. In accordance with the general formulas for the linear conditional means, we see that $E(Y|x) = \mu_2$ if $x = \mu_1$ and $E(X|y) = \mu_1$ if $y = \mu_2$. Accordingly, in this special case, we have $\mu_2 = 4\mu_1 + 3$ and $\mu_1 = \frac{1}{16}\mu_2 - 3$ so that $\mu_1 = -\frac{15}{4}$ and $\mu_2 = -12$. The general formulas for the linear conditional means also show that the product of the coefficients of x and y, respectively, is equal to ρ^2 and that the quotient of these coefficients is equal to σ_2^2/σ_1^2. Here $\rho^2 = 4(\frac{1}{16}) = \frac{1}{4}$ with $\rho = \frac{1}{2}$ (not $-\frac{1}{2}$), and $\sigma_2^2/\sigma_1^2 = 64$. Thus, from the two linear conditional means, we are able to find the values of μ_1, μ_2, ρ, and σ_2/σ_1, but not the values of σ_1 and σ_2.

This section will conclude with a definition and an illustrative example. Let $f(x, y)$ denote the joint p.d.f. of the two random variables X and Y. If $E(e^{t_1X + t_2Y})$ exists for $-h_1 < t_1 < h_1$, $-h_2 < t_2 < h_2$, where h_1 and h_2 are positive, it is denoted by $M(t_1, t_2)$ and is called the *moment-generating function* of the joint distribution of X and Y. As in the case of one random variable, the moment-generating function $M(t_1, t_2)$ completely determines the joint distribution of X and Y, and hence the marginal distributions of X and Y. In fact,

$$M(t_1, 0) = E(e^{t_1X}) = M(t_1)$$

and

$$M(0, t_2) = E(e^{t_2Y}) = M(t_2).$$

In addition, in the case of random variables of the continuous type,

$$\frac{\partial^{k+m}M(t_1, t_2)}{\partial t_1^k \partial t_2^m} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^m e^{t_1x + t_2y} f(x, y)\, dx\, dy,$$

so that

$$\left.\frac{\partial^{k+m}M(t_1, t_2)}{\partial t_1^k \partial t_2^m}\right|_{t_1 = t_2 = 0} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^m f(x, y)\, dx\, dy = E(X^kY^m).$$

For instance, in a simplified notation which appears to be clear,

$$\mu_1 = E(X) = \frac{\partial M(0, 0)}{\partial t_1}, \qquad \mu_2 = E(Y) = \frac{\partial M(0, 0)}{\partial t_2},$$

$$\sigma_1^2 = E(X^2) - \mu_1^2 = \frac{\partial^2 M(0, 0)}{\partial t_1^2} - \mu_1^2,$$

(5)

$$\sigma_2^2 = E(Y^2) - \mu_2^2 = \frac{\partial^2 M(0, 0)}{\partial t_2^2} - \mu_2^2,$$

$$E[(X - \mu_1)(Y - \mu_2)] = \frac{\partial^2 M(0, 0)}{\partial t_1\, \partial t_2} - \mu_1\mu_2.$$

It is fairly obvious that the results of Equations (5) hold if X and Y are random variables of the discrete type. Thus the correlation coefficients may be computed by using the moment-generating function of the joint distribution if that function is readily available. An illustrative example follows.

Example 3. Let the continuous-type random variables X and Y have the joint p.d.f.

$$f(x, y) = e^{-y}, \qquad 0 < x < y < \infty,$$
$$= 0 \text{ elsewhere.}$$

The moment-generating function of this joint distribution is

$$M(t_1, t_2) = \int_0^\infty \int_x^\infty \exp(t_1 x + t_2 y - y)\, dy\, dx$$
$$= \frac{1}{(1 - t_1 - t_2)(1 - t_2)},$$

provided $t_1 + t_2 < 1$ and $t_2 < 1$. For this distribution, Equations (5) become

$$\mu_1 = 1, \qquad \mu_2 = 2,$$
$$\sigma_1^2 = 1, \qquad \sigma_2^2 = 2, \tag{6}$$
$$E[(X - \mu_1)(Y - \mu_2)] = 1.$$

Verification of results of Equations (6) is left as an exercise. If, momentarily, we accept these results, the correlation coefficient of X and Y is $\rho = 1/\sqrt{2}$. Furthermore, the moment-generating functions of the marginal distributions of X and Y are, respectively,

$$M(t_1, 0) = \frac{1}{1 - t_1}, \qquad t_1 < 1,$$

$$M(0, t_2) = \frac{1}{(1 - t_2)^2}, \qquad t_2 < 1.$$

These moment-generating functions are, of course, respectively, those of the marginal probability density functions,

$$f_1(x) = \int_x^\infty e^{-y}\, dy = e^{-x}, \qquad 0 < x < \infty,$$

zero elsewhere, and

$$f_2(y) = e^{-y} \int_0^y dx = y e^{-y}, \qquad 0 < y < \infty,$$

zero elsewhere.

EXERCISES

2.18. Let the random variables X and Y have the joint p.d.f.

(a) $f(x, y) = \frac{1}{3}$, $(x, y) = (0, 0), (1, 1), (2, 2)$, zero elsewhere.

(b) $f(x, y) = \frac{1}{3}$, $(x, y) = (0, 2), (1, 1), (2, 0)$, zero elsewhere.

(c) $f(x, y) = \frac{1}{3}$, $(x, y) = (0, 0), (1, 1), (2, 0)$, zero elsewhere.

In each case compute the correlation coefficient of X and Y.

2.19. Let X and Y have the joint p.d.f. described as follows:

(x, y)	$(1, 1)$	$(1, 2)$	$(1, 3)$	$(2, 1)$	$(2, 2)$	$(2, 3)$
$f(x, y)$	$\frac{2}{15}$	$\frac{4}{15}$	$\frac{3}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{4}{15}$

and $f(x, y)$ is equal to zero elsewhere. Find the correlation coefficient ρ.

2.20. Let $f(x, y) = 2$, $0 < x < y$, $0 < y < 1$, zero elsewhere, be the joint p.d.f. of X and Y. Show that the conditional means are, respectively, $(1 + x)/2$, $0 < x < 1$, and $y/2$, $0 < y < 1$. Show that the correlation coefficient of X and Y is $\rho = \frac{1}{2}$.

2.21. Show that the variance of the conditional distribution of Y, given $X = x$, in Exercise 2.20, is $(1 - x)^2/12$, $0 < x < 1$, and that the variance of the conditional distribution of X, given $Y = y$, is $y^2/12$, $0 < y < 1$.

2.22. Verify the results of Equations (6) of this section.

2.23. Let X and Y have the joint p.d.f. $f(x, y) = 1$, $-x < y < x$, $0 < x < 1$, zero elsewhere. Show that, on the set of positive probability density, the graph of $E(Y|x)$ is a straight line, whereas that of $E(X|y)$ is not a straight line.

2.24. If the correlation coefficient ρ of X and Y exists, show that $-1 \leq \rho \leq 1$. *Hint.* Consider the discriminant of the nonnegative quadratic function $h(v) = E\{[(X - \mu_1) + v(Y - \mu_2)]^2\}$, where v is real and is not a function of X nor of Y.

2.25. Let $\psi(t_1, t_2) = \ln M(t_1, t_2)$, where $M(t_1, t_2)$ is the moment-generating function of X and Y. Show that

$$\frac{\partial \psi(0, 0)}{\partial t_i}, \qquad \frac{\partial^2 \psi(0, 0)}{\partial t_i^2}, \qquad i = 1, 2,$$

and

$$\frac{\partial^2 \psi(0, 0)}{\partial t_1 \, \partial t_2}$$

yield the means, the variances, and the covariance of the two random variables.

2.26. Let X_1, X_2, and X_3 be three random variables with means, variances, and correlation coefficients, denoted by μ_1, μ_2, μ_3; σ_1^2, σ_2^2, σ_3^2; and

ρ_{12}, ρ_{13}, ρ_{23}, respectively. If $E(X_1 - \mu_1|x_2, x_3) = b_2(x_2 - \mu_2) + b_3(x_3 - \mu_3)$, where b_2 and b_3 are constants, determine b_2 and b_3 in terms of the variances and the correlation coefficients.

2.4 Stochastic Independence

Let X_1 and X_2 denote random variables of either the continuous or the discrete type which have the joint p.d.f. $f(x_1, x_2)$ and marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, respectively. In accordance with the definition of the conditional p.d.f. $f(x_2|x_1)$, we may write the joint p.d.f. $f(x_1, x_2)$ as

$$f(x_1, x_2) = f(x_2|x_1)f_1(x_1).$$

Suppose we have an instance where $f(x_2|x_1)$ does not depend upon x_1. Then the marginal p.d.f. of X_2 is, for random variables of the continuous type,

$$\begin{aligned} f_2(x_2) &= \int_{-\infty}^{\infty} f(x_2|x_1)f_1(x_1)\,dx_1 \\ &= f(x_2|x_1)\int_{-\infty}^{\infty} f_1(x_1)\,dx_1 \\ &= f(x_2|x_1). \end{aligned}$$

Accordingly,

$$f_2(x_2) = f(x_2|x_1) \qquad \text{and} \qquad f(x_1, x_2) = f_1(x_1)f_2(x_2),$$

when $f(x_2|x_1)$ does not depend upon x_1. That is, if the conditional distribution of X_2, given $X_1 = x_1$, is independent of any assumption about x_1, then $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. These considerations motivate the following definition.

Definition 1. Let the random variables X_1 and X_2 have the joint p.d.f. $f(x_1, x_2)$ and the marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, respectively. The random variables X_1 and X_2 are said to be stochastically independent if, and only if, $f(x_1, x_2) \equiv f_1(x_1)f_2(x_2)$. Random variables that are not stochastically independent are said to be stochastically dependent.

Remarks. Two comments should be made about the preceding definition. First the product of two nonnegative functions $f_1(x_1)f_2(x_2)$ means a function that is positive on a product space. That is, if $f_1(x_1)$ and $f_2(x_2)$ are positive on, and only on, the respective spaces $\mathcal{A}_1$ and $\mathcal{A}_2$, then the product of $f_1(x_1)$ and $f_2(x_2)$ is positive on, and only on, the product space $\mathcal{A} = \{(x_1, x_2); x_1 \in \mathcal{A}_1, x_2 \in \mathcal{A}_2\}$. For instance, if $\mathcal{A}_1 = \{x_1; 0 < x_1 < 1\}$ and $\mathcal{A}_2 = \{x_2; 0 < x_2 < 3\}$, then $\mathcal{A} = \{(x_1, x_2); 0 < x_1 < 1, 0 < x_2 < 3\}$. The

second remark pertains to the identity. The identity in Definition 1 should be interpreted as follows. There may be certain points $(x_1, x_2) \in \mathscr{A}$ at which $f(x_1, x_2) \neq f_1(x_1)f_2(x_2)$. However, if A is the set of points (x_1, x_2) at which the equality does not hold, then $P(A) = 0$. In the subsequent theorems and the subsequent generalizations, a product of nonnegative functions and an identity should be interpreted in an analogous manner.

Example 1. Let the joint p.d.f. of X_1 and X_2 be

$$\begin{aligned} f(x_1, x_2) &= x_1 + x_2, \qquad 0 < x_1 < 1,\ 0 < x_2 < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

It will be shown that X_1 and X_2 are stochastically dependent. Here the marginal probability density functions are

$$\begin{aligned} f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2 &= \int_0^1 (x_1 + x_2)\, dx_2 = x_1 + \tfrac{1}{2}, \qquad 0 < x_1 < 1, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

and

$$\begin{aligned} f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1 &= \int_0^1 (x_1 + x_2)\, dx_1 = \tfrac{1}{2} + x_2, \qquad 0 < x_2 < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Since $f(x_1, x_2) \not\equiv f_1(x_1)f_2(x_2)$, the random variables X_1 and X_2 are stochastically dependent.

The following theorem makes it possible to assert, without computing the marginal probability density functions, that the random variables X_1 and X_2 of Example 1 are stochastically dependent.

Theorem 1. *Let the random variables X_1 and X_2 have the joint p.d.f. $f(x_1, x_2)$. Then X_1 and X_2 are stochastically independent if and only if $f(x_1, x_2)$ can be written as a product of a nonnegative function of x_1 alone and a nonnegative function of x_2 alone. That is,*

$$f(x_1, x_2) \equiv g(x_1)h(x_2),$$

where $g(x_1) > 0$, $x_1 \in \mathscr{A}_1$, zero elsewhere, and $h(x_2) > 0$, $x_2 \in \mathscr{A}_2$, zero elsewhere.

Proof. If X_1 and X_2 are stochastically independent, then $f(x_1, x_2) \equiv f_1(x_1)f_2(x_2)$, where $f_1(x_1)$ and $f_2(x_2)$ are the marginal probability density functions of X_1 and X_2, respectively. Thus, the condition $f(x_1, x_2) \equiv g(x_1)h(x_2)$ is fulfilled.

Conversely, if $f(x_1, x_2) \equiv g(x_1)h(x_2)$, then, for random variables of the continuous type, we have

$$f_1(x_1) = \int_{-\infty}^{\infty} g(x_1)h(x_2)\,dx_2 = g(x_1)\int_{-\infty}^{\infty} h(x_2)\,dx_2 = c_1 g(x_1)$$

and

$$f_2(x_2) = \int_{-\infty}^{\infty} g(x_1)h(x_2)\,dx_1 = h(x_2)\int_{-\infty}^{\infty} g(x_1)\,dx_1 = c_2 h(x_2),$$

where c_1 and c_2 are constants, not functions of x_1 or x_2. Moreover, $c_1 c_2 = 1$ because

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1)h(x_2)\,dx_1\,dx_2 = \left[\int_{-\infty}^{\infty} g(x_1)\,dx_1\right]\left[\int_{-\infty}^{\infty} h(x_2)\,dx_2\right] = c_2 c_1.$$

These results imply that

$$f(x_1, x_2) \equiv g(x_1)h(x_2) \equiv c_1 g(x_1) c_2 h(x_2) \equiv f_1(x_1) f_2(x_2).$$

Accordingly, X_1 and X_2 are stochastically independent.

If we now refer to Example 1, we see that the joint p.d.f.

$$\begin{aligned} f(x_1, x_2) &= x_1 + x_2, \qquad 0 < x_1 < 1, 0 < x_2 < 1, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

cannot be written as the product of a nonnegative function of x_1 alone and a nonnegative function of x_2 alone. Accordingly, X_1 and X_2 are stochastically dependent.

Example 2. Let the p.d.f. of the random variables X_1 and X_2 be $f(x_1, x_2) = 8x_1x_2$, $0 < x_1 < x_2 < 1$, zero elsewhere. The formula $8x_1x_2$ might suggest to some that X_1 and X_2 are stochastically independent. However, if we consider the space $\mathcal{A} = \{(x_1, x_2); 0 < x_1 < x_2 < 1\}$, we see that it is not a product space. This should make it clear that, in general, X_1 and X_2 must be stochastically dependent if the space of positive probability density of X_1 and X_2 is bounded by a curve that is neither a horizontal nor a vertical line.

We now give a theorem that frequently simplifies the calculations of probabilities of events which involve stochastically independent variables.

Thereom 2. *If X_1 and X_2 are stochastically independent random variables with marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, respectively, then*

$$\Pr(a < X_1 < b, c < X_2 < d) = \Pr(a < X_1 < b)\Pr(c < X_2 < d)$$

for every $a < b$ and $c < d$, where a, b, c, and d are constants.

Proof. From the stochastic independence of X_1 and X_2, the joint p.d.f. of X_1 and X_2 is $f_1(x_1)f_2(x_2)$. Accordingly, in the continuous case,

$$\begin{aligned}\Pr(a < X_1 < b, c < X_2 < d) &= \int_a^b \int_c^d f_1(x_1)f_2(x_2)\,dx_2\,dx_1 \\ &= \left[\int_a^b f_1(x_1)\,dx_1\right]\left[\int_c^d f_2(x_2)\,dx_2\right] \\ &= \Pr(a < X_1 < b)\Pr(c < X_2 < d);\end{aligned}$$

or, in the discrete case,

$$\begin{aligned}\Pr(a < X_1 < b, c < X_2 < d) &= \sum_{a<x_1<b}\ \sum_{c<x_2<d} f_1(x_1)f_2(x_2) \\ &= \left[\sum_{a<x_1<b} f_1(x_1)\right]\left[\sum_{c<x_2<d} f_2(x_2)\right] \\ &= \Pr(a < X_1 < b)\Pr(c < X_2 < d),\end{aligned}$$

as was to be shown.

Example 3. In Example 1, X_1 and X_2 were found to be stochastically dependent. There, in general,

$$\Pr(a < X_1 < b, c < X_2 < d) \neq \Pr(a < X_1 < b)\Pr(c < X_2 < d).$$

For instance,

$$\Pr(0 < X_1 < \tfrac{1}{2}, 0 < X_2 < \tfrac{1}{2}) = \int_0^{1/2}\int_0^{1/2} (x_1 + x_2)\,dx_1\,dx_2 = \tfrac{1}{8},$$

whereas

$$\Pr(0 < X_1 < \tfrac{1}{2}) = \int_0^{1/2} (x_1 + \tfrac{1}{2})\,dx_1 = \tfrac{3}{8}$$

and

$$\Pr(0 < X_2 < \tfrac{1}{2}) = \int_0^{1/2} (\tfrac{1}{2} + x_2)\,dx_2 = \tfrac{3}{8}.$$

Not merely are calculations of some probabilities usually simpler when we have stochastically independent random variables, but many mathematical expectations, including certain moment-generating functions, have comparably simpler computations. The following result will prove so useful that we state it in form of a theorem.

Theorem 3. *Let the stochastically independent random variables X_1 and X_2 have the marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, respectively. The expected value of the product of a function $u(X_1)$ of X_1 alone and a function $v(X_2)$ of X_2 alone is, subject to their existence, equal to the product of the expected value of $u(X_1)$ and the expected value of $v(X_2)$; that is,*

$$E[u(X_1)v(X_2)] = E[u(X_1)]E[v(X_2)].$$

Proof. The stochastic independence of X_1 and X_2 implies that the joint p.d.f. of X_1 and X_2 is $f_1(x_1)f_2(x_2)$. Thus, we have, by definition of mathematical expectation, in the continuous case,

$$\begin{aligned} E[u(X_1)v(X_2)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x_1)v(x_2)f_1(x_1)f_2(x_2)\,dx_1\,dx_2 \\ &= \left[\int_{-\infty}^{\infty} u(x_1)f_1(x_1)\,dx_1\right]\left[\int_{-\infty}^{\infty} v(x_2)f_2(x_2)\,dx_2\right] \\ &= E[u(X_1)]E[v(X_2)]; \end{aligned}$$

or, in the discrete case,

$$\begin{aligned} E[u(X_1)v(X_2)] &= \sum_{x_2}\sum_{x_1} u(x_1)v(x_2)f_1(x_1)f_2(x_2) \\ &= \left[\sum_{x_1} u(x_1)f_1(x_1)\right]\left[\sum_{x_2} v(x_2)f_2(x_2)\right] \\ &= E[u(X_1)]E[v(X_2)], \end{aligned}$$

as stated in the theorem.

Example 4. Let X and Y be two stochastically independent random variables with means μ_1 and μ_2 and positive variances σ_1^2 and σ_2^2, respectively. We shall show that the stochastic independence of X and Y implies that the correlation coefficient of X and Y is zero. This is true because the covariance of X and Y is equal to

$$E[(X - \mu_1)(Y - \mu_2)] = E(X - \mu_1)E(Y - \mu_2) = 0.$$

We shall now prove a very useful theorem about stochastically independent random variables. The proof of the theorem relies heavily upon our assertion that a moment-generating function, when it exists, is unique and that it uniquely determines the distribution of probability.

Thereom 4. *Let X_1 and X_2 denote random variables that have the joint p.d.f. $f(x_1, x_2)$ and the marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, respectively. Furthermore, let $M(t_1, t_2)$ denote the moment-generating function of the distribution. Then X_1 and X_2 are stochastically independent if and only if $M(t_1, t_2) = M(t_1, 0)M(0, t_2)$.*

Proof. If X_1 and X_2 are stochastically independent, then

$$\begin{aligned} M(t_1, t_2) &= E(e^{t_1X_1 + t_2X_2}) \\ &= E(e^{t_1X_1}e^{t_2X_2}) \\ &= E(e^{t_1X_1})E(e^{t_2X_2}) \\ &= M(t_1, 0)M(0, t_2). \end{aligned}$$

Thus the stochastic independence of X_1 and X_2 implies that the moment-generating function of the joint distribution factors into the

product of the moment-generating functions of the two marginal distributions.

Suppose next that the moment-generating function of the joint distribution of X_1 and X_2 is given by $M(t_1, t_2) = M(t_1, 0)M(0, t_2)$. Now X_1 has the unique moment-generating function which, in the continuous case, is given by

$$M(t_1, 0) = \int_{-\infty}^{\infty} e^{t_1 x_1} f_1(x_1)\, dx_1.$$

Similarly, the unique moment-generating function of X_2, in the continuous case, is given by

$$M(0, t_2) = \int_{-\infty}^{\infty} e^{t_2 x_2} f_2(x_2)\, dx_2.$$

Thus we have

$$\begin{aligned} M(t_1, 0)M(0, t_2) &= \left[\int_{-\infty}^{\infty} e^{t_1 x_1} f_1(x_1)\, dx_1\right]\left[\int_{-\infty}^{\infty} e^{t_2 x_2} f_2(x_2)\, dx_2\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1 x_1 + t_2 x_2} f_1(x_1) f_2(x_2)\, dx_1\, dx_2. \end{aligned}$$

We are given that $M(t_1, t_2) = M(t_1, 0)M(0, t_2)$; so

$$M(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1 x_1 + t_2 x_2} f_1(x_1) f_2(x_2)\, dx_1\, dx_2.$$

But $M(t_1, t_2)$ is the moment-generating function of X_1 and X_2. Thus also

$$M(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1 x_1 + t_2 x_2} f(x_1, x_2)\, dx_1\, dx_2.$$

The uniqueness of the moment-generating function implies that the two distributions of probability that are described by $f_1(x_1)f_2(x_2)$ and $f(x_1, x_2)$ are the same. Thus

$$f(x_1, x_2) \equiv f_1(x_1)f_2(x_2).$$

That is, if $M(t_1, t_2) = M(t_1, 0)M(0, t_2)$, then X_1 and X_2 are stochastically independent. This completes the proof when the random variables are of the continuous type. With random variables of the discrete type, the proof is made by using summation instead of integration.

Let the random variables $X_1, X_2, \ldots, X_n$ have the joint p.d.f. $f(x_1, x_2, \ldots, x_n)$ and the marginal probability density functions $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$, respectively. The definition of the stochastic independence of X_1 and X_2 is generalized to the mutual stochastic independence of $X_1, X_2, \ldots, X_n$ as follows: The random variables

$X_1, X_2, \ldots, X_n$ are said to be mutually stochastically independent if and only if $f(x_1, x_2, \ldots, x_n) \equiv f_1(x_1)f_2(x_2)\cdots f_n(x_n)$. It follows immediately from this definition of the mutual stochastic independence of $X_1, X_2, \ldots, X_n$ that

$$\begin{aligned}\Pr(a_1 < X_1 < b_1, a_2 < X_2 < b_2, \ldots, a_n < X_n < b_n) &= \Pr(a_1 < X_1 < b_1)\Pr(a_2 < X_2 < b_2)\cdots\Pr(a_n < X_n < b_n) \\ &= \prod_{i=1}^{n} \Pr(a_i < X_i < b_i),\end{aligned}$$

where the symbol $\prod_{i=1}^{n} \varphi(i)$ is defined to be

$$\prod_{i=1}^{n} \varphi(i) = \varphi(1)\varphi(2)\cdots\varphi(n).$$

The theorem that $E[u(X_1)v(X_2)] = E[u(X_1)]E[v(X_2)]$ for stochastically independent random variables X_1 and X_2 becomes, for mutually stochastically independent random variables $X_1, X_2, \ldots, X_n$,

$$E[u_1(X_1)u_2(X_2)\cdots u_n(X_n)] = E[u_1(X_1)]E[u_2(X_2)]\cdots E[u_n(X_n)],$$

or

$$E\left[\prod_{i=1}^{n} u_i(X_i)\right] = \prod_{i=1}^{n} E[u_i(X_i)].$$

The moment-generating function of the joint distribution of n random variables $X_1, X_2, \ldots, X_n$ is defined as follows. Let

$$E[\exp(t_1X_1 + t_2X_2 + \cdots + t_nX_n)]$$

exist for $-h_i < t_i < h_i$, $i = 1, 2, \ldots, n$, where each h_i is positive. This expectation is denoted by $M(t_1, t_2, \ldots, t_n)$ and it is called the *moment-generating* function of the joint distribution of $X_1, \ldots, X_n$ (or simply the moment-generating function of $X_1, \ldots, X_n$). As in the cases of one and two variables, this moment-generating function is unique and uniquely determines the joint distribution of the n variables (and hence all marginal distributions). For example, the moment-generating function of the marginal distribution of X_i is $M(0, \ldots, 0, t_i, 0, \ldots, 0)$, $i = 1, 2, \ldots, n$; that of the marginal distribution of X_i and X_j is $M(0, \ldots, 0, t_i, 0, \ldots, 0, t_j, 0, \ldots, 0)$; and so on. Theorem 4 of this chapter can be generalized, and the factorization

$$M(t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} M(0, \ldots, 0, t_i, 0, \ldots, 0)$$

is a necessary and sufficient condition for the mutual stochastic independence of $X_1, X_2, \ldots, X_n$.

Remark. If X_1, X_2, and X_3 are mutually stochastically independent, they are *pairwise* stochastically *independent* (that is, X_i and X_j, $i \neq j$, where $i, j = 1, 2, 3$ are stochastically independent). However, the following example, due to S. Bernstein, shows that pairwise independence does not necessarily imply mutual independence. Let X_1, X_2, and X_3 have the joint p.d.f.

$$\begin{aligned} f(x_1, x_2, x_3) &= \tfrac{1}{4}, \qquad (x_1, x_2, x_3) \in \{(1, 0, 0), (0, 1, 0), (0, 0, 1), (1, 1, 1)\}, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

The joint p.d.f. of X_i and X_j, $i \neq j$, is

$$\begin{aligned} f_{ij}(x_i, x_j) &= \tfrac{1}{4}, \qquad (x_i, x_j) \in \{(0, 0), (1, 0), (0, 1), (1, 1)\}, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

whereas the marginal p.d.f. of X_i is

$$\begin{aligned} f_i(x_i) &= \tfrac{1}{2}, \qquad x_i = 0, 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Obviously, if $i \neq j$, we have

$$f_{ij}(x_i, x_j) \equiv f_i(x_i) f_j(x_j),$$

and thus X_i and X_j are stochastically independent. However,

$$f(x_1, x_2, x_3) \not\equiv f_1(x_1) f_2(x_2) f_3(x_3).$$

Thus X_1, X_2, and X_3 are not mutually stochastically independent.

***Example* 5.** Let X_1, X_2, and X_3 be three mutually stochastically independent random variables and let each have the p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. The joint p.d.f. of X_1, X_2, X_3 is $f(x_1)f(x_2)f(x_3) = 8x_1x_2x_3$, $0 < x_i < 1$, $i = 1, 2, 3$, zero elsewhere. Let Y be the maximum of X_1, X_2, and X_3. Then, for instance, we have

$$\begin{aligned} \Pr(Y \leq \tfrac{1}{2}) &= \Pr(X_1 \leq \tfrac{1}{2}, X_2 \leq \tfrac{1}{2}, X_3 \leq \tfrac{1}{2}) \\ &= \int_0^{1/2} \int_0^{1/2} \int_0^{1/2} 8x_1x_2x_3 \, dx_1 \, dx_2 \, dx_3 \\ &= (\tfrac{1}{2})^6 = \tfrac{1}{64}. \end{aligned}$$

In a similar manner, we find that the distribution function of Y is

$$\begin{aligned} G(y) = \Pr(Y \leq y) &= 0, \qquad y < 0, \\ &= y^6, \qquad 0 \leq y < 1, \\ &= 1, \qquad 1 \leq y. \end{aligned}$$

Accordingly, the p.d.f. of Y is

$$\begin{aligned} g(y) &= 6y^5, \qquad 0 < y < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Example 6. Let a fair coin be tossed at random on successive independent trials. Let the random variable $X_i = 1$ or $X_i = 0$ according to whether the outcome on the ith toss is a head or a tail, $i = 1, 2, 3, \ldots$. Let the p.d.f. of each X_i be $f(x) = \frac{1}{2}$, $x = 0, 1$, zero elsewhere. Since the trials are independent, we say that the random variables $X_1, X_2, X_3, \ldots$ are mutually stochastically independent. Then, for example, the probability that the first head appears on the third trial is

$$\Pr(X_1 = 0, X_2 = 0, X_3 = 1) = \Pr(X_1 = 0)\Pr(X_2 = 0)\Pr(X_3 = 1) = (\tfrac{1}{2})^3 = \tfrac{1}{8}.$$

In general, if Y is the number of the trial on which the first head appears, then the p.d.f. of Y is

$$g(y) = (\tfrac{1}{2})^y, \qquad y = 1, 2, 3, \ldots,$$
$$= 0 \text{ elsewhere.}$$

In particular, $\Pr(Y = 3) = g(3) = \frac{1}{8}$.

EXERCISES

2.27. Show that the random variables X_1 and X_2 with joint p.d.f. $f(x_1, x_2) = 12x_1x_2(1 - x_2)$, $0 < x_1 < 1$, $0 < x_2 < 1$, zero elsewhere, are stochastically independent.

2.28. If the random variables X_1 and X_2 have the joint p.d.f. $f(x_1, x_2) = 2e^{-x_1-x_2}$, $0 < x_1 < x_2$, $0 < x_2 < \infty$, zero elsewhere, show that X_1 and X_2 are stochastically dependent.

2.29. Let $f(x_1, x_2) = \frac{1}{16}$, $x_1 = 1, 2, 3, 4$, and $x_2 = 1, 2, 3, 4$, zero elsewhere, be the joint p.d.f. of X_1 and X_2. Show that X_1 and X_2 are stochastically independent.

2.30. Find $\Pr(0 < X_1 < \frac{1}{3}, 0 < X_2 < \frac{1}{3})$ if the random variables X_1 and X_2 have the joint p.d.f. $f(x_1, x_2) = 4x_1(1 - x_2)$, $0 < x_1 < 1$, $0 < x_2 < 1$, zero elsewhere.

2.31. Find the probability of the union of the events $a < X_1 < b$, $-\infty < X_2 < \infty$ and $-\infty < X_1 < \infty$, $c < X_2 < d$ if X_1 and X_2 are two stochastically independent variables with $\Pr(a < X_1 < b) = \frac{2}{3}$ and $\Pr(c < X_2 < d) = \frac{5}{8}$.

2.32. If $f(x_1, x_2) = e^{-x_1-x_2}$, $0 < x_1 < \infty$, $0 < x_2 < \infty$, zero elsewhere, is the joint p.d.f. of the random variables X_1 and X_2, show that X_1 and X_2 are stochastically independent and that

$$E(e^{t(X_1+X_2)}) = (1 - t)^{-2}, \qquad t < 1.$$

2.33. Let X_1, X_2, X_3, and X_4 be four mutually stochastically independent random variables, each with p.d.f. $f(x) = 3(1 - x)^2$, $0 < x < 1$, zero

elsewhere. If Y is the minimum of these four variables, find the distribution function and the p.d.f. of Y.

2.34. A fair die is cast at random three independent times. Let the random variable X_i be equal to the number of spots which appear on the ith trial, $i = 1, 2, 3$. Let the random variable Y be equal to max (X_i). Find the distribution function and the p.d.f. of Y. *Hint.* Pr $(Y \leq y) =$ Pr $(X_i \leq y, i = 1, 2, 3)$.

2.35. Suppose a man leaves for work between 8:00 A.M. and 8:30 A.M. and takes between 40 and 50 minutes to get to the office. Let X denote the time of departure and let Y denote the time of travel. If we assume that these random variables are stochastically independent and uniformly distributed, find the probability that he arrives at the office before 9:00 A.M.

2.36. Let $M(t_1, t_2, t_3)$ be the moment-generating function of the random variables X_1, X_2, and X_3 of Bernstein's example, described in the final remark of this section. Show that $M(t_1, t_2, 0) = M(t_1, 0, 0)M(0, t_2, 0)$, $M(t_1, 0, t_3) = M(t_1, 0, 0)M(0, 0, t_3)$, $M(0, t_2, t_3) = M(0, t_2, 0)M(0, 0, t_3)$, but $M(t_1, t_2, t_3) \neq M(t_1, 0, 0)M(0, t_2, 0)M(0, 0, t_3)$. Thus X_1, X_2, X_3 are pairwise stochastically independent but not mutually stochastically independent.

2.37. Generalize Theorem 1 of this chapter to the case of n mutually stochastically independent random variables.

2.38. Generalize Theorem 4 of this chapter to the case of n mutually stochastically independent random variables.

Chapter 3

Some Special Distributions

3.1 The Binomial Distribution

In Chapter 1 we introduced the *uniform distribution* and the *hypergeometric distribution*. In this chapter we shall discuss some other important distributions of random variables frequently used in statistics. We begin with the binomial distribution.

Recall, if n is a positive integer, that

$$(a + b)^n = \sum_{x=0}^{n} \binom{n}{x} b^x a^{n-x}.$$

Consider the function defined by

$$f(x) = \binom{n}{x} p^x (1 - p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n,$$
$$= 0 \text{ elsewhere},$$

where n is a positive integer and $0 < p < 1$. Under these conditions it is clear that $f(x) \geq 0$ and that

$$\sum_x f(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1 - p)^{n-x}$$
$$= [(1 - p) + p]^n = 1.$$

That is, $f(x)$ satisfies the conditions of being a p.d.f. of a random variable X of the discrete type. A random variable X that has a p.d.f. of the form of $f(x)$ is said to have a *binomial* distribution, and any such $f(x)$ is called a *binomial p.d.f.* A binomial distribution will be denoted by the symbol $b(n, p)$. The constants n and p are called the *parameters*

of the binomial distribution. Thus, if we say X is $b(5, \frac{1}{3})$, we mean that X has the binomial p.d.f.

$$f(x) = \binom{5}{x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{5-x}, \qquad x = 0, 1, \ldots, 5,$$
$$= 0 \text{ elsewhere.}$$

Remark. The binomial distribution serves as an excellent mathematical model in a number of experimental situations. Consider a random experiment, the outcome of which can be classified in but one of two mutually exclusive and exhaustive ways, say, success or failure (for example, head or tail, life or death, effective or noneffective, etc.). Let the random experiment be repeated n independent times. Assume further that the probability of success, say p, is the same on each repetition; thus, the probability of failure on each repetition is $1 - p$. Define the random variable X_i, $i = 1, 2, \ldots, n$, to be zero, if the outcome of the ith performance is a failure, and to be one if that outcome is a success. We then have $\Pr(X_i = 0) = 1 - p$ and $\Pr(X_i = 1) = p$, $i = 1, 2, \ldots, n$. Since it has been assumed that the experiment is to be repeated n independent times, the random variables $X_1, X_2, \ldots, X_n$ are mutually stochastically independent. According to the definition of X_i, the sum $Y = X_1 + X_2 + \cdots + X_n$ is the number of successes throughout the n repetitions of the random experiment. The following argument shows that Y has a binomial distribution: Let y be an element of $\{y; y = 0, 1, 2, \ldots, n\}$. Then $Y = y$ if and only if exactly y of the variables $X_1, X_2, \ldots, X_n$ have the value one, and each of the remaining $n - y$ variables is equal to zero. There are $\binom{n}{y}$ ways in which exactly y *ones* can be assigned to y of the variables $X_1, X_2, \ldots, X_n$. Since $X_1, X_2, \ldots, X_n$ are mutually stochastically independent the probability of each of these ways is $p^y(1 - p)^{n-y}$. Now $\Pr(Y = y)$ is the sum of the probabilities of these $\binom{n}{y}$ mutually exclusive events; that is,

$$\Pr(Y = y) = \binom{n}{y}p^y(1 - p)^{n-y}, \qquad y = 0, 1, 2, \ldots, n,$$

zero elsewhere. This is the p.d.f. of a binomial distribution.

The moment-generating function of a binomial distribution is easily found. It is

$$M(t) = \sum_x e^{tx}f(x) = \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x(1 - p)^{n-x}$$
$$= \sum_{x=0}^{n}\binom{n}{x}(pe^t)^x(1 - p)^{n-x}$$
$$= [(1 - p) + pe^t]^n$$

for all real values of t. The mean μ and the variance σ^2 of X may be computed from $M(t)$. Since

$$M'(t) = n[(1 - p) + pe^t]^{n-1}(pe^t)$$

and

$$M''(t) = n[(1 - p) + pe^t]^{n-1}(pe^t) + n(n - 1)[(1 - p) + pe^t]^{n-2}(pe^t)^2,$$

it follows that

$$\mu = M'(0) = np$$

and

$$\sigma^2 = M''(0) - \mu^2 = np + n(n - 1)p^2 - (np)^2 = np(1 - p).$$

Example 1. The binomial distribution with p.d.f.

$$f(x) = \binom{7}{x}\left(\frac{1}{2}\right)^x\left(1 - \frac{1}{2}\right)^{7-x}, \qquad x = 0, 1, 2, \ldots, 7,$$
$$= 0 \text{ elsewhere},$$

has the moment-generating function

$$M(t) = (\tfrac{1}{2} + \tfrac{1}{2}e^t)^7,$$

has mean $\mu = np = \frac{7}{2}$, and has variance $\sigma^2 = np(1 - p) = \frac{7}{4}$. Furthermore, if X is the random variable with this distribution, we have

$$\Pr(0 \leq X \leq 1) = \sum_{x=0}^{1} f(x) = \frac{1}{128} + \frac{7}{128} = \frac{8}{128}$$

and

$$\Pr(X = 5) = f(5)$$
$$= \frac{7!}{5!\,2!}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^2 = \frac{21}{128}.$$

Example 2. If the moment-generating function of a random variable X is

$$M(t) = (\tfrac{2}{3} + \tfrac{1}{3}e^t)^5,$$

then X has a binomial distribution with $n = 5$ and $p = \frac{1}{3}$; that is, the p.d.f. of X is

$$f(x) = \binom{5}{x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{5-x}, \qquad x = 0, 1, 2, \ldots, 5,$$
$$= 0 \text{ elsewhere}.$$

Here $\mu = np = \frac{5}{3}$ and $\sigma^2 = np(1 - p) = \frac{10}{9}$.

Example 3. If Y is $b(n, \frac{1}{3})$, then $\Pr(Y \geq 1) = 1 - \Pr(Y = 0) = 1 - (\frac{2}{3})^n$. Suppose we wish to find the smallest value of n that yields $\Pr(Y \geq 1) > 0.80$. We have $1 - (\frac{2}{3})^n > 0.80$ and $0.20 > (\frac{2}{3})^n$. Either by inspection or by use of logarithms, we see that $n = 4$ is the solution. That is,

the probability of at least one success throughout $n = 4$ independent repetitions of a random experiment with probability of success $p = \frac{1}{3}$ is greater than 0.80.

Example 4. Let the random variable Y be equal to the number of successes throughout n independent repetitions of a random experiment with probability p of success. That is, Y is $b(n, p)$. The ratio Y/n is called the relative frequency of success. For every $\epsilon > 0$, we have

$$\Pr\left(\left|\frac{Y}{n} - p\right| \geq \epsilon\right) = \Pr\left(|Y - np| \geq \epsilon n\right)$$

$$= \Pr\left(|Y - \mu| \geq \epsilon\sqrt{\frac{n}{p(1-p)}}\,\sigma\right),$$

where $\mu = np$ and $\sigma^2 = np(1 - p)$. In accordance with Chebyshev's inequality with $k = \epsilon\sqrt{n/p(1-p)}$, we have

$$\Pr\left(|Y - \mu| \geq \epsilon\sqrt{\frac{n}{p(1-p)}}\,\sigma\right) \leq \frac{p(1-p)}{n\epsilon^2}$$

and hence

$$\Pr\left(\left|\frac{Y}{n} - p\right| \geq \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2}.$$

Now, for every fixed $\epsilon > 0$, the right-hand member of the preceding inequality is close to zero for sufficiently large n. That is,

$$\lim_{n\to\infty} \Pr\left(\left|\frac{Y}{n} - p\right| \geq \epsilon\right) = 0$$

and

$$\lim_{n\to\infty} \Pr\left(\left|\frac{Y}{n} - p\right| < \epsilon\right) = 1.$$

Since this is true for every fixed $\epsilon > 0$, we see, in a certain sense, that the relative frequency of success is for large values of n, close to the probability p of success. This result is one form of the *law of large numbers*. It was alluded to in the initial discussion of probability in Chapter 1 and will be considered again, along with related concepts, in Chapter 5.

Example 5. Let the mutually stochastically independent random variables X_1, X_2, X_3 have the same distribution function $F(x)$. Let Y be the middle value of X_1, X_2, X_3. To determine the distribution function of Y, say $G(y) = \Pr(Y \leq y)$, we note that $Y \leq y$ if and only if at least two of the random variables X_1, X_2, X_3 are less than or equal to y. Let us say that the ith "trial" is a success if $X_i \leq y$, $i = 1, 2, 3$; here each "trial" has the probability of success $F(y)$. In this terminology, $G(y) = \Pr(Y \leq y)$ is then the probability of at least two successes in three independent trials. Thus,

$$G(y) = \binom{3}{2}[F(y)]^2[1 - F(y)] + [F(y)]^3.$$

If $F(x)$ is a continuous type of distribution function so that the p.d.f. of X is $F'(x) = f(x)$, then the p.d.f. of Y is

$$g(y) = G'(y) = 6[F(y)][1 - F(y)]f(y).$$

Example 6. Consider a sequence of independent repetitions of a random experiment with constant probability p of success. Let the random variable Y denote the total number of failures in this sequence before the rth success; that is, $Y + r$ is equal to the number of trials necessary to produce exactly r successes. Here r is a fixed positive integer. To determine the p.d.f. of Y, let y be an element of $\{y; y = 0, 1, 2, \ldots\}$. Then, by the multiplication rule of probabilities, $\Pr(Y = y) = g(y)$ is equal to the product of the probability

$$\binom{y + r - 1}{r - 1} p^{r-1}(1 - p)^y$$

of obtaining exactly $r - 1$ successes in the first $y + r - 1$ trials and the probability p of a success on the $(y + r)$th trial. Thus the p.d.f. $g(y)$ of Y is given by

$$g(y) = \binom{y + r - 1}{r - 1} p^r(1 - p)^y, \qquad y = 0, 1, 2, \ldots,$$
$$= 0 \text{ elsewhere.}$$

A distribution with a p.d.f. of the form $g(y)$ is called a *negative binomial distribution*; and any such $g(y)$ is called a negative binomial p.d.f. The distribution derives its name from the fact that $g(y)$ is a general term in the expansion of $p^r[1 - (1 - p)]^{-r}$. It is left as an exercise to show that the moment-generating function of this distribution is $M(t) = p^r[1 - (1 - p)e^t]^{-r}$, for $t < -\ln(1 - p)$. If $r = 1$, then Y has the p.d.f.

$$g(y) = p(1 - p)^y, \qquad y = 0, 1, 2, \ldots,$$

zero elsewhere, and the moment-generating function $M(t) = p[1 - (1 - p)e^t]^{-1}$. In this special case, $r = 1$, we say that Y has a *geometric distribution.*

The binomial distribution can be generalized to a distribution of more than one random variable. If n is a positive integer and a_1, a_2, a_3 are fixed constants, we have

$$(1) \quad \sum_{x=0}^{n} \sum_{y=0}^{n-x} \frac{n!}{x!\, y!\, (n - x - y)!} a_1^x a_2^y a_3^{n-x-y}$$
$$= \sum_{x=0}^{n} \frac{n!}{x!\, (n - x)!} a_1^x \sum_{y=0}^{n-x} \frac{(n - x)!}{y!\, (n - x - y)!} a_2^y a_3^{n-x-y}$$
$$= \sum_{x=0}^{n} \frac{n!}{x!\, (n - x)!} a_1^x (a_2 + a_3)^{n-x}$$
$$= (a_1 + a_2 + a_3)^n.$$

Let the function $f(x, y)$ be given by

$$f(x, y) = \frac{n!}{x!\, y!\, (n - x - y)!}\, p_1^x p_2^y p_3^{n-x-y},$$

where x and y are nonnegative integers with $x + y \leq n$, and p_1, p_2, and p_3 are positive proper fractions with $p_1 + p_2 + p_3 = 1$; and let $f(x, y) = 0$ elsewhere. Accordingly, $f(x, y)$ satisfies the conditions of being a joint p.d.f. of two random variables X and Y of the discrete type; that is, $f(x, y)$ is nonnegative and its sum over all points (x, y) at which $f(x, y)$ is positive is equal to $(p_1 + p_2 + p_3)^n = 1$. The random variables X and Y which have a joint p.d.f. of the form $f(x, y)$ are said to have a *trinomial* distribution and any such $f(x, y)$ is called a trinomial p.d.f. This distribution and the binomial distribution are special cases of the *multinomial* distribution. The moment-generating function of a trinomial distribution, in accordance with Equation (1), is given by

$$\begin{aligned} M(t_1, t_2) &= \sum_{x=0}^{n} \sum_{y=0}^{n-x} \frac{n!}{x!\, y!\, (n - x - y)!}\, (p_1 e^{t_1})^x (p_2 e^{t_2})^y p_3^{n-x-y} \\ &= (p_1 e^{t_1} + p_2 e^{t_2} + p_3)^n \end{aligned}$$

for all real values of t_1 and t_2. The moment-generating functions of the marginal distributions of X and Y are, respectively,

$$M(t_1, 0) = (p_1 e^{t_1} + p_2 + p_3)^n = [(1 - p_1) + p_1 e^{t_1}]^n$$

and

$$M(0, t_2) = (p_1 + p_2 e^{t_2} + p_3)^n = [(1 - p_2) + p_2 e^{t_2}]^n.$$

We see immediately, from Theorem 4, Section 2.4, that X and Y are stochastically dependent. In addition, X is $b(n, p_1)$ and Y is $b(n, p_2)$. Accordingly, the means and the variances of X and Y are, respectively, $\mu_1 = np_1$, $\mu_2 = np_2$, $\sigma_1^2 = np_1(1 - p_1)$, and $\sigma_2^2 = np_2(1 - p_2)$.

Consider next the conditional p.d.f. of Y, given $X = x$. We have

$$f(y|x) = \frac{(n - x)!}{y!\, (n - x - y)!} \left(\frac{p_2}{1 - p_1}\right)^y \left(\frac{p_3}{1 - p_1}\right)^{n-x-y}, \qquad y = 0, 1, \ldots, n - x,$$

$$= 0 \text{ elsewhere.}$$

Thus the conditional distribution of Y, given $X = x$, is $b[n - x, p_2/(1 - p_1)]$. Hence the conditional mean of Y, given $X = x$, is the linear function

$$E(Y|x) = (n - x)\left(\frac{p_2}{1 - p_1}\right).$$

Likewise, we find that the conditional distribution of X, given $Y = y$, is $b[n - y, p_1/(1 - p_2)]$ and thus

$$E(X|y) = (n - y)\left(\frac{p_1}{1 - p_2}\right).$$

Now recall (p. 73) that the square of the correlation coefficient, say ρ^2, is equal to the product of $-p_2/(1 - p_1)$ and $-p_1/(1 - p_2)$, the coefficients of x and y in the respective conditional means. Since both of these coefficients are negative (and thus ρ is negative), we have

$$\rho = -\sqrt{\frac{p_1 p_2}{(1 - p_1)(1 - p_2)}}.$$

Remark. Let a random experiment be repeated n independent times. On each repetition the experiment terminates in but one of k mutually exclusive and exhaustive ways, say $C_1, C_2, \ldots, C_k$. Let p_i be the probability that the outcome is an element of C_i and let p_i remain constant throughout the n independent repetitions, $i = 1, 2, \ldots, k$. Define the random variable X_i to be equal to the number of outcomes which are elements of C_i, $i = 1, 2, \ldots, k - 1$. Furthermore, let $x_1, x_2, \ldots, x_{k-1}$ be nonnegative integers so that $x_1 + x_2 + \cdots + x_{k-1} \leq n$. Then the probability that exactly x_1 terminations of the experiment are in $C_1, \ldots$, exactly x_{k-1} terminations are in C_{k-1}, and hence exactly $n - (x_1 + \cdots + x_{k-1})$ terminations are in C_k is

$$\frac{n!}{x_1! \cdots x_{k-1}!\, x_k!} p_1^{x_1} \cdots p_{k-1}^{x_{k-1}} p_k^{x_k},$$

where x_k is merely an abbreviation for $n - (x_1 + \cdots + x_{k-1})$. This is the *multinomial p.d.f.* of $k - 1$ random variables $X_1, X_2, \ldots, X_{k-1}$ of the discrete type. The moment-generating function of a multinomial distribution is given by

$$M(t_1, \ldots, t_{k-1}) = (p_1 e^{t_1} + \cdots + p_{k-1} e^{t_{k-1}} + p_k)^n$$

for all real values of $t_1, t_2, \ldots, t_{k-1}$. Thus each one-variable marginal p.d.f. is binomial, each two-variable marginal p.d.f. is trinomial, and so on.

EXERCISES

3.1. If the moment-generating function of a random variable X is $(\frac{1}{3} + \frac{2}{3}e^t)^5$, find Pr ($X = 2$ or 3).

3.2. The moment-generating function of a random variable X is $(\frac{2}{3} + \frac{1}{3}e^t)^9$. Show that

$$\Pr(\mu - 2\sigma < X < \mu + 2\sigma) = \sum_{x=1}^{5} \binom{9}{x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{9-x}.$$

3.3. If X is $b(n, p)$, show that

$$E\left(\frac{X}{n}\right) = p \qquad \text{and} \qquad E\left[\left(\frac{X}{n} - p\right)^2\right] = \frac{p(1-p)}{n}.$$

3.4. Let the mutually stochastically independent random variables X_1, X_2, X_3 have the same p.d.f. $f(x) = 3x^2$, $0 < x < 1$, zero elsewhere. Find the probability that exactly two of these three variables exceed $\frac{1}{2}$.

3.5. Let Y be the number of successes in n independent repetitions of a random experiment having the probability of success $p = \frac{2}{3}$. If $n = 3$, compute Pr $(2 \leq Y)$; if $n = 5$, compute Pr $(3 \leq Y)$.

3.6. Let Y be the number of successes throughout n independent repetitions of a random experiment having probability of success $p = \frac{1}{4}$. Determine the smallest value of n so that Pr $(1 \leq Y) \geq 0.70$.

3.7. Let the stochastically independent random variables X_1 and X_2 have binomial distributions with parameters $n_1 = 3$, $p_1 = \frac{2}{3}$ and $n_2 = 4$, $p_2 = \frac{1}{2}$, respectively. Compute Pr $(X_1 = X_2)$.

3.8. The joint moment-generating function of a multinomial distribution with $k = 4$ is given by

$$\sum_{x_3=0}^{n} \sum_{x_2=0}^{n-x_3} \sum_{x_1=0}^{n-x_2-x_3} \frac{n!}{x_1!\, x_2!\, x_3!\, x_4!} (p_1 e^{t_1})^{x_1} (p_2 e^{t_2})^{x_2} (p_3 e^{t_3})^{x_3} p_4^{x_4},$$

where $x_4 = n - x_1 - x_2 - x_3$ and $p_4 = 1 - p_1 - p_2 - p_3$. Show that this moment-generating function is $(p_4 + p_1 e^{t_1} + p_2 e^{t_2} + p_3 e^{t_3})^n$. Set $t_2 = t_3 = 0$, and show that the marginal distribution of X_1 is binomial with parameters n and p_1, and hence $E(X_1) = np_1$. Set $t_3 = 0$ and show that the marginal distribution of X_1 and X_2 is trinomial. Find $E(X_3 | x_1, x_2)$.

3.9. Let X be $b(2, p)$ and let Y be $b(4, p)$. If Pr $(X \geq 1) = \frac{5}{9}$, find Pr $(Y \geq 1)$.

3.10. If $x = r$ is the unique mode of a distribution which is $b(n, p)$, show that

$$(n + 1)p - 1 < r < (n + 1)p.$$

3.11. One of the numbers $1, 2, \ldots, 6$ is to be chosen by casting an unbiased die. Let this random experiment be repeated five independent times. Let the random variable X_1 be the number of terminations in the set $\{x; x = 1, 2, 3\}$ and let the random variable X_2 be the number of terminations in the set $\{x; x = 4, 5\}$. Compute Pr $(X_1 = 2, X_2 = 1)$.

3.12. Show that the moment-generating function of the negative binomial distribution is $M(t) = p^r[1 - (1 - p)e^t]^{-r}$. Find the mean and the variance of this distribution.

3.13. Let X_1 and X_2 have a trinomial distribution. Differentiate the moment-generating function to show that their covariance is $-np_1p_2$.

3.14. If a fair coin is tossed at random five independent times, find the conditional probability of five heads relative to the hypothesis that there are at least four heads.

3.15. Let an unbiased die be cast at random seven independent times. Compute the conditional probability that each side appears at least once relative to the hypothesis that the side "one" appears exactly twice.

3.16. Let X be $b(n, p)$ and let k be a nonnegative integer. If the kth moment about the origin is denoted by $\mu_k' = E(X^k)$ and if the kth moment about the mean is denoted by $\mu_k = E[(X - np)^k]$, prove that

(a) $$\mu_{k+1}' = np\mu_k' + p(1 - p)\frac{d\mu_k'}{dp},$$

(b) $$\mu_{k+1} = p(1 - p)\left[nk\mu_{k-1} + \frac{d\mu_k}{dp}\right].$$

Hint. Compute $d\mu_k'/dp$ and $d\mu_k/dp$.

3.17. Let

$$f(x_1, x_2) = \binom{x_1}{x_2}\left(\frac{1}{2}\right)^{x_1}\left(\frac{x_1}{15}\right), \qquad \begin{matrix} x_2 = 0, 1, \ldots, x_1, \\ x_1 = 1, 2, 3, 4, 5, \end{matrix}$$

zero elsewhere, be the joint p.d.f. of X_1 and X_2. Determine: (a) $E(X_2)$, (b) $u(x_1) = E(X_2|x_1)$, and (c) $E[u(X_1)]$. Compare the answers to parts (a) and (c).

3.18. Three fair dice are cast. In 10 independent casts, let X be the number of times all three faces are alike and let Y be the number of times only two faces are alike. Find the joint p.d.f. of X and Y and compute $E(6XY)$.

3.2 The Poisson Distribution

Recall that the series

$$1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \cdots = \sum_{x=0}^{\infty} \frac{m^x}{x!}$$

converges, for all values of m, to e^m. Consider the function $f(x)$ defined by

$$f(x) = \frac{m^x e^{-m}}{x!}, \qquad x = 0, 1, 2, \ldots,$$

$$= 0 \text{ elsewhere},$$

where $m > 0$. Since $m > 0$, then $f(x) \geq 0$ and

$$\sum_x f(x) = \sum_{x=0}^{\infty} \frac{m^x e^{-m}}{x!} = e^{-m} \sum_{x=0}^{\infty} \frac{m^x}{x!} = e^{-m}e^m = 1;$$

that is, $f(x)$ satisfies the conditions of being a p.d.f. of a discrete type of random variable. A random variable X which has a p.d.f. of the form of $f(x)$ is said to have a *Poisson* distribution, and any such $f(x)$ is called a *Poisson p.d.f.*

Remarks. Experience indicates that the Poisson p.d.f. may be used in a number of applications with quite satisfactory results. For example, let the random variable X denote the number of alpha particles emitted by a radioactive substance that enter a prescribed region during a prescribed interval of time. With a suitable value of m, it is found that X may be assumed to have a Poisson distribution. Again let the random variable X denote the number of defects on a manufactured article, such as a refrigerator door. Upon examining many of these doors, it is found, with an appropriate value of m, that X may be said to have a Poisson distribution. The number of railroad accidents in some unit of time (or the number of insurance claims in some unit of time) is often assumed to be a random variable which has a Poisson distribution. Each of these instances can be thought of as a process that generates a number of changes (accidents, claims, etc.) in a fixed interval (of time or space and so on). If a process leads to a Poisson distribution, that process is called a *Poisson process*. Some assumptions that ensure a Poisson process will now be enumerated.

Let $g(x, w)$ denote the probability of x changes in each interval of length w. Furthermore, let the symbol $o(h)$ represent any function such that $\lim_{h \to 0} [o(h)/h] = 0$; for example, $h^2 = o(h)$ and $o(h) + o(h) = o(h)$. The Poisson postulates are the following:

(a) $g(1, h) = \lambda h + o(h)$, where λ is a positive constant and $h > 0$.

(b) $\sum_{x=2}^{\infty} g(x, h) = o(h)$.

(c) The numbers of changes in nonoverlapping intervals are stochastically independent.

Postulates (a) and (c) state, in effect, that the probability of one change in a short interval h is independent of changes in other nonoverlapping intervals and is approximately proportional to the length of the interval. The substance of (b) is that the probability of two or more changes in the same short interval h is essentially equal to zero. If $x = 0$, we take $g(0, 0) = 1$. In accordance with postulates (a) and (b), the probability of at least one change in an interval of length h is $\lambda h + o(h) + o(h) = \lambda h + o(h)$. Hence the probability of zero changes in this interval of length h is $1 - \lambda h - o(h)$. Thus the probability $g(0, w + h)$ of zero changes in an interval of length $w + h$ is, in accordance with postulate (c), equal to the product of the probability $g(0, w)$ of zero changes in an interval of length w and the probability $[1 - \lambda h - o(h)]$ of zero changes in a nonoverlapping interval of length h. That is,

$$g(0, w + h) = g(0, w)[1 - \lambda h - o(h)].$$

Then

$$\frac{g(0, w + h) - g(0, w)}{h} = -\lambda g(0, w) - \frac{o(h)g(0, w)}{h}.$$

If we take the limit as $h \to 0$, we have

$$D_w[g(0, w)] = -\lambda g(0, w).$$

The solution of this differential equation is

$$g(0, w) = ce^{-\lambda w}.$$

The condition $g(0, 0) = 1$ implies that $c = 1$; so

$$g(0, w) = e^{-\lambda w}.$$

If x is a positive integer, we take $g(x, 0) = 0$. The postulates imply that

$$g(x, w + h) = [g(x, w)][1 - \lambda h - o(h)] + [g(x - 1, w)][\lambda h + o(h)] + o(h).$$

Accordingly, we have

$$\frac{g(x, w + h) - g(x, w)}{h} = -\lambda g(x, w) + \lambda g(x - 1, w) + \frac{o(h)}{h}$$

and

$$D_w[g(x, w)] = -\lambda g(x, w) + \lambda g(x - 1, w),$$

for $x = 1, 2, 3, \ldots$. It can be shown, by mathematical induction, that the solutions to these differential equations, with boundary conditions $g(x, 0) = 0$ for $x = 1, 2, 3, \ldots$, are, respectively,

$$g(x, w) = \frac{(\lambda w)^x e^{-\lambda w}}{x!}, \qquad x = 1, 2, 3, \ldots.$$

Hence, the number of changes X in an interval of length w has a Poisson distribution with parameter $m = \lambda w$.

The moment-generating function of a Poisson distribution is given by

$$\begin{aligned} M_x(t) &= \sum_x e^{tx} f(x) = \sum_{x=0}^{\infty} e^{tx} \frac{m^x e^{-m}}{x!} \\ &= e^{-m} \sum_{x=0}^{\infty} \frac{(me^t)^x}{x!} \\ &= e^{-m} e^{me^t} = e^{m(e^t - 1)} \end{aligned}$$

for all real values of t. Since

$$M'(t) = e^{m(e^t - 1)}(me^t)$$

and

$$M''(t) = e^{m(e^t - 1)}(me^t) + e^{m(e^t - 1)}(me^t)^2,$$

then

$$\mu = M'(0) = m$$

and

$$\sigma^2 = M''(0) - \mu^2 = m + m^2 - m^2 = m.$$

That is, a Poisson distribution has $\mu = \sigma^2 = m > 0$. On this account, a Poisson p.d.f. is frequently written

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}, \qquad x = 0, 1, 2, \ldots,$$
$$= 0 \text{ elsewhere.}$$

Thus the parameter m in a Poisson p.d.f. is the mean μ. Table I in the Appendix gives approximately the distribution function of the Poisson distribution for various values of the parameter $m = \mu$.

Example 1. Suppose X has a Poisson distribution with $\mu = 2$. Then the p.d.f. of X is

$$f(x) = \frac{2^x e^{-2}}{x!}, \qquad x = 0, 1, 2, \ldots,$$
$$= 0 \text{ elsewhere.}$$

The variance of this distribution is $\sigma^2 = \mu = 2$. If we wish to compute $\Pr(1 \leq X)$, we have

$$\Pr(1 \leq X) = 1 - \Pr(X = 0)$$
$$= 1 - f(0) = 1 - e^{-2} = 0.865,$$

approximately, by Table I of the Appendix.

Example 2. If the moment-generating function of a random variable X is

$$M(t) = e^{4(e^t - 1)},$$

then X has a Poisson distribution with $\mu = 4$. Accordingly, by way of example,

$$\Pr(X = 3) = \frac{4^3 e^{-4}}{3!} = \frac{32}{3} e^{-4};$$

or, by Table I,

$$\Pr(X = 3) = \Pr(X \leq 3) - \Pr(X \leq 2) = 0.433 - 0.238 = 0.195.$$

Example 3. Let the probability of exactly one blemish in 1 foot of wire be about $\frac{1}{1000}$ and let the probability of two or more blemishes in that length be, for all practical purposes, zero. Let the random variable X be the number of blemishes in 3000 feet of wire. If we assume the stochastic independence of the numbers of blemishes in nonoverlapping intervals, then the postulates of the Poisson process are approximated, with $\lambda = \frac{1}{1000}$ and

$w = 3000$. Thus X has an approximate Poisson distribution with mean $3000(\frac{1}{1000}) = 3$. Thus, for illustration, the probability that there are exactly five blemishes in 3000 feet of wire is

$$\Pr(X = 5) = \frac{3^5 e^{-3}}{5!}$$

and, by Table I of the Appendix,

$$\Pr(X = 5) = \Pr(X \leq 5) - \Pr(X \leq 4) = 0.101,$$

approximately.

EXERCISES

3.19. If the random variable X has a Poisson distribution such that $\Pr(X = 1) = \Pr(X = 2)$, find $\Pr(X = 4)$.

3.20. The moment-generating function of a random variable X is $e^{4(e^t - 1)}$. Show that $\Pr(\mu - 2\sigma < X < \mu + 2\sigma) = 0.931$.

3.21. In a lengthy manuscript, it is discovered that only 13.5 per cent of the pages contain no typing errors. If we assume that the number of errors per page is a random variable with a Poisson distribution, find the percentage of pages which have exactly one error.

3.22. Let the p.d.f. $f(x)$ be positive on and only on the nonnegative integers. Given that $f(x) = (4/x)f(x - 1)$, $x = 1, 2, 3, \ldots$. Find $f(x)$.

3.23. Let X have a Poisson distribution with $\mu = 100$. Use Chebyshev's inequality to determine a lower bound for $\Pr(75 < X < 125)$.

3.24. Given that $g(x, 0) = 0$ and that

$$D_w[g(x, w)] = -\lambda g(x, w) + \lambda g(x - 1, w)$$

for $x = 1, 2, 3, \ldots$. If $g(0, w) = e^{-\lambda w}$, show, by mathematical induction, that

$$g(x, w) = \frac{(\lambda w)^x e^{-\lambda w}}{x!}, \qquad x = 1, 2, 3, \ldots.$$

3.25. Let the number of chocolate drops in a certain type of cookie have a Poisson distribution. We want the probability that a cookie of this type contains at least two chocolate drops to be greater than 0.99. Find the smallest value that the mean of the distribution can take.

3.26. Let X have a Poisson distribution with parameter $\mu > 0$. If k is a nonnegative integer and if $\mu'_k = E(X^k)$, prove that

$$\mu'_{k+1} = \mu\left(\mu'_k + \frac{d\mu'_k}{d\mu}\right).$$

3.27. Let X and Y have the joint p.d.f. $f(x, y) = e^{-2}/[x!(y - x)!]$, $y = 0, 1, 2, \ldots$; $x = 0, 1, \ldots, y$, zero elsewhere. (a) Find the moment-generating function $M(t_1, t_2)$ of this joint distribution. (b) Compute the means, the variances, and the correlation coefficient of X and Y. (c) Determine the conditional mean $E(X|y)$.

3.3 The Gamma and Chi-Square Distributions

In this section we introduce the gamma and chi-square distributions. It is proved in books on advanced calculus that the integral

$$\int_0^\infty y^{\alpha-1}e^{-y}\,dy$$

exists for $\alpha > 0$ and that the value of the integral is a positive number. The integral is called the gamma function of α, and we write

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}\,dy.$$

If $\alpha = 1$, clearly

$$\Gamma(1) = \int_0^\infty e^{-y}\,dy = 1.$$

If $\alpha > 1$, an integration by parts shows that

$$\Gamma(\alpha) = (\alpha - 1)\int_0^\infty y^{\alpha-2}e^{-y}\,dy = (\alpha - 1)\Gamma(\alpha - 1).$$

Accordingly, if α is a positive integer greater than one,

$$\Gamma(\alpha) = (\alpha - 1)(\alpha - 2)\cdots(3)(2)(1)\Gamma(1) = (\alpha - 1)!$$

Since $\Gamma(1) = 1$, this suggests that we take $0! = 1$ as we have done.

In the integral which defines $\Gamma(\alpha)$, let us introduce a new variable x by writing $y = x/\beta$, where $\beta > 0$. Then

$$\Gamma(\alpha) = \int_0^\infty \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-x/\beta}\left(\frac{1}{\beta}\right) dx,$$

or, equivalently,

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

Since $\alpha > 0$, $\beta > 0$, and $\Gamma(\alpha) > 0$, we see that

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

is a p.d.f. of a random variable of the continuous type. A random variable X that has a p.d.f. of this form is said to have a *gamma* distribution with parameters α and β; and any such $f(x)$ is called a gamma-type p.d.f.

Remark. The gamma distribution is frequently the probability model for waiting times; for instance, in life-testing, the waiting time until "death" is the random variable which frequently has a gamma distribution. To see this, let us assume the postulates of a Poisson process and let the interval of length w be a time interval. Specifically, let the random variable W be the time that is needed to obtain exactly k changes (possibly deaths), where k is a fixed positive integer. Then the distribution function of W is

$$G(w) = \Pr(W \leq w) = 1 - \Pr(W > w).$$

However, the event $W > w$, for $w > 0$, is equivalent to the event in which there are less than k changes in a time interval of length w. That is, if the random variable X is the number of changes in an interval of length w, then

$$\Pr(W > w) = \sum_{x=0}^{k-1} \Pr(X = x) = \sum_{x=0}^{k-1} \frac{(\lambda w)^x e^{-\lambda w}}{x!}.$$

It is left as an exercise to verify that

$$\int_{\lambda w}^{\infty} \frac{z^{k-1}e^{-z}}{(k-1)!}\,dz = \sum_{x=0}^{k-1} \frac{(\lambda w)^x e^{-\lambda w}}{x!}.$$

If, momentarily, we accept this result, we have, for $w > 0$,

$$G(w) = 1 - \int_{\lambda w}^{\infty} \frac{z^{k-1}e^{-z}}{\Gamma(k)}\,dz = \int_{0}^{\lambda w} \frac{z^{k-1}e^{-z}}{\Gamma(k)}\,dz,$$

and for $w \leq 0$, $G(w) = 0$. If we change the variable of integration in the integral that defines $G(w)$ by writing $z = \lambda y$, then

$$G(w) = \int_0^w \frac{\lambda^k y^{k-1} e^{-\lambda y}}{\Gamma(k)}\,dy, \qquad w > 0,$$

and $G(w) = 0$, $w \leq 0$. Accordingly, the p.d.f. of W is

$$g(w) = G'(w) = \frac{\lambda^k w^{k-1} e^{-\lambda w}}{\Gamma(k)}, \qquad 0 < w < \infty,$$
$$= 0 \text{ elsewhere.}$$

That is, W has a gamma distribution with $\alpha = k$ and $\beta = 1/\lambda$. If W is the waiting time until the first change, that is, if $k = 1$, the p.d.f. of W is

$$g(w) = \lambda e^{-\lambda w}, \qquad 0 < w < \infty,$$
$$= 0 \text{ elsewhere.}$$

We now find the moment-generating function of a gamma distribution. Since

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

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

we may set $y = x(1 - \beta t)/\beta$, $t < 1/\beta$, or $x = \beta y/(1 - \beta t)$, to obtain

$$M(t) = \int_0^\infty \frac{\beta/(1-\beta t)}{\Gamma(\alpha)\beta^\alpha} \left(\frac{\beta y}{1-\beta t}\right)^{\alpha-1} e^{-y}\, dy.$$

That is,

$$M(t) = \left(\frac{1}{1-\beta t}\right)^\alpha \int_0^\infty \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y}\, dy$$

$$= \frac{1}{(1-\beta t)^\alpha}, \qquad t < \frac{1}{\beta}.$$

Now

$$M'(t) = (-\alpha)(1 - \beta t)^{-\alpha-1}(-\beta)$$

and

$$M''(t) = (-\alpha)(-\alpha - 1)(1 - \beta t)^{-\alpha-2}(-\beta)^2.$$

Hence, for a gamma distribution, we have

$$\mu = M'(0) = \alpha\beta$$

and

$$\sigma^2 = M''(0) - \mu^2 = \alpha(\alpha + 1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2.$$

Example 1. Let the waiting time W have a gamma p.d.f. with $\alpha = k$ and $\beta = 1/\lambda$. Accordingly, $E(W) = k/\lambda$. If $k = 1$, then $E(W) = 1/\lambda$; that is, the expected waiting time for $k = 1$ changes is equal to the reciprocal of λ.

Example 2. Let X be a random variable such that

$$E(X^m) = \frac{(m+3)!}{3!}\, 3^m, \qquad m = 1, 2, 3, \ldots.$$

Then the moment-generating function of X is given by the series

$$M(t) = 1 + \frac{4!\,3}{3!\,1!}\, t + \frac{5!\,3^2}{3!\,2!}\, t^2 + \frac{6!\,3^3}{3!\,3!}\, t^3 + \cdots.$$

This, however, is the Maclaurin's series for $(1 - 3t)^{-4}$, provided $-1 < 3t < 1$ Accordingly, X has a gamma distribution with $\alpha = 4$ and $\beta = 3$.

Let us now consider the special case of the gamma distribution in which $\alpha = r/2$, where r is a positive integer, and $\beta = 2$. A random variable X of the continuous type that has the p.d.f.

$$f(x) = \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1}e^{-x/2}, \qquad 0 < x < \infty,$$
$$= 0 \text{ elsewhere},$$

and the moment-generating function

$$M(t) = (1 - 2t)^{-r/2}, \qquad t < \tfrac{1}{2},$$

is said to have a *chi-square distribution*; and any $f(x)$ of this form is called a chi-square p.d.f. The mean and the variance of a chi-square distribution are $\mu = \alpha\beta = (r/2)2 = r$ and $\sigma^2 = \alpha\beta^2 = (r/2)2^2 = 2r$, respectively. For no obvious reason, we call the parameter r the number of degrees of freedom of the chi-square distribution (or of the chi-square p.d.f.). Because the chi-square distribution has an important role in statistics and occurs so frequently, we write, for brevity, that X is $\chi^2(r)$ to mean that the random variable X has a chi-square distribution with r degrees of freedom.

Example 3. If X has the p.d.f.

$$f(x) = \tfrac{1}{4}xe^{-x/2}, \qquad 0 < x < \infty,$$
$$= 0 \text{ elsewhere},$$

then X is $\chi^2(4)$. Hence $\mu = 4$, $\sigma^2 = 8$, and $M(t) = (1 - 2t)^{-2}$, $t < \frac{1}{2}$.

Example 4. If X has the moment-generating function $M(t) = (1 - 2t)^{-8}$, $t < \frac{1}{2}$, then X is $\chi^2(16)$.

If the random variable X is $\chi^2(r)$, then, with $c_1 < c_2$, we have

$$\Pr(c_1 \leq X \leq c_2) = \Pr(X \leq c_2) - \Pr(X \leq c_1),$$

since $\Pr(X = c_1) = 0$. To compute such a probability, we need the value of an integral like

$$\Pr(X \leq x) = \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1}e^{-w/2}\,dw.$$

Tables of this integral for selected values of r and x have been prepared and are partially reproduced in Table II in the Appendix.

Example 5. Let X be $\chi^2(10)$. Then, by Table II of the Appendix, with $r = 10$,

$$\Pr(3.25 \leq X \leq 20.5) = \Pr(X \leq 20.5) - \Pr(X \leq 3.25)$$
$$= 0.975 - 0.025 = 0.95.$$

Again, by way of example, if $\Pr(a < X) = 0.05$, then $\Pr(X \leq a) = 0.95$ and, thus, $a = 18.3$ from Table II with $r = 10$.

Example 6. Let X have a gamma distribution with $\alpha = r/2$, where r is a positive integer, and $\beta > 0$. Define the random variable $Y = 2X/\beta$. We seek the p.d.f. of Y. Now the distribution function of Y is

$$G(y) = \Pr(Y \leq y) = \Pr\left(X \leq \frac{\beta y}{2}\right).$$

If $y \leq 0$, then $G(y) = 0$; but if $y > 0$, then

$$G(y) = \int_0^{\beta y/2} \frac{1}{\Gamma(r/2)\beta^{r/2}} x^{r/2-1} e^{-x/\beta}\, dx.$$

Accordingly, the p.d.f. of Y is

$$g(y) = G'(y) = \frac{\beta/2}{\Gamma(r/2)\beta^{r/2}} (\beta y/2)^{r/2-1} e^{-y/2}$$

$$= \frac{1}{\Gamma(r/2)2^{r/2}} y^{r/2-1} e^{-y/2}$$

if $y > 0$. That is, Y is $\chi^2(r)$.

EXERCISES

3.28. If $(1 - 2t)^{-6}$, $t < \frac{1}{2}$, is the moment-generating function of the random variable X, find $\Pr(X < 5.23)$.

3.29. If X is $\chi^2(5)$, determine the constants c and d so that $\Pr(c < X < d) = 0.95$ and $\Pr(X < c) = 0.025$.

3.30. If X has a gamma distribution with $\alpha = 3$ and $\beta = 4$, find $\Pr(3.28 < X < 25.2)$. *Hint.* Consider the probability of the equivalent event $1.64 < Y < 12.6$, where $Y = 2X/4 = X/2$.

3.31. Let X be a random variable such that $E(X^m) = (m + 1)!\, 2^m$, $m = 1, 2, 3, \ldots$. Determine the distribution of X.

3.32. Show that

$$\int_\mu^\infty \frac{1}{\Gamma(k)} z^{k-1} e^{-z}\, dz = \sum_{x=0}^{k-1} \frac{\mu^x e^{-\mu}}{x!}, \qquad k = 1, 2, 3, \ldots.$$

3.33. Let X_1, X_2, X_3 be mutually stochastically independent random variables, each with p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Find the distribution of $Y = \text{minimum}(X_1, X_2, X_3)$.

3.34. Let X have a gamma distribution with p.d.f.

$$f(x) = \frac{1}{\beta^2} x e^{-x/\beta}, \qquad 0 < x < \infty,$$

zero elsewhere. If $x = 2$ is the unique mode of the distribution, find the parameter β and $\Pr(X < 9.49)$.

3.35. Let X have the uniform distribution with p.d.f.

$$f(x) = \frac{1}{2\theta_2}, \qquad \theta_1 - \theta_2 < x < \theta_1 + \theta_2, \ 0 < \theta_2,$$

zero elsewhere. Find θ_1 and θ_2 so that the mean and the variance of X are, respectively, equal to the mean and the variance of a distribution which is $\chi^2(8)$.

3.36. Let X have a Poisson distribution with parameter m. If m is an experimental value of a random variable having a gamma distribution with $\alpha = 2$ and $\beta = 1$, compute $\Pr(X = 0, 1, 2)$.

3.37. Let X have the uniform distribution with p.d.f. $f(x) = 1, 0 < x < 1$, zero elsewhere. Find the distribution function of $Y = -2 \ln X$. What is the p.d.f. of Y?

3.4 The Normal Distribution

Consider the integral

$$I = \int_{-\infty}^{\infty} \exp(-y^2/2)\, dy.$$

This integral exists because the integrand is a positive continuous function which is bounded by an integrable function; that is,

$$0 < \exp(-y^2/2) < \exp(-|y| + 1), \qquad -\infty < y < \infty,$$

and

$$\int_{-\infty}^{\infty} \exp(-|y| + 1)\, dy = 2e.$$

To evaluate the integral I, we note that $I > 0$ and that I^2 may be written

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-\frac{y^2 + z^2}{2}\right) dy\, dz.$$

This iterated integral can be evaluated by changing to polar coordinates. If we set $y = r\cos\theta$ and $z = r\sin\theta$, we have

$$\begin{aligned} I^2 &= \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r\, dr\, d\theta \\ &= \int_0^{2\pi} d\theta = 2\pi. \end{aligned}$$

Accordingly, $I = \sqrt{2\pi}$ and

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy = 1.$$

If we introduce a new variable of integration, say x, by writing

$$y = \frac{x - a}{b}, \qquad b > 0,$$

the preceding integral becomes

$$\int_{-\infty}^{\infty} \frac{1}{b\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2b^2}\right] dx = 1.$$

Since $b > 0$, this implies that

$$f(x) = \frac{1}{b\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2b^2}\right], \qquad -\infty < x < \infty$$

satisfies the conditions of being a p.d.f. of a continuous type of random variable. A random variable of the continuous type that has a p.d.f. of the form of $f(x)$ is said to have a *normal distribution*, and any $f(x)$ of this form is called a normal p.d.f.

We can find the moment-generating function of a normal distribution as follows. In

$$M(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{b\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2b^2}\right] dx$$

let $x = by + b^2t + a$. Then $M(t)$ becomes

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} e^{t(by+b^2t+a)} \frac{1}{b\sqrt{2\pi}} \exp\left[-\frac{(by+b^2t)^2}{2b^2}\right] b\, dy \\ &= \exp\left(at + \frac{b^2t^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy \\ &= \exp\left(at + \frac{b^2t^2}{2}\right) \end{aligned}$$

for all real values of t.

The mean μ and variance σ^2 of a normal distribution will be calculated from $M(t)$. Now

$$M'(t) = M(t)(a + b^2t)$$

and

$$M''(t) = M(t)(b^2) + M(t)(a + b^2t)^2.$$

Thus

$$\mu = M'(0) = a$$

and

$$\sigma^2 = M''(0) - \mu^2 = b^2 + a^2 - a^2 = b^2.$$

This permits us to write a normal p.d.f. in the form of

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad -\infty < x < \infty,$$

a form which shows explicitly the values of μ and σ^2. The moment-generating function $M(t)$ can be written

$$M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Example 1. If X has the moment-generating function

$$M(t) = e^{2t+32t^2},$$

then X has a normal distribution with $\mu = 2$, $\sigma^2 = 64$.

The normal p.d.f. occurs so frequently in certain parts of statistics that we denote it, for brevity, by $n(\mu, \sigma^2)$. Thus, if we say the random variable X is $n(0, 1)$, we mean that X has a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$, so that the p.d.f. of X is

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

If we say that X is $n(5, 4)$, we mean that X has a normal distribution with mean $\mu = 5$ and variance $\sigma^2 = 4$, so that the p.d.f. of X is

$$f(x) = \frac{1}{2\sqrt{2\pi}} \exp\left[-\frac{(x-5)^2}{2(4)}\right], \qquad -\infty < x < \infty.$$

Moreover, if

$$M(t) = e^{t^2/2},$$

then X is $n(0, 1)$.

The graph of

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad -\infty < x < \infty,$$

is seen (1) to be symmetric about a vertical axis through $x = \mu$, (2) to have its maximum of $1/(\sigma\sqrt{2\pi})$ at $x = \mu$, and (3) to have the x-axis as a horizontal asymptote. It should be verified that (4) there are points of inflection at $x = \mu \pm \sigma$.

Remark. Each of the special distributions considered thus far has been "justified" by some derivation which is based upon certain concepts found in elementary probability theory. Such a motivation for the normal distribution is not given at this time; a motivation is presented in Chapter 5. How-

ever, the normal distribution is one of the more widely used distributions in applications of statistical methods. Variables that are often assumed to be random variables having normal distributions (with appropriate values of μ and σ) are the diameter of a hole made by a drill press, the score on a test, the yield of a grain on a plot of ground, and the length of a newborn child.

We now prove a very useful theorem.

Theorem 1. *If the random variable X is $n(\mu, \sigma^2)$, $\sigma^2 > 0$, then the random variable $W = (X - \mu)/\sigma$ is $n(0, 1)$.*

Proof. The distribution function $G(w)$ of W is, since $\sigma > 0$,

$$G(w) = \Pr\left(\frac{X - \mu}{\sigma} \leq w\right) = \Pr(X \leq w\sigma + \mu).$$

That is,

$$G(w) = \int_{-\infty}^{w\sigma+\mu} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right] dx.$$

If we change the variable of integration by writing $y = (x - \mu)/\sigma$, then

$$G(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy.$$

Accordingly, the p.d.f. $g(w) = G'(w)$ of the continuous-type random variable W is

$$g(w) = \frac{1}{\sqrt{2\pi}} e^{-w^2/2}, \qquad -\infty < w < \infty.$$

Thus W is $n(0, 1)$, which is the desired result.

This fact considerably simplifies calculations of probabilities concerning normally distributed variables, as will be seen presently. Suppose that X is $n(\mu, \sigma^2)$. Then, with $c_1 < c_2$ we have, since $\Pr(X = c_1) = 0$,

$$\begin{aligned}\Pr(c_1 < X < c_2) &= \Pr(X < c_2) - \Pr(X < c_1)\\ &= \Pr\left(\frac{X - \mu}{\sigma} < \frac{c_2 - \mu}{\sigma}\right) - \Pr\left(\frac{X - \mu}{\sigma} < \frac{c_1 - \mu}{\sigma}\right)\\ &= \int_{-\infty}^{(c_2-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw - \int_{-\infty}^{(c_1-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw\end{aligned}$$

because $W = (X - \mu)/\sigma$ is $n(0, 1)$. That is, probabilities concerning X, which is $n(\mu, \sigma^2)$, can be expressed in terms of probabilities concerning W, which is $n(0, 1)$. However, an integral like

$$\int_{-\infty}^{k} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw$$

cannot be evaluated by the fundamental theorem of calculus because an "antiderivative" of $e^{-w^2/2}$ is not expressible as an elementary function. Instead, tables of the approximate value of this integral for various values of k have been prepared and are partially reproduced in Table III in the Appendix. We use the notation (for normal)

$$N(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw;$$

thus, if X is $n(\mu, \sigma^2)$, then

$$\Pr(c_1 < X < c_2) = \Pr\left(\frac{X - \mu}{\sigma} < \frac{c_2 - \mu}{\sigma}\right) - \Pr\left(\frac{X - \mu}{\sigma} < \frac{c_1 - \mu}{\sigma}\right)$$

$$= N\left(\frac{c_2 - \mu}{\sigma}\right) - N\left(\frac{c_1 - \mu}{\sigma}\right).$$

It is left as an exercise to show that $N(-x) = 1 - N(x)$.

Example 2. Let X be $n(2, 25)$. Then, by Table III,

$$\Pr(0 < X < 10) = N\left(\frac{10 - 2}{5}\right) - N\left(\frac{0 - 2}{5}\right)$$

$$= N(1.6) - N(-0.4)$$

$$= 0.945 - (1 - 0.655) = 0.600$$

and

$$\Pr(-8 < X < 1) = N\left(\frac{1 - 2}{5}\right) - N\left(\frac{-8 - 2}{5}\right)$$

$$= N(-0.2) - N(-2)$$

$$= (1 - 0.579) - (1 - 0.977) = 0.398.$$

Example 3. Let X be $n(\mu, \sigma^2)$. Then, by Table III,

$$\Pr(\mu - 2\sigma < X < \mu + 2\sigma) = N\left(\frac{\mu + 2\sigma - \mu}{\sigma}\right) - N\left(\frac{\mu - 2\sigma - \mu}{\sigma}\right)$$

$$= N(2) - N(-2)$$

$$= 0.977 - (1 - 0.977) = 0.954.$$

Example 4. Suppose that 10 per cent of the probability for a certain distribution which is $n(\mu, \sigma^2)$ is below 60 and that 5 per cent is above 90.

What are the values of μ and σ? We are given that the random variable X is $n(\mu, \sigma^2)$ and that $\Pr(X \leq 60) = 0.10$ and $\Pr(X \leq 90) = 0.95$. Thus $N[(60 - \mu)/\sigma] = 0.10$ and $N[(90 - \mu)/\sigma] = 0.95$. From Table III in the Appendix we have

$$\frac{60 - \mu}{\sigma} = -1.282, \qquad \frac{90 - \mu}{\sigma} = 1.645.$$

These conditions require that $\mu = 73.1$ and $\sigma = 10.2$ approximately.

We close this section with an important theorem.

Theorem 2. *If the random variable X is $n(\mu, \sigma^2)$, $\sigma^2 > 0$, then the random variable $V = (X - \mu)^2/\sigma^2$ is $\chi^2(1)$.*

Proof. Because $V = W^2$, where $W = (X - \mu)/\sigma$ is $n(0, 1)$, the distribution function $G(v)$ of V is, for $v \geq 0$,

$$G(v) = \Pr(W^2 \leq v) = \Pr(-\sqrt{v} \leq W \leq \sqrt{v}).$$

That is,

$$G(v) = 2\int_0^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw, \qquad 0 \leq v,$$

and

$$G(v) = 0, \qquad v < 0.$$

If we change the variable of integration by writing $w = \sqrt{y}$, then

$$G(v) = \int_0^v \frac{1}{\sqrt{2\pi}\sqrt{y}} e^{-y/2}\, dy, \qquad 0 \leq v.$$

Hence the p.d.f. $g(v) = G'(v)$ of the continuous-type random variable V is

$$g(v) = \frac{1}{\sqrt{\pi}\sqrt{2}} v^{1/2-1} e^{-v/2}, \qquad 0 < v < \infty,$$

$$= 0 \text{ elsewhere.}$$

Since $g(v)$ is a p.d.f. and hence

$$\int_0^\infty g(v)\, dv = 1,$$

it must be that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and thus V is $\chi^2(1)$.

EXERCISES

3.38. If

$$N(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw,$$

show that $N(-x) = 1 - N(x)$.

3.39. If X is $n(75, 100)$, find $\Pr(X < 60)$ and $\Pr(70 < X < 100)$.

3.40. If X is $n(\mu, \sigma^2)$, find b so that $\Pr[-b < (X - \mu)/\sigma < b] = 0.90$.

3.41. Let X be $n(\mu, \sigma^2)$ so that $\Pr(X < 89) = 0.90$ and $\Pr(X < 94) = 0.95$. Find μ and σ^2.

3.42. Determine the constant c so that $f(x) = ce^{-x^2+4x}$, $-\infty < x < \infty$, satisfies the conditions of being a p.d.f.

3.43. Show that the graph of a p.d.f. $n(\mu, \sigma^2)$ has points of inflection at $x = \mu - \sigma$ and $x = \mu + \sigma$.

3.44. Determine the ninetieth percentile of the distribution which is $n(65, 25)$.

3.45. If e^{3t+8t^2} is the moment-generating function of the random variable X, find $\Pr(-1 < X < 9)$.

3.46. Let the random variable X have the p.d.f.

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, \qquad 0 < x < \infty, \text{ zero elsewhere.}$$

Find the mean and variance of X.

3.47. Let X be $n(5, 10)$. Find $\Pr[0.04 < (X - 5)^2 < 38.4]$.

3.48. If X is $n(1, 4)$, compute the probability $\Pr(1 < X^2 < 9)$.

3.49. If X is $n(75, 25)$, find the conditional probability that X is greater than 80 relative to the hypothesis that X is greater than 77. See Exercise 2.17.

3.50. Let X be a random variable such that $E(X^{2m}) = (2m)!/(2^m m!)$, $m = 1, 2, 3, \ldots$ and $E(X^{2m-1}) = 0$, $m = 1, 2, 3, \ldots$. Find the moment-generating function and the p.d.f. of X.

3.51. Let the mutually stochastically independent random variables X_1, X_2, and X_3 be $n(0, 1)$, $n(2, 4)$, and $n(-1, 1)$, respectively. Compute the probability that exactly two of these three variables are less than zero.

3.52. Let X be $n(\mu, \sigma^2)$ and let k be a nonnegative integer. If $\mu_k' = E(X^k)$ and if $\mu_{2k} = E[(X - \mu)^{2k}]$, prove that

(a) $$\mu_{k+2}' = 2\mu\mu_{k+1}' + (\sigma^2 - \mu^2)\mu_k' + \sigma^3 \frac{d\mu_k'}{d\sigma},$$

(b) $$\mu_{2k+2} = \sigma^2\mu_{2k} + \sigma^3 \frac{d\mu_{2k}}{d\sigma}.$$

3.53. Let X and Y be stochastically independent random variables, each with a distribution which is $n(0, 1)$. Let $Z = X + Y$. Find the integral that represents the distribution function $G(z) = \Pr(X + Y \leq z)$ of Z. Determine the p.d.f. of Z.

3.5 The Bivariate Normal Distribution

Let us investigate the function

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-q/2}, \qquad -\infty < x < \infty, \ -\infty < y < \infty,$$

where, with $\sigma_1 > 0$, $\sigma_2 > 0$, and $-1 < \rho < 1$,

$$q = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right].$$

At this point we do not know that the constants μ_1, μ_2, σ_1^2, σ_2^2, and ρ represent parameters of a distribution. As a matter of fact, we do not know that $f(x, y)$ has the properties of a joint p.d.f. It will now be shown that

(i) $f(x, y)$ is a joint p.d.f.;
(ii) X is $n(\mu_1, \sigma_1^2)$ and Y is $n(\mu_2, \sigma_2^2)$;
(iii) ρ is the correlation coefficient of X and Y.

A joint p.d.f. of this form is called a bivariate normal p.d.f., and the random variables X and Y are said to have a *bivariate normal distribution.*

That the nonnegative function $f(x, y)$ actually is a joint p.d.f. can be seen as follows: Define $f_1(x)$ by

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy.$$

Now

$$\begin{aligned}(1-\rho^2)q &= \left[\left(\frac{y-\mu_2}{\sigma_2}\right) - \rho\left(\frac{x-\mu_1}{\sigma_1}\right)\right]^2 + (1-\rho^2)\left(\frac{x-\mu_1}{\sigma_1}\right)^2 \\ &= \left(\frac{y-b}{\sigma_2}\right)^2 + (1-\rho^2)\left(\frac{x-\mu_1}{\sigma_1}\right)^2,\end{aligned}$$

where $b = \mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1)$. Thus

$$f_1(x) = \frac{\exp\left[-(x-\mu_1)^2/2\sigma_1^2\right]}{\sigma_1\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\exp\left\{-(y-b)^2/[2\sigma_2^2(1-\rho^2)]\right\}}{\sigma_2\sqrt{1-\rho^2}\sqrt{2\pi}}\, dy.$$

For the purpose of integration, the integrand of the integral in this expression for $f_1(x)$ may be considered a normal p.d.f. with mean b and variance $\sigma_2^2(1-\rho^2)$. Thus this integral is equal to one and

$$f_1(x) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right], \qquad -\infty < x < \infty.$$

Since

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dy\,dx = \int_{-\infty}^{\infty} f_1(x)\,dx = 1,$$

the nonnegative function $f(x, y)$ is a joint p.d.f. of two continuous-type random variables X and Y. Accordingly, the function $f_1(x)$ is the marginal p.d.f. of X, and X is seen to be $n(\mu_1, \sigma_1^2)$. In like manner, we see that Y is $n(\mu_2, \sigma_2^2)$.

Moreover, from the above development, we note that

$$f(x, y) = f_1(x)\left\{\frac{1}{\sigma_2\sqrt{1-\rho^2}\sqrt{2\pi}} \exp\left[-\frac{(y-b)^2}{2\sigma_2^2(1-\rho^2)}\right]\right\},$$

where $b = \mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1)$. Accordingly, the second factor in the right-hand member of the above equation is the conditional p.d.f. of Y, given that $X = x$. That is, the conditional p.d.f. of Y, given $X = x$, is itself normal with mean $\mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1)$ and variance $\sigma_2^2(1 - \rho^2)$. Thus, with a bivariate normal distribution, the conditional mean of Y, given $X = x$, is linear in x and is given by

$$E(Y|x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

Since the coefficient of x in this linear conditional mean $E(Y|x)$ is $\rho\sigma_2/\sigma_1$, and since σ_1 and σ_2 represent the respective standard deviations, the number ρ is in fact the correlation coefficient of X and Y. This follows from the result, established in Section 2.3, p. 71, that the coefficient of x in a general linear conditional mean $E(Y|x)$ is the product of the correlation coefficient and the ratio σ_2/σ_1.

Although the mean of the conditional distribution of Y, given $X = x$, depends upon x (unless $\rho = 0$), the variance $\sigma_2^2(1 - \rho^2)$ is the same for all real values of x. Thus, by way of example, given that $X = x$, the conditional probability that Y is within $(2.576)\sigma_2\sqrt{1-\rho^2}$ units of the conditional mean is 0.99, whatever the value of x. In this sense, most of the probability for the distribution of X and Y lies in the band

$$\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1) \pm (2.576)\sigma_2\sqrt{1-\rho^2}$$

about the graph of the linear conditional mean. For every fixed positive σ_2, the width of this band depends upon ρ. Because the band is narrow when ρ^2 is nearly one, we see that ρ does measure the intensity of the concentration of the probability for X and Y about the linear conditional mean. This is the fact to which we alluded in the remark of Section 2.3.

In a similar manner, we can show that the conditional distribution of X, given $Y = y$, is the normal distribution

$$n\left[\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2(1 - \rho^2)\right].$$

Example 1. Let us assume that in a certain population of married couples the height X_1 of the husband and the height X_2 of the wife have a bivariate normal distribution with parameters $\mu_1 = 5.8$ feet, $\mu_2 = 5.3$ feet, $\sigma_1 = \sigma_2 = 0.2$ feet, and $\rho = 0.6$. The conditional p.d.f. of X_2, given $x_1 = 6.3$, is normal with mean $5.3 + (0.6)(6.3 - 5.8) = 5.6$ and standard deviation $(0.2)\sqrt{(1 - 0.36)} = 0.16$. Accordingly, given that the height of the husband is 6.3 feet, the probability that his wife has a height between 5.28 and 5.92 feet is

$$\Pr(5.28 < X_2 < 5.92 | x_1 = 6.3) = N(2) - N(-2) = 0.954.$$

The moment-generating function of a bivariate normal distribution can be determined as follows. We have

$$\begin{aligned} M(t_1, t_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1 x + t_2 y} f(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} e^{t_1 x} f_1(x)\left[\int_{-\infty}^{\infty} e^{t_2 y} f(y|x)\, dy\right] dx \end{aligned}$$

for all real values of t_1 and t_2. The integral within the brackets is the moment-generating function of the conditional p.d.f. $f(y|x)$. Since $f(y|x)$ is a normal p.d.f. with mean $\mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1)$ and variance $\sigma_2^2(1 - \rho^2)$, then

$$\int_{-\infty}^{\infty} e^{t_2 y} f(y|x)\, dy = \exp\left\{t_2\left[\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)\right] + \frac{t_2^2\sigma_2^2(1 - \rho^2)}{2}\right\}.$$

Accordingly, $M(t_1, t_2)$ can be written in the form

$$\exp\left\{t_2\mu_2 - t_2\rho\frac{\sigma_2}{\sigma_1}\mu_1 + \frac{t_2^2\sigma_2^2(1 - \rho^2)}{2}\right\}\int_{-\infty}^{\infty}\exp\left[\left(t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}\right)x\right] f_1(x)\, dx.$$

But $E(e^{tX}) = \exp[\mu_1 t + (\sigma_1^2 t^2)/2]$ for all real values of t. Accordingly, if we set $t = t_1 + t_2\rho(\sigma_2/\sigma_1)$, we see that $M(t_1, t_2)$ is given by

$$\exp\left\{t_2\mu_2 - t_2\rho\frac{\sigma_2}{\sigma_1}\mu_1 + \frac{t_2^2\sigma_2^2(1 - \rho^2)}{2} + \mu_1\left(t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}\right) + \sigma_1^2\frac{\left(t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}\right)^2}{2}\right\}$$

or, equivalently,

$$M(t_1, t_2) = \exp\left(\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2 t_1^2 + 2\rho\sigma_1\sigma_2 t_1 t_2 + \sigma_2^2 t_2^2}{2}\right).$$

It is interesting to note that if, in this moment-generating function $M(t_1, t_2)$, the correlation coefficient ρ is set equal to zero, then

$$M(t_1, t_2) = M(t_1, 0)M(0, t_2).$$

Thus X and Y are stochastically independent when $\rho = 0$. If, conversely, $M(t_1, t_2) \equiv M(t_1, 0)M(0, t_2)$, we have $e^{\rho\sigma_1\sigma_2 t_1 t_2} = 1$. Since each of σ_1 and σ_2 is positive, then $\rho = 0$. Accordingly, we have the following theorem.

Theorem 3. *Let X and Y have a bivariate normal distribution with means μ_1 and μ_2, positive variances σ_1^2 and σ_2^2, and correlation coefficient ρ. Then X and Y are stochastically independent if and only if $\rho = 0$.*

As a matter of fact, if any two random variables are stochastically independent and have positive standard deviations, we have noted in Example 4 of Section 2.4 (p. 80) that $\rho = 0$. However, $\rho = 0$ does not in general imply that two variables are stochastically independent; this can be seen in Exercises 2.18(c) and 2.23. The importance of Theorem 3 lies in the fact that we now know when and only when two random variables that have a bivariate normal distribution are stochastically independent.

EXERCISES

3.54. Let X and Y have a bivariate normal distribution with parameters $\mu_1 = 3$, $\mu_2 = 1$, $\sigma_1^2 = 16$, $\sigma_2^2 = 25$, and $\rho = \frac{3}{5}$. Determine the following probabilities:

(a) $\Pr(3 < Y < 8)$.
(b) $\Pr(3 < Y < 8 | x = 7)$.
(c) $\Pr(-3 < X < 3)$.
(d) $\Pr(-3 < X < 3 | y = -4)$.

3.55. Let $f(x,y) = (1/2\pi)\exp[-\frac{1}{2}(x^2 + y^2)]\{1 + xy\exp[-\frac{1}{2}(x^2 + y^2 - 2)]\}$, where $-\infty < x < \infty$, $-\infty < y < \infty$. If $f(x, y)$ is a joint p.d.f., it is not a normal bivariate p.d.f. Show that $f(x, y)$ actually is a joint p.d.f. and that each marginal p.d.f. is normal. Thus, the fact that each marginal p.d.f. is normal does not imply that the joint p.d.f. is bivariate normal.

3.56. If $M(t_1, t_2)$ is the moment-generating function of a bivariate normal distribution, compute the covariance by using the formula

$$\frac{\partial^2 M(0, 0)}{\partial t_1 \, \partial t_2} - \frac{\partial M(0, 0)}{\partial t_1} \frac{\partial M(0, 0)}{\partial t_2}.$$

Now let $\psi(t_1, t_2) = \ln M(t_1, t_2)$. Show that $\partial^2 \psi(0, 0)/\partial t_1 \, \partial t_2$ gives this covariance directly.

3.57. Let X and Y have a bivariate normal distribution with parameters $\mu_1 = 5$, $\mu_2 = 10$, $\sigma_1^2 = 1$, $\sigma_2^2 = 25$, and $\rho > 0$. If $\Pr(4 < Y < 16 | x = 5) = 0.954$, determine ρ.

3.58. Let X and Y have a bivariate normal distribution with parameters $\mu_1 = 20$, $\mu_2 = 40$, $\sigma_1^2 = 9$, $\sigma_2^2 = 4$, and $\rho = 0.6$. Find the shortest interval for which 0.90 is the conditional probability that Y is in this interval, given that $X = 22$.

Chapter 4

Distributions of Functions of Random Variables

4.1 Sampling Theory

Let $X_1, X_2, \ldots, X_n$ denote n random variables that have the joint p.d.f. $f(x_1, x_2, \ldots, x_n)$. These variables may or may not be stochastically independent. Problems such as the following are very interesting in themselves; but more importantly, their solutions often provide the basis for making statistical inferences. Let Y be a random variable that is defined by a function of $X_1, X_2, \ldots, X_n$, say $Y = u(X_1, X_2, \ldots, X_n)$. Once the p.d.f. $f(x_1, x_2, \ldots, x_n)$ is given, can we find the p.d.f. of Y? In some of the preceding chapters, we have solved a few of these problems. Among them are the following two: If $n = 1$ and if X_1 is $n(\mu, \sigma^2)$, then $Y = (X_1 - \mu)/\sigma$ is $n(0, 1)$. If n is a positive integer, if the random variables X_i, $i = 1, 2, \ldots, n$, are mutually stochastically independent, and each X_i has the same p.d.f. $f(x) = p^x(1 - p)^{1-x}$, $x = 0, 1$, and zero elsewhere, and if $Y = \sum_1^n X_i$, then Y is $b(n, p)$. It should be observed that $Y = u(X_1) = (X_1 - \mu)/\sigma$ is a function of X_1 that depends upon the two parameters of the normal distribution; whereas $Y = u(X_1, X_2, \ldots, X_n) = \sum_1^n X_i$ does not depend upon p, the parameter of the common p.d.f. of the X_i, $i = 1, 2, \ldots, n$. The distinction that we make between these functions is brought out in the following definition:

Definition 1. A function of one or more random variables which does not depend upon any *unknown* parameter is called a *statistic*.

In accordance with this definition, the random variable $Y = \sum_1^n X_i$ discussed above is a statistic. But the random variable $Y = (X_1 - \mu)/\sigma$ is not a statistic unless μ and σ are known numbers. It should be noted that, although a statistic does not depend upon any unknown parameter, the *distribution* of that statistic may very well depend upon unknown parameters.

Remark. We remark, for the benefit of the more advanced student, that a statistic is usually defined to be a measurable function of the random variables. In this book, however, we wish to minimize the use of measure theoretic terminology so we have suppressed the modifier "measurable." It is quite clear that a statistic is a random variable. In fact, some probabilists avoid the use of the word "statistic" altogether, and they refer to a measureable function of random variables as a random variable. We decided to use the word "statistic" because the student will encounter it so frequently in books and journals.

We can motivate the study of the distribution of a statistic in the following way. Let a random variable X be defined on a sample space $\mathscr{C}$ and let the space of X be denoted by $\mathscr{A}$. In many situations confronting us, the distribution of X is not completely known. For instance, we may know the distribution except for the value of an unknown parameter. To obtain more information about this distribution (or the unknown parameter), we shall repeat under identical conditions the random experiment n independent times. Let the random variable X_i be a function of the ith outcome, $i = 1, 2, \ldots, n$. Then we call $X_1, X_2, \ldots, X_n$ the items of a random sample from the distribution under consideration. Suppose we can define a statistic $Y = u(X_1, X_2, \ldots, X_n)$ whose p.d.f. is found to be $g(y)$. Perhaps this p.d.f. shows that there is a great probability that Y has a value close to the unknown parameter. Once the experiment has been repeated in the manner indicated and we have $X_1 = x_1, \ldots, X_n = x_n$, then $y = u(x_1, x_2, \ldots, x_n)$ is a known number. It is to be hoped that this known number can in some manner be used to elicit information about the unknown parameter. Thus, a statistic may prove to be useful.

Remarks. Let the random variable X be defined as the diameter of a hole to be drilled by a certain drill press and let it be assumed that X has a normal distribution. Past experience with many drill presses makes this assumption plausible; but the assumption does not specify the mean μ nor the variance σ^2 of this normal distribution. The only way to obtain information about μ and σ^2 is to have recourse to experimentation. Thus we shall drill

a number, say $n = 20$, of these holes whose diameters will be $X_1, X_2, \ldots, X_{20}$. Then $X_1, X_2, \ldots, X_{20}$ is a random sample from the normal distribution under consideration. Once the holes have been drilled and the diameters measured, the 20 numbers may be used, as will be seen later, to elicit information about μ and σ^2.

The term "random sample" is now defined in a more formal manner.

Definition 2. Let $X_1, X_2, \ldots, X_n$ denote n mutually stochastically independent random variables, each of which has the same but possibly unknown p.d.f. $f(x)$; that is, the probability density functions of $X_1, X_2, \ldots, X_n$ are, respectively, $f_1(x_1) = f(x_1), f_2(x_2) = f(x_2), \ldots, f_n(x_n) = f(x_n)$, so that the joint p.d.f. is $f(x_1)f(x_2)\cdots f(x_n)$. The random variables $X_1, X_2, \ldots, X_n$ are then said to constitute a *random sample* from a distribution that has p.d.f. $f(x)$.

Later we shall define what we mean by a random sample from a distribution of more than one random variable.

Sometimes it is convenient to refer to a random sample of size n from a given distribution and, as has been remarked, to refer to $X_1, X_2, \ldots, X_n$ as the items of the random sample. A reexamination of Example 5 of Section 2.4 reveals that we found the p.d.f. of the statistic which is the maximum of the items of a random sample of size $n = 3$ from a distribution with p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. In the first Remark of Section 3.1 (p. 87) (and referred to in this section), we found the p.d.f. of the statistic which is the sum of the items of a random sample of size n from a distribution that has p.d.f. $f(x) = p^x(1 - p)^{1-x}$, $x = 0, 1$, zero elsewhere.

In this book, most of the statistics we encounter will be functions of the items of a random sample from a given distribution. We now define two important statistics of this type:

Definition 3. Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a given distribution. The statistic

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \sum_{i=1}^{n} \frac{X_i}{n}$$

is called the *mean* of the random sample, and the statistic

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n} = \sum_{i=1}^{n} \frac{X_i^2}{n} - \bar{X}^2$$

is called the *variance* of the random sample.

Remark. Many writers do not define the variance of a random sample as we have done but, instead, they take $S^2 = \sum_1^n (X_i - \bar{X})^2/(n-1)$. There are good reasons for doing this. But a certain price has to be paid, as we shall indicate. Let $x_1, x_2, \ldots, x_n$ denote experimental values of the random variable X which has the p.d.f. $f(x)$ and the distribution function $F(x)$. Thus we may look upon $x_1, x_2, \ldots, x_n$ as the experimental values of a random sample of size n from the given distribution. The *distribution of the sample* is then defined to be the distribution obtained by assigning a probability of $1/n$ to each of the points $x_1, x_2, \ldots, x_n$. This is a distribution of the discrete type. The corresponding distribution function will be denoted by $F_n(x)$ and it is a step function. If we let f_x denote the number of sample values which are less than or equal to x, then $F_n(x) = f_x/n$, so that $F_n(x)$ gives the relative frequency of the event $X \leq x$ in the set of n observations. The function $F_n(x)$ is often called the "empirical distribution function" and it has a number of uses.

Because the distribution of the sample is a discrete distribution, the mean and the variance have been defined and are, respectively, $\sum_1^n x_i/n = \bar{x}$ and $\sum_1^n (x_i - \bar{x})^2/n = s^2$. Thus if one finds the distribution of the sample and the associated empirical distribution function to be useful concepts, it would seem logically inconsistent to define the variance of a random sample in any way other than we have.

Random sampling distribution theory means the general problem of finding distributions of functions of the items of a random sample. Up to this point, the only method, other than direct probabilistic arguments, of finding the distribution of a function of one or more random variables is the *distribution function technique.* That is, if $X_1, X_2, \ldots, X_n$ are random variables, the distribution of $Y = u(X_1, X_2, \ldots, X_n)$ is determined by computing the distribution function of Y, namely,

$$G(y) = \Pr [u(X_1, X_2, \ldots, X_n) \leq y].$$

Even in what superficially appears to be a very simple problem, this can be quite tedious. This fact is illustrated in the next paragraph.

Let X_1, X_2, X_3 denote a random sample of size three from a distribution that is $n(0, 1)$. Let Y denote the statistic which is the sum of the squares of the sample items. The distribution function of Y is given by

$$G(y) = \Pr [X_1^2 + X_2^2 + X_3^2 \leq y].$$

If $y < 0$, then $G(y) = 0$. However, if $y \geq 0$, then

$$G(y) = \iiint_A \frac{1}{(2\pi)^{3/2}} \exp\left[-\frac{1}{2}(x_1^2 + x_2^2 + x_3^2)\right] dx_1\, dx_2\, dx_3,$$

where A is the set of points (x_1, x_2, x_3) interior to, or on the surface of, a sphere with center at $(0, 0, 0)$ and radius equal to $\sqrt{y}$. This is not a simple integral. We might hope to make progress by changing to spherical coordinates:

$$x_1 = \rho \cos\theta \sin\varphi, \qquad x_2 = \rho \sin\theta \sin\varphi, \qquad x_3 = \rho \cos\varphi,$$

where $\rho \geq 0$, $0 \leq \theta < 2\pi$, $0 \leq \varphi \leq \pi$. Then, for $y \geq 0$,

$$G(y) = \int_0^{\sqrt{y}} \int_0^{2\pi} \int_0^{\pi} \frac{1}{(2\pi)^{3/2}} e^{-\rho^2/2} \rho^2 \sin\varphi \, d\varphi \, d\theta \, d\rho$$

$$= \sqrt{\frac{2}{\pi}} \int_0^{\sqrt{y}} \rho^2 e^{-\rho^2/2} \, d\rho.$$

If we change the variable of integration by setting $\rho = \sqrt{w}$, we have

$$G(y) = \sqrt{\frac{2}{\pi}} \int_0^{y} \frac{\sqrt{w}}{2} e^{-w/2} \, dw,$$

for $y \geq 0$. Since Y is a random variable of the continuous type, the p.d.f. of Y is $g(y) = G'(y)$. Thus

$$g(y) = \frac{1}{\sqrt{2\pi}} y^{3/2-1} e^{-y/2}, \qquad 0 < y < \infty,$$

$$= 0 \text{ elsewhere.}$$

Because $\Gamma(\frac{3}{2}) = (\frac{1}{2})\Gamma(\frac{1}{2}) = (\frac{1}{2})\sqrt{\pi}$, and thus $\sqrt{2\pi} = \Gamma(\frac{3}{2})2^{3/2}$, we see that Y is $\chi^2(3)$.

The problem that we have just solved points up the desirability of having, if possible, various methods of determining the distribution of a function of random variables. We shall find that other techniques are available and that often a particular technique is vastly superior to the others in a given situation. These techniques will be discussed in subsequent sections.

EXERCISES

4.1. Show that

$$S^2 = \frac{1}{n}\sum_1^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_1^n X_i^2 - \bar{X}^2,$$

where $\bar{X} = \sum_1^n X_i/n$.

4.2. If a random experiment has been performed five times and it has been found that $X_1 = 2.1$, $X_2 = 2.5$, $X_3 = 2.2$, $X_4 = 2.5$, $X_5 = 2.7$, compute the mean $\bar{x}$ and variance s^2 of this sample.

4.3. Find the probability that exactly four items of a random sample of size 5 from the distribution having p.d.f. $f(x) = (x + 1)/2$, $-1 < x < 1$, zero elsewhere, exceed zero.

4.4. Let X_1, X_2, X_3 be a random sample of size 3 from a distribution that is $n(6, 4)$. Determine the probability that the largest sample item exceeds 8.

4.5. Let X_1, X_2 be a random sample from the distribution having p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. Find Pr $(X_1/X_2 \leq \frac{1}{2})$.

4.6. If the sample size is $n = 2$, find the constant c so that $S^2 = c(X_1 - X_2)^2$.

4.7. If $x_i = i$, $i = 1, 2, \ldots, n$, compute the values of $\bar{x} = \sum x_i/n$ and $s^2 = \sum (x_i - \bar{x})^2/n$.

4.8. Let $y_i = a + bx_i$, $i = 1, 2, \ldots, n$, where a and b are constants. Find $\bar{y} = \sum y_i/n$ and $s_y^2 = \sum (y_i - \bar{y})^2/n$ in terms of a, b, $\bar{x} = \sum x_i/n$, and $s_x^2 = \sum (x_i - \bar{x})^2/n$.

4.9. Let X_1 and X_2 denote a random sample of size 2 from a distribution which is $n(0, 1)$. Find the p.d.f. of $Y = X_1^2 + X_2^2$.

4.10. Let X_1, X_2 denote a random sample of size 2 from a distribution with p.d.f. $f(x) = \frac{1}{2}$, $0 < x < 2$, zero elsewhere. Find the joint p.d.f. of X_1 and X_2. Let $Y = X_1 + X_2$. Find the distribution function and the p.d.f. of Y.

4.11. Let X_1 and X_2 denote a random sample of size 2 from a distribution with p.d.f. $f(x) = 1$, $0 < x < 1$, zero elsewhere. Find the distribution function and the p.d.f. of $Y = X_1/X_2$.

4.12. Let X_1, X_2, X_3 be a random sample of size 3 from a distribution having p.d.f. $f(x) = 5x^4$, $0 < x < 1$, zero elsewhere. Let Y be the largest item in the sample. Find the distribution function and p.d.f. of Y.

4.13. Let X_1 and X_2 be items of a random sample from a distribution with p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. Evaluate the conditional probability Pr $(X_1 < X_2 | X_1 < 2X_2)$.

4.2 Transformations of Variables of the Discrete Type

An alternative method of finding the distribution of a function of one or more random variables is called the *change of variable technique.* There are some delicate questions (with particular reference to random variables of the continuous type) involved in this technique, and these make it desirable for us first to consider special cases.

Let X have the Poisson p.d.f.

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}, \qquad x = 0, 1, 2, \ldots,$$

$$= 0 \text{ elsewhere.}$$

As we have done before, let $\mathscr{A}$ denote the space $\mathscr{A} = \{x; x = 0, 1, 2, \ldots\}$, so that $\mathscr{A}$ is the set where $f(x) > 0$. Define a new random variable Y by $Y = 4X$. We wish to find the p.d.f. of Y by the change of variable technique. Let $y = 4x$. We call $y = 4x$ a transformation from x to y, and we say that the transformation maps the space $\mathscr{A}$ onto the space $\mathscr{B} = \{y; y = 0, 4, 8, 12, \ldots\}$. The space $\mathscr{B}$ is obtained by transforming each point in $\mathscr{A}$ in accordance with $y = 4x$. We note two things about this transformation: It is such that to each point in $\mathscr{A}$ there corresponds one, and only one, point in $\mathscr{B}$; and conversely, to each point in $\mathscr{B}$ there corresponds one, and only one, point in $\mathscr{A}$. That is, the transformation $y = 4x$ sets up a one-to-one correspondence between the points of $\mathscr{A}$ and those of $\mathscr{B}$. Any function $y = u(x)$ (not merely $y = 4x$) that maps a space $\mathscr{A}$ (not merely our $\mathscr{A}$) onto a space $\mathscr{B}$ (not merely our $\mathscr{B}$) such that there is a one-to-one correspondence between the points of $\mathscr{A}$ and those of $\mathscr{B}$ is called a *one-to-one transformation*. It is important to note that a one-to-one transformation, $y = u(x)$, implies that y is a single-valued function of x, and that x is a single-valued function of y. In our case this is obviously true, since $y = 4x$ and $x = (\frac{1}{4})y$.

Our problem is that of finding the p.d.f. $g(y)$ of the discrete type of random variable $Y = 4X$. Now $g(y) = \Pr(Y = y)$. Because there is a one-to-one correspondence between the points of $\mathscr{A}$ and those of $\mathscr{B}$, the event $Y = y$ or $4X = y$ can occur when, and only when, the event $X = (\frac{1}{4})y$ occurs. That is, the two events are equivalent and have the same probability. Hence

$$g(y) = \Pr(Y = y) = \Pr\left(X = \frac{y}{4}\right) = \frac{\mu^{y/4} e^{-\mu}}{(y/4)!}, \qquad y = 0, 4, 8, \ldots,$$

$$= 0 \text{ elsewhere.}$$

The foregoing detailed discussion should make the subsequent text easier to read. Let X be a random variable of the discrete type, having p.d.f. $f(x)$. Let $\mathscr{A}$ denote the set of discrete points, at each of which $f(x) > 0$, and let $y = u(x)$ define a one-to-one transformation that maps $\mathscr{A}$ onto $\mathscr{B}$. If we solve $y = u(x)$ for x in terms of y, say, $x = w(y)$, then for each $y \in \mathscr{B}$, we have $x = w(y) \in \mathscr{A}$. Consider the random variable

$Y = u(X)$. If $y \in \mathscr{B}$, then $x = w(y) \in \mathscr{A}$, and the events $Y = y$ [or $u(X) = y$] and $X = w(y)$ are equivalent. Accordingly, the p.d.f. of Y is

$$g(y) = \Pr(Y = y) = \Pr[X = w(y)] = f[w(y)], \qquad y \in \mathscr{B},$$
$$= 0 \text{ elsewhere.}$$

Example 1. Let X have the binomial p.d.f.

$$f(x) = \frac{3!}{x!\,(3-x)!}\left(\frac{2}{3}\right)^x\left(\frac{1}{3}\right)^{3-x}, \qquad x = 0, 1, 2, 3,$$
$$= 0 \text{ elsewhere.}$$

We seek the p.d.f. $g(y)$ of the random variable $Y = X^2$. The transformation $y = u(x) = x^2$ maps $\mathscr{A} = \{x; x = 0, 1, 2, 3\}$ onto $\mathscr{B} = \{y; y = 0, 1, 4, 9\}$. In general, $y = x^2$ does not define a one-to-one transformation; here, however, it does, for there are no negative values of x in $\mathscr{A} = \{x; x = 0, 1, 2, 3\}$. That is, we have the single-valued inverse function $x = w(y) = \sqrt{y}$ (not $-\sqrt{y}$), and so

$$g(y) = f(\sqrt{y}) = \frac{3!}{(\sqrt{y})!\,(3-\sqrt{y})!}\left(\frac{2}{3}\right)^{\sqrt{y}}\left(\frac{1}{3}\right)^{3-\sqrt{y}}, \qquad y = 0, 1, 4, 9,$$
$$= 0 \text{ elsewhere.}$$

There are no essential difficulties involved in a problem like the following: Let $f(x_1, x_2)$ be the joint p.d.f. of two discrete-type random variables X_1 and X_2 with $\mathscr{A}$ the (two-dimensional) set of points at which $f(x_1, x_2) > 0$. Let $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ define a one-to-one transformation which maps $\mathscr{A}$ onto $\mathscr{B}$. The joint p.d.f. of the two new random variables $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ is given by

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)], \qquad (y_1, y_2) \in \mathscr{B},$$
$$= 0 \text{ elsewhere,}$$

where $x_1 = w_1(y_1, y_2)$ and $x_2 = w_2(y_1, y_2)$ are the single-valued inverses of $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$. From this joint p.d.f. $g(y_1, y_2)$ we may obtain the marginal p.d.f. of Y_1 by summing on y_2 or the marginal p.d.f. of Y_2 by summing on y_1.

Perhaps it should be emphasized that the technique of change of variables involves the introduction of as many "new" variables as there were "old" variables. That is, suppose $f(x_1, x_2, x_3)$ is the joint p.d.f. of X_1, X_2, and X_3, with $\mathscr{A}$ the set where $f(x_1, x_2, x_3) > 0$. Let us say we seek the p.d.f. of $Y_1 = u_1(X_1, X_2, X_3)$. We would then define (if possible) $Y_2 = u_2(X_1, X_2, X_3)$ and $Y_3 = u_3(X_1, X_2, X_3)$, so that

$y_1 = u_1(x_1, x_2, x_3)$, $y_2 = u_2(x_1, x_2, x_3)$, $y_3 = u_3(x_1, x_2, x_3)$ define a one-to-one transformation of $\mathscr{A}$ onto $\mathscr{B}$. This would enable us to find the joint p.d.f. of Y_1, Y_2, and Y_3 from which we would get the marginal p.d.f. of Y_1 by summing on y_2 and y_3.

Example 2. Let X_1 and X_2 be two stochastically independent random variables which have Poisson distributions with means μ_1 and μ_2, respectively. The joint p.d.f. of X_1 and X_2 is

$$\frac{\mu_1^{x_1}\mu_2^{x_2}e^{-\mu_1-\mu_2}}{x_1!\,x_2!}, \qquad x_1 = 0, 1, 2, 3, \ldots, \quad x_2 = 0, 1, 2, 3, \ldots,$$

and is zero elsewhere. Thus, the space $\mathscr{A}$ is the set of points (x_1, x_2), where each of x_1 and x_2 is a nonnegative integer. We wish to find the p.d.f. of $Y_1 = X_1 + X_2$. If we use the change of variable technique, we need to define a second random variable Y_2. Because Y_2 is of no interest to us, let us choose it in such a way that we have a simple one-to-one transformation. For example, take $Y_2 = X_2$. Then $y_1 = x_1 + x_2$ and $y_2 = x_2$ represent a one-to-one transformation that maps $\mathscr{A}$ onto

$$\mathscr{B} = \{(y_1, y_2);\ y_2 = 0, 1, \ldots, y_1 \quad \text{and} \quad y_1 = 0, 1, 2, \ldots\}.$$

Note that, if $(y_1, y_2) \in \mathscr{B}$, then $0 \leq y_2 \leq y_1$. The inverse functions are given by $x_1 = y_1 - y_2$ and $x_2 = y_2$. Thus the joint p.d.f. of Y_1 and Y_2 is

$$g(y_1, y_2) = \frac{\mu_1^{y_1-y_2}\mu_2^{y_2}e^{-\mu_1-\mu_2}}{(y_1 - y_2)!\,y_2!}, \qquad (y_1, y_2) \in \mathscr{B},$$

and is zero elsewhere. Consequently, the marginal p.d.f. of Y_1 is given by

$$\begin{aligned} g_1(y_1) &= \sum_{y_2=0}^{y_1} g(y_1, y_2) \\ &= \frac{e^{-\mu_1-\mu_2}}{y_1!} \sum_{y_2=0}^{y_1} \frac{y_1!}{(y_1 - y_2)!\,y_2!}\,\mu_1^{y_1-y_2}\mu_2^{y_2} \\ &= \frac{(\mu_1 + \mu_2)^{y_1}e^{-\mu_1-\mu_2}}{y_1!}, \qquad y_1 = 0, 1, 2, \ldots, \end{aligned}$$

and is zero elsewhere. That is, $Y_1 = X_1 + X_2$ has a Poisson distribution with parameter $\mu_1 + \mu_2$.

EXERCISES

4.14. Let X have the p.d.f. $f(x) = \frac{1}{3}$, $x = 1, 2, 3$, zero elsewhere. Find the p.d.f. of $Y = 2X + 1$.

***4.15.** If $f(x_1, x_2) = (\frac{2}{3})^{x_1+x_2}(\frac{1}{3})^{2-x_1-x_2}$, $(x_1, x_2) = (0, 0), (0, 1), (1, 0), (1, 1)$, zero elsewhere, is the joint p.d.f. of X_1 and X_2, find the joint p.d.f. of $Y_1 = X_1 - X_2$ and $Y_2 = X_1 + X_2$.

4.16. Let X have the p.d.f. $f(x) = (\frac{1}{2})^x$, $x = 1, 2, 3, \ldots$, zero elsewhere. Find the p.d.f. of $Y = X^3$.

4.17. Let X_1 and X_2 have the joint p.d.f. $f(x_1, x_2) = x_1x_2/36$, $x_1 = 1, 2, 3$, and $x_2 = 1, 2, 3$, zero elsewhere. Find first the joint p.d.f. of $Y_1 = X_1X_2$ and $Y_2 = X_2$, and then find the marginal p.d.f. of Y_1.

4.18. Let the stochastically independent random variables X_1 and X_2 be $b(n_1, p)$ and $b(n_2, p)$, respectively. Find the joint p.d.f. of $Y_1 = X_1 + X_2$ and $Y_2 = X_2$, and then find the marginal p.d.f. of Y_1. *Hint.* Use the fact that

$$\sum_{w=0}^{k} \binom{n_1}{w}\binom{n_2}{k-w} = \binom{n_1 + n_2}{k}.$$

This can be proved by comparing the coefficients of x^k in each member of the identity $(1 + x)^{n_1}(1 + x)^{n_2} \equiv (1 + x)^{n_1+n_2}$.

4.19. Let X_1, X_2, X_3 be three mutually stochastically independent random variables which have Poisson distributions with means μ_1, μ_2, and μ_3, respectively. Let $Y_1 = X_1 + X_2 + X_3$. Show that Y_1 has a Poisson distribution with mean $\mu_1 + \mu_2 + \mu_3$. *Hint.* Show first that $X_1 + X_2$ is stochastically independent of X_3 and use the result of Example 2.

4.20. Let X_1 and X_2 be stochastically independent random variables of the discrete type with joint p.d.f. $f_1(x_1)f_2(x_2)$, $(x_1, x_2) \in \mathcal{A}$. Let $y_1 = u_1(x_1)$ and $y_2 = u_2(x_2)$ denote a one-to-one transformation which maps $\mathcal{A}$ onto $\mathcal{B}$. Show that $Y_1 = u_1(X_1)$ and $Y_2 = u_2(X_2)$ are stochastically independent.

4.3 Transformations of Variables of the Continuous Type

In the preceding section, we introduced the notion of a one-to-one transformation and the mapping of a set $\mathcal{A}$ onto a set $\mathcal{B}$ under that transformation. Those ideas were sufficient to enable us to find the distribution of a function of several random variables of the discrete type. In this section we shall examine the same problem when the random variables are of the continuous type. It is again helpful to begin with a special problem.

Example 1. Let X be a random variable of the continuous type, having p.d.f.

$$\begin{aligned} f(x) &= 2x, \qquad 0 < x < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Here $\mathcal{A}$ is the space $\{x; 0 < x < 1\}$ where $f(x) > 0$. Define the random variable Y by $Y = 8X^3$ and consider the transformation $y = 8x^3$. Under the

transformation $y = 8x^3$, the set $\mathscr{A}$ is mapped onto the set $\mathscr{B} = \{y; 0 < y < 8\}$, and, moreover, the transformation is one-to-one. For every $0 < a < b < 8$, the event $a < Y < b$ will occur when, and only when, the event $\frac{1}{2}\sqrt[3]{a} < X < \frac{1}{2}\sqrt[3]{b}$ occurs because there is a one-to-one correspondence between the points of $\mathscr{A}$ and $\mathscr{B}$. Thus

$$\begin{aligned} \Pr(a < Y < b) &= \Pr\left(\tfrac{1}{2}\sqrt[3]{a} < X < \tfrac{1}{2}\sqrt[3]{b}\right) \\ &= \int_{\sqrt[3]{a}/2}^{\sqrt[3]{b}/2} 2x\,dx. \end{aligned}$$

Let us rewrite this integral by changing the variable of integration from x to y by writing $y = 8x^3$ or $x = \frac{1}{2}\sqrt[3]{y}$. Now

$$\frac{dx}{dy} = \frac{1}{6y^{2/3}},$$

and, accordingly, we have

$$\begin{aligned} \Pr(a < Y < b) &= \int_a^b 2\left(\frac{\sqrt[3]{y}}{2}\right)\left(\frac{1}{6y^{2/3}}\right) dy \\ &= \int_a^b \frac{1}{6y^{1/3}}\,dy. \end{aligned}$$

Since this is true for every $0 < a < b < 8$, the p.d.f. $g(y)$ of Y is the integrand; that is,

$$\begin{aligned} g(y) &= \frac{1}{6y^{1/3}}, \qquad 0 < y < 8, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

It is worth noting that we found the p.d.f. of the random variable $Y = 8X^3$ by using a theorem on the change of variable in a definite integral. However, to obtain $g(y)$ we actually need only two things: (1) the set $\mathscr{B}$ of points y where $g(y) > 0$ and (2) the integrand of the integral on y to which $\Pr(a < Y < b)$ is equal. These can be found by two simple rules:

(a) Verify that the transformation $y = 8x^3$ maps $\mathscr{A} = \{x; 0 < x < 1\}$ onto $\mathscr{B} = \{y; 0 < y < 8\}$ and that the transformation is one-to-one.

(b) Determine $g(y)$ on this set $\mathscr{B}$ by substituting $\frac{1}{2}\sqrt[3]{y}$ for x in $f(x)$ and then multiplying this result by the derivative of $\frac{1}{2}\sqrt[3]{y}$. That is,

$$\begin{aligned} g(y) &= f\left(\frac{\sqrt[3]{y}}{2}\right)\frac{d[(\frac{1}{2})\sqrt[3]{y}]}{dy} = \frac{1}{6y^{1/3}}, \qquad 0 < y < 8, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

We shall accept a theorem in analysis on the change of variable in a definite integral to enable us to state a more general result. Let X be a random variable of the continuous type having p.d.f. $f(x)$. Let $\mathscr{A}$ be the one-dimensional space where $f(x) > 0$. Consider the random variable $Y = u(X)$, where $y = u(x)$ defines a one-to-one transformation which maps the set $\mathscr{A}$ onto the set $\mathscr{B}$. Let the inverse of $y = u(x)$ be denoted by $x = w(y)$, and let the derivative $dx/dy = w'(y)$ be continuous and not vanish for all points y in $\mathscr{B}$. Then the p.d.f. of the random variable $Y = u(X)$ is given by

$$\begin{aligned} g(y) &= f[w(y)]|w'(y)|, \qquad y \in \mathscr{B}, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

where $|w'(y)|$ represents the absolute value of $w'(y)$. This is precisely what we did in Example 1 of this section, except there we deliberately chose $y = 8x^3$ to be an increasing function so that

$$\frac{dx}{dy} = w'(y) = \frac{1}{6y^{2/3}}, \qquad 0 < y < 8,$$

is positive, and hence

$$\left|\frac{1}{6y^{2/3}}\right| = \frac{1}{6y^{2/3}}, \qquad 0 < y < 8.$$

Henceforth, we shall refer to $dx/dy = w'(y)$ as the Jacobian (denoted by J) of the transformation. In most mathematical areas, $J = w'(y)$ is referred to as the Jacobian of the inverse transformation $x = w(y)$, but in this book it will be called the Jacobian of the transformation, simply for convenience.

Example 2. Let X have the p.d.f.

$$\begin{aligned} f(x) &= 1, \qquad 0 < x < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

We are to show that the random variable $Y = -2 \ln X$ has a chi-square distribution with two degrees of freedom. Here the transformation is $y = u(x) = -2 \ln x$, so that $x = w(y) = e^{-y/2}$. The space $\mathscr{A}$ is $\mathscr{A} = \{x; 0 < x < 1\}$, which the one-to-one transformation $y = -2 \ln x$ maps onto $\mathscr{B} = \{y; 0 < y < \infty\}$. The Jacobian of the transformation is

$$J = \frac{dx}{dy} = w'(y) = -\tfrac{1}{2}e^{-y/2}.$$

Accordingly, the p.d.f. $g(y)$ of $Y = -2 \ln X$ is

$$\begin{aligned} g(y) &= f(e^{-y/2})|J| = \tfrac{1}{2}e^{-y/2}, \qquad 0 < y < \infty, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

a p.d.f. which is chi-square with two degrees of freedom. Note that this problem was first proposed in Exercise 3.37 (p. 104).

This method of finding the p.d.f. of a function of one random variable of the continuous type will now be extended to functions of two random variables of this type. Again, only functions which define a one-to-one transformation will be considered at this time. Let $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ define a one-to-one transformation which maps a (two-dimensional) set $\mathcal{A}$ in the x_1x_2-plane onto a (two-dimensional) set $\mathcal{B}$ in the y_1y_2-plane. If we express each of x_1 and x_2 in terms of y_1 and y_2, we can write $x_1 = w_1(y_1, y_2)$, $x_2 = w_2(y_1, y_2)$. The determinant of order two,

$$\begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix},$$

is called the Jacobian of the transformation and will be denoted by the symbol J. It will be assumed that these first-order partial derivatives are continuous and that the Jacobian J is not identically equal to zero in $\mathcal{B}$. An illustrative example may be desirable before we proceed with the extension of the change of variable technique to two random variables of the continuous type.

Example 3. Let $\mathcal{A}$ be the set $\mathcal{A} = \{(x_1, x_2), 0 < x_1 < 1, 0 < x_2 < 1\}$ which is depicted in Figure 4.1. We wish to determine the set $\mathcal{B}$ in the y_1y_2-plane which is the mapping of $\mathcal{A}$ under the one-to-one transformation

$$y_1 = u_1(x_1, x_2) = x_1 + x_2,$$
$$y_2 = u_2(x_1, x_2) = x_1 - x_2,$$

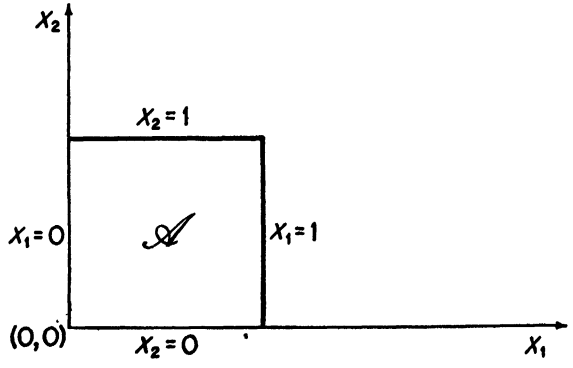


FIGURE 4.1

and we wish to compute the Jacobian of the transformation. Now

$$x_1 = w_1(y_1, y_2) = \tfrac{1}{2}(y_1 + y_2),$$
$$x_2 = w_2(y_1, y_2) = \tfrac{1}{2}(y_1 - y_2).$$

To determine the set $\mathscr{B}$ in the y_1y_2-plane onto which $\mathscr{A}$ is mapped under the transformation, note that the boundaries of $\mathscr{A}$ are transformed as follows into the boundaries of $\mathscr{B}$:

$$\begin{aligned} x_1 &= 0 \quad \text{into} \quad 0 = \tfrac{1}{2}(y_1 + y_2), \\ x_1 &= 1 \quad \text{into} \quad 1 = \tfrac{1}{2}(y_1 + y_2), \\ x_2 &= 0 \quad \text{into} \quad 0 = \tfrac{1}{2}(y_1 - y_2), \\ x_2 &= 1 \quad \text{into} \quad 1 = \tfrac{1}{2}(y_1 - y_2). \end{aligned}$$

Accordingly, $\mathscr{B}$ is as shown in Figure 4.2.

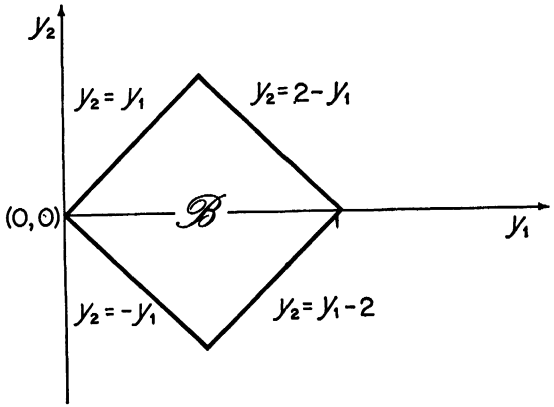


FIGURE 4.2

Finally,

$$J = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} \\ \dfrac{\partial x_2}{\partial y_1} & \dfrac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} \dfrac{1}{2} & \dfrac{1}{2} \\ \dfrac{1}{2} & -\dfrac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

We now proceed with the problem of finding the joint p.d.f. of two functions of two continuous-type random variables. Let X_1 and X_2 be random variables of the continuous type, having joint p.d.f. $\varphi(x_1, x_2)$. Let $\mathscr{A}$ be the two-dimensional set in the x_1x_2-plane where $\varphi(x_1, x_2) > 0$. Let $Y_1 = u_1(X_1, X_2)$ be a random variable whose p.d.f. is to be found. If $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ define a one-to-one transformation of $\mathscr{A}$ onto a set $\mathscr{B}$ in the y_1y_2-plane (with nonidentically vanishing

Jacobian), we can find, by use of a theorem in analysis, the joint p.d.f. of $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$. Let A be a subset of $\mathscr{A}$, and let B denote the mapping of A under the one-to-one transformation (see Figure 4.3).

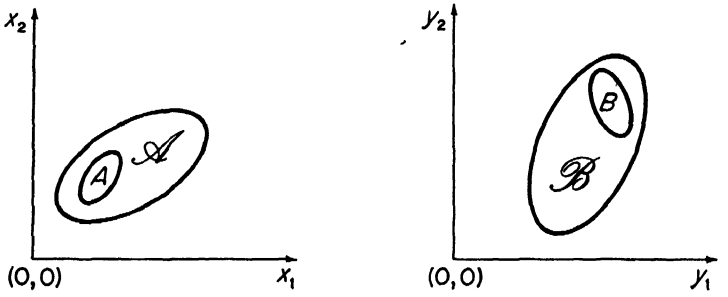


FIGURE 4.3

The events $(X_1, X_2) \in A$ and $(Y_1, Y_2) \in B$ are equivalent. Hence

$$\begin{aligned}\Pr[(Y_1, Y_2) \in B] &= \Pr[(X_1, X_2) \in A] \\ &= \int_A \int \varphi(x_1, x_2)\, dx_1\, dx_2.\end{aligned}$$

We wish now to change variables of integration by writing $y_1 = u_1(x_1, x_2)$, $y_2 = u_2(x_1, x_2)$, or $x_1 = w_1(y_1, y_2)$, $x_2 = w_2(y_1, y_2)$. It has been proved in analysis that this change of variables requires

$$\int_A \int \varphi(x_1, x_2)\, dx_1\, dx_2 = \int_B \int \varphi[w_1(y_1, y_2), w_2(y_1, y_2)]|J|\, dy_1\, dy_2.$$

Thus, for every set B in $\mathscr{B}$,

$$\Pr[(Y_1, Y_2) \in B] = \int_B \int \varphi[w_1(y_1, y_2), w_2(y_1, y_2)]|J|\, dy_1\, dy_2,$$

which implies that the joint p.d.f. $g(y_1, y_2)$ of Y_1 and Y_2 is

$$\begin{aligned}g(y_1, y_2) &= \varphi[w_1(y_1, y_2), w_2(y_1, y_2)]|J|, \qquad (y_1, y_2) \in \mathscr{B}, \\ &= 0 \text{ elsewhere.}\end{aligned}$$

Accordingly, the marginal p.d.f. $g_1(y_1)$ of Y_1 can be obtained from the joint p.d.f. $g(y_1, y_2)$ in the usual manner by integrating on y_2. Four examples of this result will be given.

Example 4. Let the random variable X have the p.d.f.

$$\begin{aligned}f(x) &= 1, \qquad 0 < x < 1, \\ &= 0 \text{ elsewhere,}\end{aligned}$$

and let X_1, X_2 denote a random sample from this distribution. The joint p.d.f. of X_1 and X_2 is then

$$\begin{aligned}\varphi(x_1, x_2) &= f(x_1)f(x_2) = 1, \qquad 0 < x_1 < 1, 0 < x_2 < 1,\\ &= 0 \text{ elsewhere.}\end{aligned}$$

Consider the two random variables $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. We wish to find the joint p.d.f. of Y_1 and Y_2. Here the two-dimensional space $\mathscr{A}$ in the x_1x_2-plane is that of Example 3 of this section. The one-to-one transformation $y_1 = x_1 + x_2$, $y_2 = x_1 - x_2$ maps $\mathscr{A}$ onto the space $\mathscr{B}$ of that example. Moreover, the Jacobian of that transformation has been shown to be $J = -\frac{1}{2}$. Thus

$$\begin{aligned}g(y_1, y_2) &= \varphi[\tfrac{1}{2}(y_1 + y_2), \tfrac{1}{2}(y_1 - y_2)]|J|\\ &= f[\tfrac{1}{2}(y_1 + y_2)]f[\tfrac{1}{2}(y_1 - y_2)]|J| = \tfrac{1}{2}, \qquad (y_1, y_2) \in \mathscr{B},\\ &= 0 \text{ elsewhere.}\end{aligned}$$

Because $\mathscr{B}$ is not a product space, the random variables Y_1 and Y_2 are stochastically dependent. The marginal p.d.f. of Y_1 is given by

$$g_1(y_1) = \int_{-\infty}^{\infty} g(y_1, y_2)\, dy_2.$$

If we refer to Figure 4.2, it is seen that

$$\begin{aligned}g_1(y_1) &= \int_{-y_1}^{y_1} \tfrac{1}{2}\, dy_2 = y_1, \qquad 0 < y_1 \leq 1,\\ &= \int_{y_1-2}^{2-y_1} \tfrac{1}{2}\, dy_2 = 2 - y_1, \qquad 1 < y_1 < 2,\\ &= 0 \text{ elsewhere.}\end{aligned}$$

In a similar manner, the marginal p.d.f. $g_2(y_2)$ is given by

$$\begin{aligned}g_2(y_2) &= \int_{-y_2}^{y_2+2} \tfrac{1}{2}\, dy_1 = y_2 + 1, \qquad -1 < y_2 \leq 0,\\ &= \int_{y_2}^{2-y_2} \tfrac{1}{2}\, dy_1 = 1 - y_2, \qquad 0 < y_2 < 1,\\ &= 0 \text{ elsewhere.}\end{aligned}$$

Example 5. Let X_1, X_2 be a random sample of size 2 from the distribution having p.d.f.

$$\begin{aligned}f(x) &= e^{-x}, \qquad 0 < x < \infty,\\ &= 0 \text{ elsewhere.}\end{aligned}$$

Let $Y_1 = X_1 + X_2$ and $Y_2 = X_1/(X_1 + X_2)$. We shall show that Y_1 and Y_2 are stochastically independent.

Since the joint p.d.f. of X_1 and X_2 is

$$\begin{aligned}\varphi(x_1, x_2) &= f(x_1)f(x_2) = e^{-x_1-x_2}, \qquad 0 < x_1 < \infty, \ 0 < x_2 < \infty,\\ &= 0 \text{ elsewhere,}\end{aligned}$$

the space $\mathscr{A}$ is, exclusive of the points on the coordinate axes, the first quadrant of the x_1x_2-plane. Now

$$y_1 = u_1(x_1, x_2) = x_1 + x_2,$$

$$y_2 = u_2(x_1, x_2) = \frac{x_1}{x_1 + x_2}$$

may be written $x_1 = y_1y_2$, $x_2 = y_1(1 - y_2)$ so that

$$J = \begin{vmatrix} y_2 & y_1 \\ 1 - y_2 & -y_1 \end{vmatrix} = -y_1 \neq 0.$$

The transformation is one-to-one, and it maps $\mathscr{A}$ onto $\mathscr{B} = \{(y_1, y_2); 0 < y_1 < \infty, 0 < y_2 < 1\}$ in the y_1y_2-plane. The joint p.d.f. of Y_1 and Y_2 is then

$$g(y_1, y_2) = y_1e^{-y_1}, \qquad 0 < y_1 < \infty, \ 0 < y_2 < 1,$$
$$= 0 \text{ elsewhere.}$$

In accordance with Theorem 1, Chapter 2 (p. 77) the random variables Y_1 and Y_2 are stochastically independent. The marginal probability density functions of Y_1 and Y_2 are, respectively, given by

$$g_1(y_1) = \int_0^1 y_1e^{-y_1}\,dy_2 = y_1e^{-y_1}, \qquad 0 < y_1 < \infty,$$
$$= 0 \text{ elsewhere,}$$

and

$$g_2(y_2) = \int_0^\infty y_1e^{-y_1}\,dy_1 = \Gamma(2) = 1, \qquad 0 < y_2 < 1,$$
$$= 0 \text{ elsewhere.}$$

Example 6. Let $Y_1 = \frac{1}{2}(X_1 - X_2)$, where X_1 and X_2 are stochastically independent random variables, each being $\chi^2(2)$. The joint p.d.f. of X_1 and X_2 is

$$f(x_1)f(x_2) = \tfrac{1}{4}\exp\left(-\frac{x_1 + x_2}{2}\right), \qquad 0 < x_1 < \infty, \ 0 < x_2 < \infty,$$
$$= 0 \text{ elsewhere.}$$

Let $Y_2 = X_2$ so that $y_1 = \frac{1}{2}(x_1 - x_2)$, $y_2 = x_2$, or $x_1 = 2y_1 + y_2$, $x_2 = y_2$ define a one-to-one transformation from $\mathscr{A} = \{(x_1, x_2); 0 < x_1 < \infty, 0 < x_2 < \infty\}$ onto $\mathscr{B} = \{(y_1, y_2); -2y_1 < y_2$ and $0 < y_2, -\infty < y_1 < \infty\}$. The Jacobian of the transformation is

$$J = \begin{vmatrix} 2 & 1 \\ 0 & 1 \end{vmatrix} = 2;$$

hence the joint p.d.f. of Y_1 and Y_2 is

$$g(y_1, y_2) = \frac{|2|}{4} e^{-y_1 - y_2}, \qquad (y_1, y_2) \in \mathscr{B},$$
$$= 0 \text{ elsewhere.}$$

Thus the p.d.f. of Y_1 is given by

$$g(y_1) = \int_{-2y_1}^{\infty} \tfrac{1}{2} e^{-y_1 - y_2}\, dy_2 = \tfrac{1}{2} e^{y_1}, \qquad -\infty < y_1 < 0,$$
$$= \int_{0}^{\infty} \tfrac{1}{2} e^{-y_1 - y_2}\, dy_2 = \tfrac{1}{2} e^{-y_1}, \qquad 0 \leq y_1 < \infty,$$

or

$$g_1(y_1) = \tfrac{1}{2} e^{-|y_1|}, \qquad -\infty < y_1 < \infty.$$

This p.d.f. is now frequently called the double exponential p.d.f.

Example 7. In this example, a rather important result is established. Let X_1 and X_2 be stochastically independent random variables of the continuous type with joint p.d.f. $f_1(x_1)f_2(x_2)$ that is positive on the two-dimensional space $\mathscr{A}$. Let $Y_1 = u_1(X_1)$, a function of X_1 alone, and $Y_2 = u_2(X_2)$, a function of X_2 alone. We assume for the present that $y_1 = u_1(x_1)$, $y_2 = u_2(x_2)$ define a one-to-one transformation from $\mathscr{A}$ onto a two-dimensional set $\mathscr{B}$ in the y_1y_2-plane. Thus, solving for x_1 and x_2 in terms of y_1 and y_2, we have $x_1 = w_1(y_1)$ and $x_2 = w_2(y_2)$ so that

$$J = \begin{vmatrix} w_1'(y_1) & 0 \\ 0 & w_2'(y_2) \end{vmatrix} = w_1'(y_1)w_2'(y_2) \neq 0.$$

Hence the joint p.d.f. of Y_1 and Y_2 is

$$g(y_1, y_2) = f_1[w_1(y_1)]f_2[w_2(y_2)]|w_1'(y_1)w_2'(y_2)|, \qquad (y_1, y_2) \in \mathscr{B},$$
$$= 0 \text{ elsewhere.}$$

However, from the procedure for changing variables in the case of one random variable, we see that the marginal probability density functions of Y_1 and Y_2 are, respectively, $g_1(y_1) = f_1[w_1(y_1)]|w_1'(y_1)|$ and $g_2(y_2) = f_2[w_2(y_2)]|w_2'(y_2)|$ for y_1 and y_2 in some appropriate sets. Consequently,

$$g(y_1, y_2) \equiv g_1(y_1)g_2(y_2).$$

Thus, summarizing, we note that, if X_1 and X_2 are stochastically independent random variables, then $Y_1 = u_1(X_1)$ and $Y_2 = u_2(X_2)$ are also stochastically independent random variables. It has been seen that the result holds if X_1 and X_2 are of the discrete type; see Exercise 4.20.

EXERCISES

4.21. Let X have the p.d.f. $f(x) = x^2/9$, $0 < x < 3$, zero elsewhere. Find the p.d.f. of $Y = X^3$.

4.22. If the p.d.f. of X is $f(x) = 2xe^{-x^2}$, $0 < x < \infty$, zero elsewhere, determine the p.d.f. of $Y = X^2$.

4.23. Let X be a random variable of the continuous type with p.d.f. $f(x)$. (a) Find the p.d.f. of $Y = X^2$. (b) If $f(x) = f(-x)$, for all x, simplify the answer of part (a). (c) If $f(x) = 0$ when $x \leq 0$, simplify the answer of part (a). Note that the transformation $y = x^2$ is not necessarily one-to-one. If this is too difficult now, return to the problem after Section 4.5.

4.24. Let X_1, X_2 be a random sample from the normal distribution $n(0, 1)$. Show that the marginal p.d.f. of $Y_1 = X_1/X_2$ is the *Cauchy p.d.f.*

$$g_1(y_1) = \frac{1}{\pi(1 + y_1^2)}, \qquad -\infty < y_1 < \infty.$$

Hint. Let $Y_2 = X_2$ and take the p.d.f. of X_2 to be equal to zero at $x_2 = 0$. Then determine the joint p.d.f. of Y_1 and Y_2. Be sure to multiply by the absolute value of the Jacobian.

4.25. Let X_1 and X_2 be two stochastically independent random variables that have gamma distributions and joint p.d.f.

$$f(x_1, x_2) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x_1^{\alpha-1} x_2^{\beta-1} e^{-x_1-x_2}, \qquad 0 < x_1 < \infty, 0 < x_2 < \infty,$$

zero elsewhere, where $\alpha > 0$, $\beta > 0$. Show that the marginal p.d.f. of $Y_1 = X_1/(X_1 + X_2)$ is the *beta p.d.f.*

$$g_1(y_1) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y_1^{\alpha-1}(1 - y_1)^{\beta-1}, \qquad 0 < y_1 < 1,$$

zero elsewhere. *Hint.* Let $Y_2 = X_1 + X_2$ and find first the joint p.d.f. of Y_1 and Y_2.

4.26. Find the mean and variance of the beta distribution considered in Exercise 4.25. *Hint.* From that exercise, we know

$$\int_0^1 y^{\alpha-1}(1 - y)^{\beta-1}\, dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

for all $\alpha > 0$, $\beta > 0$.

4.27. Determine the constant c in each of the following so that each $f(x)$ is a *beta p.d.f.*

(a) $f(x) = cx(1 - x)^3$, $0 < x < 1$, zero elsewhere.

(b) $f(x) = cx^4(1 - x)^5$, $0 < x < 1$, zero elsewhere.

(c) $f(x) = cx^2(1 - x)^8$, $0 < x < 1$, zero elsewhere.

4.28. Determine the constant c so that $f(x) = cx(3 - x)^4$, $0 < x < 3$, zero elsewhere, is a p.d.f.

4.29. Show that the graph of the beta p.d.f. is symmetric about the vertical line through $x = \frac{1}{2}$ if $\alpha = \beta$.

4.30. Let X have the uniform distribution over the interval $(-\pi/2, \pi/2)$. Show that $Y = \tan X$ has a Cauchy distribution.

4.31. Let X have the p.d.f. $f(x) = (x - 1)/2$, $1 < x < 3$, zero elsewhere. Find a monotone function $u(x)$ so that $Y = u(X)$ has a uniform distribution over the interval $(0, 1)$.

4.32. Let X_1 and X_2 denote a random sample of size 2 from a uniform distribution with p.d.f. $f(x) = \frac{1}{9}$, $1 < x < 10$, zero elsewhere. Find the p.d.f. of $Y_1 = X_1X_2$. Show that $\Pr(Y_1 \in A) > \frac{1}{9}$, where $A = \{y_1; 1 < y_1 < 2 \text{ or } 10 < y_1 < 20\}$. This problem is of interest in analyzing the random error in floating decimal-point computations.

4.33. Let X_1 and X_2 denote a random sample of size 2 from a distribution which is $n(\mu, \sigma^2)$. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Find the joint p.d.f. of Y_1 and Y_2 and show that these random variables are stochastically independent.

4.34. Let X_1 and X_2 denote a random sample of size 2 from a distribution which is $n(\mu, \sigma^2)$. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_1 + 2X_2$. Show that the joint p.d.f. of Y_1 and Y_2 is bivariate normal with correlation coefficient $3/\sqrt{10}$.

4.4 The t and F Distributions

It is the purpose of this section to define two additional distributions quite useful in certain problems of statistical inference. These are called, respectively, the (Student's) t distribution and the F distribution.

Let W denote a random variable which is $n(0, 1)$; let V denote a random variable which is $\chi^2(r)$; and let W and V be stochastically independent. Then the joint p.d.f. of W and V, say $\varphi(w, v)$, is the product of the p.d.f. of W and that of V or

$$\varphi(w, v) = \frac{1}{\sqrt{2\pi}} e^{-w^2/2} \frac{1}{\Gamma(r/2)2^{r/2}} v^{r/2-1}e^{-v/2},$$

$$-\infty < w < \infty, 0 < v < \infty,$$

$$= 0 \text{ elsewhere.}$$

Define a new random variable T by writing

$$T = \frac{W}{\sqrt{V/r}}.$$

The change of variable technique will be used to obtain the p.d.f. $g_1(t)$ of T. The equations

$$t = \frac{w}{\sqrt{v/r}} \qquad \text{and} \qquad u = v$$

define a one-to-one transformation which maps $\mathcal{A} = \{(w, v); -\infty < w < \infty, 0 < v < \infty\}$ onto $\mathcal{B} = \{(t, u); -\infty < t < \infty, 0 < u < \infty\}$. Since $w = t\sqrt{u}/\sqrt{r}$, $v = u$, the absolute value of the Jacobian of the transformation is $|J| = \sqrt{u}/\sqrt{r}$. Accordingly, the joint p.d.f. of T and $U = V$ is given by

$$\begin{aligned} g(t, u) &= \varphi\left(\frac{t\sqrt{u}}{\sqrt{r}}, u\right)|J| \\ &= \frac{1}{\sqrt{2\pi}\Gamma(r/2)2^{r/2}} u^{r/2-1} \exp\left[-\frac{u}{2}\left(1 + \frac{t^2}{r}\right)\right]\frac{\sqrt{u}}{\sqrt{r}}, \quad -\infty < t < \infty, 0 < u < \infty, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

The marginal p.d.f. of T is then

$$\begin{aligned} g_1(t) &= \int_{-\infty}^{\infty} g(t, u)\, du \\ &= \int_0^{\infty} \frac{1}{\sqrt{2\pi r}\Gamma(r/2)2^{r/2}} u^{(r+1)/2-1} \exp\left[-\frac{u}{2}\left(1 + \frac{t^2}{r}\right)\right] du. \end{aligned}$$

In this integral let $z = u[1 + (t^2/r)]/2$, and it is seen that

$$\begin{aligned} g_1(t) &= \int_0^{\infty} \frac{1}{\sqrt{2\pi r}\Gamma(r/2)2^{r/2}} \left(\frac{2z}{1 + t^2/r}\right)^{(r+1)/2-1} e^{-z}\left(\frac{2}{1 + t^2/r}\right) dz \\ &= \frac{\Gamma[(r + 1)/2]}{\sqrt{\pi r}\Gamma(r/2)} \frac{1}{(1 + t^2/r)^{(r+1)/2}}, \qquad -\infty < t < \infty. \end{aligned}$$

Thus, if W is $n(0, 1)$, if V is $\chi^2(r)$, and if W and V are stochastically independent, then

$$T = \frac{W}{\sqrt{V/r}}$$

has the immediately preceding p.d.f. $g_1(t)$. The distribution of the random variable T is usually called a t distribution. It should be observed that a t distribution is completely determined by the parameter r, the number of degrees of freedom of the random variable that has the chi-square distribution. Some approximate values of

$$\Pr(T \leq t) = \int_{-\infty}^{t} g_1(w)\, dw$$

for selected values of r and t, can be found in Table IV in the Appendix.

Next consider two stochastically independent chi-square random variables U and V having r_1 and r_2 degrees of freedom, respectively. The joint p.d.f. $\varphi(u, v)$ of U and V is then

$$\varphi(u, v) = \frac{1}{\Gamma(r_1/2)\Gamma(r_2/2)2^{(r_1+r_2)/2}} u^{r_1/2-1}v^{r_2/2-1}e^{-(u+v)/2}, \quad 0 < u < \infty, 0 < v < \infty,$$

$$= 0 \text{ elsewhere.}$$

We define the new random variable

$$F = \frac{U/r_1}{V/r_2}$$

and we propose finding the p.d.f. $g_1(f)$ of F. The equations

$$f = \frac{u/r_1}{v/r_2}, \qquad z = v,$$

define a one-to-one transformation which maps the set $\mathcal{A} = \{(u, v); 0 < u < \infty, 0 < v < \infty\}$ onto the set $\mathcal{B} = \{(f, z); 0 < f < \infty, 0 < z < \infty\}$. Since $u = (r_1/r_2)zf$, $v = z$, the absolute value of the Jacobian of the transformation is $|J| = (r_1/r_2)z$. The joint p.d.f. $g(f, z)$ of the random variables F and $Z = V$ is then

$$g(f, z) = \frac{1}{\Gamma(r_1/2)\Gamma(r_2/2)2^{(r_1+r_2)/2}} \left(\frac{r_1zf}{r_2}\right)^{r_1/2-1} z^{r_2/2-1} \times \exp\left[-\frac{z}{2}\left(\frac{r_1f}{r_2} + 1\right)\right]\frac{r_1z}{r_2},$$

provided $(f, z) \in \mathcal{B}$, and zero elsewhere. The marginal p.d.f. $g_1(f)$ of F is then

$$g_1(f) = \int_{-\infty}^{\infty} g(f, z)\, dz$$

$$= \int_0^{\infty} \frac{(r_1/r_2)^{r_1/2}(f)^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)2^{(r_1+r_2)/2}} z^{(r_1+r_2)/2-1} \exp\left[-\frac{z}{2}\left(\frac{r_1f}{r_2} + 1\right)\right] dz.$$

If we change the variable of integration by writing

$$y = \frac{z}{2}\left(\frac{r_1f}{r_2} + 1\right),$$

it can be seen that

$$g_1(f) = \int_0^\infty \frac{(r_1/r_2)^{r_1/2}(f)^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)2^{(r_1+r_2)/2}} \left(\frac{2y}{r_1 f/r_2 + 1}\right)^{(r_1+r_2)/2-1} e^{-y} \times \left(\frac{2}{r_1 f/r_2 + 1}\right) dy$$

$$= \frac{\Gamma[(r_1 + r_2)/2](r_1/r_2)^{r_1/2}}{\Gamma(r_1/2)\Gamma(r_2/2)} \frac{(f)^{r_1/2-1}}{(1 + r_1 f/r_2)^{(r_1+r_2)/2}}, \qquad 0 < f < \infty,$$

$$= 0 \text{ elsewhere.}$$

Accordingly, if U and V are stochastically independent chi-square variables with r_1 and r_2 degrees of freedom, respectively, then

$$F = \frac{U/r_1}{V/r_2}$$

has the immediately preceding p.d.f. $g_1(f)$. The distribution of this random variable is usually called an F distribution. It should be observed that an F distribution is completely determined by the two parameters r_1 and r_2. Table V in the Appendix gives some approximate values of

$$\Pr(F \leq f) = \int_0^f g_1(w)\, dw$$

for selected values of r_1, r_2, and f.

EXERCISES

4.35. Let T have a t distribution with 10 degrees of freedom. Find $\Pr(|T| > 2.228)$ from Table IV.

4.36. Let T have a t distribution with 14 degrees of freedom. Determine b so that $\Pr(-b < T < b) = 0.90$.

4.37. Let F have an F distribution with parameters r_1 and r_2. Prove that $1/F$ has an F distribution with parameters r_2 and r_1.

4.38. If F has an F distribution with parameters $r_1 = 5$ and $r_2 = 10$, find a and b so that $\Pr(F \leq a) = 0.05$ and $\Pr(F \leq b) = 0.95$, and, accordingly, $\Pr(a < F < b) = 0.90$. *Hint.* Write $\Pr(F \leq a) = \Pr(1/F \geq 1/a) = 1 - \Pr(1/F \leq 1/a)$, and use the result of Exercise 4.37 and Table V.

4.39. Let $T = W/\sqrt{V/r}$, where the stochastically independent variables W and V are, respectively, normal with mean zero and variance one and chi-square with r degrees of freedom. Show that T^2 has an F distribution with parameters $r_1 = 1$ and $r_2 = r$. *Hint.* What is the distribution of the numerator of T^2?

4.40. Show that

$$Y = \frac{1}{1 + (r_1/r_2)F},$$

where F has an F distribution with parameters r_1 and r_2, has a beta distribution (Exercise 4.25).

4.41. Let X_1, X_2 be a random sample from a distribution having the p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Show that $Z = X_1/X_2$ has an F distribution.

4.5 Extensions of the Change of Variable Technique

In Section 4.3 it was seen that the determination of the joint p.d.f. of two functions of two random variables of the continuous type was essentially a corollary to a theorem in analysis having to do with the change of variables in a twofold integral. This theorem has a natural extension to n-fold integrals. This extension is as follows: Consider an integral of the form

$$\int \cdots \int_A \varphi(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n$$

taken over a subset A of an n-dimensional space $\mathscr{A}$. Let

$$y_1 = u_1(x_1, x_2, \ldots, x_n), \qquad y_2 = u_2(x_1, x_2, \ldots, x_n), \ldots,$$
$$y_n = u_n(x_1, \ldots, x_n),$$

together with the inverse functions

$$x_1 = w_1(y_1, y_2, \ldots, y_n), \qquad x_2 = w_2(y_1, y_2, \ldots, y_n), \ldots,$$
$$x_n = w_n(y_1, y_2, \ldots, y_n)$$

define a one-to-one transformation which maps $\mathscr{A}$ onto $\mathscr{B}$ in the $y_1, y_2, \ldots, y_n$ space (and hence maps the subset A of $\mathscr{A}$ onto a subset B of $\mathscr{B}$). Let the first partial derivatives of the inverse functions be continuous and let the n by n determinant (called the Jacobian)

$$J = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} & \cdots & \dfrac{\partial x_1}{\partial y_n} \\ \dfrac{\partial x_2}{\partial y_1} & \dfrac{\partial x_2}{\partial y_2} & \cdots & \dfrac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial x_n}{\partial y_1} & \dfrac{\partial x_n}{\partial y_2} & \cdots & \dfrac{\partial x_n}{\partial y_n} \end{vmatrix}$$

not vanish identically in $\mathscr{B}$. Then

$$\int\cdots\int_{A}\varphi(x_1, x_2, \ldots, x_n)\,dx_1\,dx_2\cdots dx_n$$
$$= \int\cdots\int_{B}\varphi[w_1(y_1, \ldots, y_n), w_2(y_1, \ldots, y_n), \cdots, w_n(y_1, \ldots, y_n)] \times |J|\,dy_1\,dy_2\cdots dy_n.$$

Whenever the conditions of this theorem are satisfied, we can determine the joint p.d.f. of n functions of n random variables. Appropriate changes of notation in Section 4.3 (to indicate n-space as opposed to two-space) is all that is needed to show that the joint p.d.f. of the random variables $Y_1 = u_1(X_1, X_2, \ldots, X_n)$, $Y_2 = u_2(X_1, X_2, \ldots, X_n)$, $\ldots$, $Y_n = u_n(X_1, X_2, \ldots, X_n)$—where the joint p.d.f. of $X_1, X_2, \ldots, X_n$ is $\varphi(x_1, \ldots, x_n)$—is given by

$$g(y_1, y_2, \ldots, y_n) = |J|\varphi[w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n)],$$

when $(y_1, y_2, \ldots, y_n) \in \mathscr{B}$, and is zero elsewhere.

Example 1. Let $X_1, X_2, \ldots, X_{k+1}$ be mutually stochastically independent random variables, each having a gamma distribution with $\beta = 1$. The joint p.d.f. of these variables may be written as

$$\varphi(x_1, x_2, \ldots, x_{k+1}) = \prod_{i=1}^{k+1}\frac{1}{\Gamma(\alpha_i)}x_i^{\alpha_i-1}e^{-x_i}, \qquad 0 < x_i < \infty,$$
$$= 0 \text{ elsewhere.}$$

Let

$$Y_i = \frac{X_i}{X_1 + X_2 + \cdots + X_{k+1}}, \qquad i = 1, 2, \ldots, k,$$

and $Y_{k+1} = X_1 + X_2 + \cdots + X_{k+1}$ denote $k + 1$ new random variables. The associated transformation maps $\mathscr{A} = \{(x_1, \ldots, x_{k+1});\ 0 < x_i < \infty,\ i = 1, \ldots, k + 1\}$ onto the space

$$\mathscr{B} = \{(y_1, \ldots, y_k, y_{k+1});\ 0 < y_i,\ i = 1, \ldots, k,\ y_1 + \cdots + y_k < 1,\ 0 < y_{k+1} < \infty\}.$$

The single-valued inverse functions are $x_1 = y_1y_{k+1}, \ldots, x_k = y_ky_{k+1}$, $x_{k+1} = y_{k+1}(1 - y_1 - \cdots - y_k)$, so that the Jacobian is

$$J = \begin{vmatrix} y_{k+1} & 0 & \cdots & 0 & y_1 \\ 0 & y_{k+1} & \cdots & 0 & y_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & y_{k+1} & y_k \\ -y_{k+1} & -y_{k+1} & \cdots & -y_{k+1} & (1 - y_1 - \cdots - y_k) \end{vmatrix} = y_{k+1}^k.$$

Hence the joint p.d.f. of $Y_1, \ldots, Y_k, Y_{k+1}$ is given by

$$\frac{y_{k+1}^{\alpha_1+\cdots+\alpha_{k+1}-1}y_1^{\alpha_1-1}\cdots y_k^{\alpha_k-1}(1 - y_1 - \cdots - y_k)^{\alpha_{k+1}-1}e^{-y_{k+1}}}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)\Gamma(\alpha_{k+1})}$$

provided $(y_1, \ldots, y_k, y_{k+1}) \in \mathscr{B}$ and is equal to zero elsewhere. The joint p.d.f. of $Y_1, \ldots, Y_k$ is seen by inspection to be given by

$$g(y_1, \ldots, y_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_{k+1})}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{k+1})} y_1^{\alpha_1 - 1} \cdots y_k^{\alpha_k - 1} (1 - y_1 - \cdots - y_k)^{\alpha_{k+1} - 1},$$

when $0 < y_i$, $i = 1, \ldots, k$, $y_1 + \cdots + y_k < 1$, while the function g is equal to zero elsewhere. Random variables $Y_1, \ldots, Y_k$ that have a joint p.d.f. of this form are said to have a *Dirichlet distribution* with parameters $\alpha_1, \ldots, \alpha_k, \alpha_{k+1}$, and any such $g(y_1, \ldots, y_k)$ is called a *Dirichlet p.d.f.* It is seen, in the special case of $k = 1$, that the Dirichlet p.d.f. becomes a beta p.d.f.

We now consider some other problems which are encountered when transforming variables. Let X have the Cauchy p.d.f.

$$f(x) = \frac{1}{\pi(1 + x^2)}, \qquad -\infty < x < \infty,$$

and let $Y = X^2$. We seek the p.d.f. $g(y)$ of Y. Consider the transformation $y = x^2$. This transformation maps the space of X, $\mathscr{A} = \{x; -\infty < x < \infty\}$, onto $\mathscr{B} = \{y; 0 \leq y < \infty\}$. However, the transformation is not one-to-one. To each $y \in \mathscr{B}$, with the exception of $y = 0$, there correspond two points $x \in \mathscr{A}$. For example, if $y = 4$, we may have either $x = 2$ or $x = -2$. In such an instance, we represent $\mathscr{A}$ as the union of two disjoint sets A_1 and A_2 such that $y = x^2$ defines a one-to-one transformation which maps each of A_1 and A_2 onto $\mathscr{B}$. If we take A_1 to be $\{x; -\infty < x < 0\}$ and A_2 to be $\{x; 0 \leq x < \infty\}$, we see that A_1 is mapped onto $\{y; 0 < y < \infty\}$, whereas A_2 is mapped onto $\{y; 0 \leq y < \infty\}$, and these sets are not the same. Our difficulty is caused by the fact that $x = 0$ is an element of $\mathscr{A}$. Why then do we not return to the Cauchy p.d.f. and take $f(0) = 0$? Then our new $\mathscr{A}$ is $\mathscr{A} = \{-\infty < x < \infty$ but $x \neq 0\}$. We then take $A_1 = \{x; -\infty < x < 0\}$ and $A_2 = \{x; 0 < x < \infty\}$. Thus $y = x^2$, with the inverse $x = -\sqrt{y}$, maps A_1 onto $\mathscr{B} = \{y; 0 < y < \infty\}$ and the transformation is one-to-one. Moreover, the transformation $y = x^2$, with inverse $x = \sqrt{y}$, maps A_2 onto $\mathscr{B} = \{y; 0 < y < \infty\}$ and the transformation is one-to-one. Consider the probability $\Pr[Y \in B]$, where $B \subset \mathscr{B}$. Let $A_3 = \{x; x = -\sqrt{y}, y \in B\} \subset A_1$ and let $A_4 = \{x; x = \sqrt{y}, y \in B\} \subset A_2$. Then $Y \in B$ when and only when $X \in A_3$ or $X \in A_4$. Thus we have

$$\begin{aligned} \Pr(Y \in B) &= \Pr(X \in A_3) + \Pr(X \in A_4) \\ &= \int_{A_3} f(x)\, dx + \int_{A_4} f(x)\, dx. \end{aligned}$$

In the first of these integrals, let $x = -\sqrt{y}$. Thus the Jacobian, say J_1, is $-1/2\sqrt{y}$; moreover, the set A_3 is mapped onto B. In the second

integral let $x = \sqrt{y}$. Thus the Jacobian, say J_2, is $1/2\sqrt{y}$; moreover, the set A_4 is also mapped onto B. Finally,

$$\Pr(Y \in B) = \int_B f(-\sqrt{y})|-1/2\sqrt{y}|\,dy + \int_B f(\sqrt{y})(1/2\sqrt{y})\,dy$$
$$= \int_B [f(-\sqrt{y}) + f(\sqrt{y})](1/2\sqrt{y})\,dy.$$

Hence the p.d.f. of Y is given by

$$g(y) = (1/2\sqrt{y})[f(-\sqrt{y}) + f(\sqrt{y})], \qquad y \in B.$$

With $f(x)$ the Cauchy p.d.f. we have

$$g(y) = \frac{1}{\pi(1+y)\sqrt{y}}, \qquad 0 < y < \infty,$$
$$= 0 \text{ elsewhere.}$$

In the preceding discussion of a random variable of the continuous type, we had two inverse functions $x = -\sqrt{y}$ and $x = \sqrt{y}$. That is why we sought to partition $\mathscr{A}$ (or a modification of $\mathscr{A}$) into two disjoint subsets such that the transformation $y = x^2$ maps each onto the same $\mathscr{B}$. Had there been three inverse functions, we would have sought to partition $\mathscr{A}$ (or a modified form of $\mathscr{A}$) into three disjoint subsets, and so on. It is hoped that this detailed discussion will make the following paragraph easier to read.

Let $\varphi(x_1, x_2, \ldots, x_n)$ be the joint p.d.f. of $X_1, X_2, \ldots, X_n$ which are random variables of the continuous type. Let $\mathscr{A}$ be the n-dimensional space where $\varphi(x_1, x_2, \ldots, x_n) > 0$, and consider the transformation $y_1 = u_1(x_1, x_2, \ldots, x_n)$, $y_2 = u_2(x_1, x_2, \ldots, x_n), \ldots, y_n = u_n(x_1, x_2, \ldots, x_n)$ which maps $\mathscr{A}$ onto $\mathscr{B}$ in the $y_1, y_2, \ldots, y_n$ space. To each point of $\mathscr{A}$ there will correspond, of course, but one point in $\mathscr{B}$; but to a point in $\mathscr{B}$ there may correspond more than one point in $\mathscr{A}$. That is, the transformation may not be one-to-one. Suppose, however, that we can represent $\mathscr{A}$ as the union of a finite number, say k, of mutually disjoint sets $A_1, A_2, \ldots, A_k$ so that

$$y_1 = u_1(x_1, x_2, \ldots, x_n), \ldots, \qquad y_n = u_n(x_1, x_2, \ldots, x_n)$$

define a one-to-one transformation of each A_i onto $\mathscr{B}$. Thus, to each point in $\mathscr{B}$, there will correspond exactly one point in each of $A_1, A_2, \ldots, A_k$. Let

$$\begin{aligned} x_1 &= w_{1i}(y_1, y_2, \ldots, y_n), \\ x_2 &= w_{2i}(y_1, y_2, \ldots, y_n), \\ &\vdots \\ x_n &= w_{ni}(y_1, y_2, \ldots, y_n), \end{aligned} \qquad i = 1, 2, \ldots, k,$$

denote the k groups of n inverse functions, one group for each of these k transformations. Let the first partial derivatives be continuous and let each

$$J_i = \begin{vmatrix} \frac{\partial w_{1i}}{\partial y_1} & \frac{\partial w_{1i}}{\partial y_2} & \cdots & \frac{\partial w_{1i}}{\partial y_n} \\ \frac{\partial w_{2i}}{\partial y_1} & \frac{\partial w_{2i}}{\partial y_2} & \cdots & \frac{\partial w_{2i}}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial w_{ni}}{\partial y_1} & \frac{\partial w_{ni}}{\partial y_2} & \cdots & \frac{\partial w_{ni}}{\partial y_n} \end{vmatrix}, \qquad i = 1, 2, \ldots, k,$$

be not identically equal to zero in $\mathscr{B}$. From a consideration of the probability of the union of k mutually exclusive events and by applying the change of variable technique to the probability of each of these events, it can be seen that the joint p.d.f. of $Y_1 = u_1(X_1, X_2, \ldots, X_n)$, $Y_2 = u_2(X_1, X_2, \ldots, X_n), \ldots, Y_n = u_n(X_1, X_2, \ldots, X_n)$, is given by

$$g(y_1, y_2, \ldots, y_n) = \sum_{i=1}^{k} |J_i| \varphi[w_{1i}(y_1, \ldots, y_n), \ldots, w_{ni}(y_1, \ldots, y_n)],$$

provided $(y_1, y_2, \ldots, y_n) \in \mathscr{B}$, and equals zero elsewhere. The p.d.f. of any Y_i, say Y_1, is then

$$g_1(y_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1, y_2, \ldots, y_n)\, dy_2 \cdots dy_n.$$

An illustrative example follows.

Example 2. To illustrate the result just obtained, take $n = 2$ and let X_1, X_2 denote a random sample of size 2 from a distribution that is $n(0, 1)$. The joint p.d.f. of X_1 and X_2 is

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{x_1^2 + x_2^2}{2}\right), \qquad -\infty < x_1 < \infty, -\infty < x_2 < \infty.$$

Let Y_1 denote the mean and let Y_2 denote twice the variance of the random sample. The associated transformation is

$$y_1 = \frac{x_1 + x_2}{2},$$

$$y_2 = \frac{(x_1 - x_2)^2}{2}.$$

This transformation maps $\mathscr{A} = \{(x_1, x_2); -\infty < x_1 < \infty, -\infty < x_2 < \infty\}$ onto $\mathscr{B} = \{(y_1, y_2); -\infty < y_1 < \infty, 0 \leq y_2 < \infty\}$. But the transformation is not one-to-one because, to each point in $\mathscr{B}$, exclusive of points where

$y_2 = 0$, there correspond two points in $\mathscr{A}$. In fact, the two groups of inverse functions are

$$x_1 = y_1 - \sqrt{\frac{y_2}{2}}, \qquad x_1 = y_1 + \sqrt{\frac{y_2}{2}},$$

and

$$x_2 = y_1 + \sqrt{\frac{y_2}{2}}, \qquad x_2 = y_1 - \sqrt{\frac{y_2}{2}}.$$

Moreover, the set $\mathscr{A}$ cannot be represented as the union of two disjoint sets, each of which under our transformation maps onto $\mathscr{B}$. Our difficulty is caused by those points of $\mathscr{A}$ that lie on the line whose equation is $x_2 = x_1$. At each of these points we have $y_2 = 0$. However, we can define $f(x_1, x_2)$ to be zero at each point where $x_1 = x_2$. We can do this without altering the distribution of probability because the probability measure of this set is zero. Thus we have a new $\mathscr{A} = \{(x_1, x_2); -\infty < x_1 < \infty, -\infty < x_2 < \infty, \text{ but } x_1 \neq x_2\}$. This space is the union of the two disjoint sets $A_1 = \{(x_1, x_2); x_2 > x_1\}$ and $A_2 = \{(x_1, x_2); x_2 < x_1\}$. Moreover, our transformation now defines a one-to-one transformation of each A_i, $i = 1, 2$, onto the new $\mathscr{B} = \{(y_1, y_2); -\infty < y_1 < \infty, 0 < y_2 < \infty\}$. We can now find the joint p.d.f., say $g(y_1, y_2)$, of the mean Y_1 and twice the variance Y_2 of our random sample. An easy computation shows that $|J_1| = |J_2| = 1/\sqrt{2y_2}$. Thus

$$\begin{aligned} g(y_1, y_2) &= \frac{1}{2\pi} \exp\left[-\frac{(y_1 - \sqrt{y_2/2})^2}{2} - \frac{(y_1 + \sqrt{y_2/2})^2}{2}\right] \frac{1}{\sqrt{2y_2}} \\ &\quad + \frac{1}{2\pi} \exp\left[-\frac{(y_1 + \sqrt{y_2/2})^2}{2} - \frac{(y_1 - \sqrt{y_2/2})^2}{2}\right] \frac{1}{\sqrt{2y_2}} \\ &= \sqrt{\frac{1}{2\pi}} e^{-y_1^2} \frac{1}{\sqrt{2}\Gamma(\frac{1}{2})} y_2^{\frac{1}{2}-1} e^{-y_2/2}, \\ &\qquad -\infty < y_1 < \infty, 0 < y_2 < \infty. \end{aligned}$$

We can make three interesting observations. The mean Y_1 of our random sample is $n(0, \frac{1}{2})$; Y_2, which is twice the variance of our sample, is $\chi^2(1)$; and the two are stochastically independent. Thus the mean and the variance of our sample are stochastically independent.

EXERCISES

4.42. Let X_1, X_2, X_3 denote a random sample from a normal distribution $n(0, 1)$. Let the random variables Y_1, Y_2, Y_3 be defined by

$$X_1 = Y_1 \cos Y_2 \sin Y_3, \quad X_2 = Y_1 \sin Y_2 \sin Y_3, \quad \text{and} \quad X_3 = Y_1 \cos Y_3,$$

where $0 \leq Y_1 < \infty$, $0 \leq Y_2 < 2\pi$, $0 \leq Y_3 \leq \pi$. Show that Y_1, Y_2, Y_3 are mutually stochastically independent.

4.43. Let X_1, X_2, X_3 denote a random sample from the distribution having p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Show that

$$Y_1 = \frac{X_1}{X_1 + X_2}, \qquad Y_2 = \frac{X_1 + X_2}{X_1 + X_2 + X_3}, \qquad Y_3 = X_1 + X_2 + X_3$$

are mutually stochastically independent.

4.44. Let $X_1, X_2, \ldots, X_r$ be r mutually stochastically independent gamma variables with parameters $\alpha = \alpha_i$ and $\beta = 1$, $i = 1, 2, \ldots, r$, respectively. Show that $Y_1 = X_1 + X_2 + \cdots + X_r$ has a gamma distribution with parameters $\alpha = \alpha_1 + \cdots + \alpha_r$ and $\beta = 1$. *Hint.* Let $Y_2 = X_2 + \cdots + X_r$, $Y_3 = X_3 + \cdots + X_r, \ldots, Y_r = X_r$.

4.45. Let $Y_1, \ldots, Y_k$ have a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_k, \alpha_{k+1}$. (a) Show that Y_1 has a beta distribution with parameters $\alpha = \alpha_1$ and $\beta = \alpha_2 + \cdots + \alpha_{k+1}$. (b) Show that $Y_1 + \cdots + Y_r$, $r \leq k$, has a beta distribution with parameters $\alpha = \alpha_1 + \cdots + \alpha_r$ and $\beta = \alpha_{r+1} + \cdots + \alpha_{k+1}$. (c) Show that $Y_1 + Y_2$, $Y_3 + Y_4$, $Y_5, \ldots, Y_k$, $k \geq 5$, have a Dirichlet distribution with parameters $\alpha_1 + \alpha_2$, $\alpha_3 + \alpha_4$, $\alpha_5, \ldots, \alpha_k$, α_{k+1}. *Hint.* Recall the definition of Y_i in Example 1 and use the fact that the sum of several stochastically independent gamma variables with $\beta = 1$ is a gamma variable (Exercise 4.44).

4.46. Let X_1, X_2, and X_3 be three mutually stochastically independent chi-square variables with r_1, r_2, and r_3 degrees of freedom, respectively.

(a) Show that $Y_1 = X_1/X_2$ and $Y_2 = X_1 + X_2$ are stochastically independent and that Y_2 is $\chi^2(r_1 + r_2)$.

(b) Deduce that

$$\frac{X_1/r_1}{X_2/r_2} \quad \text{and} \quad \frac{X_3/r_3}{(X_1 + X_2)/(r_1 + r_2)}$$

are stochastically independent F variables.

4.47. If $f(x) = \frac{1}{2}$, $-1 < x < 1$, zero elsewhere, is the p.d.f. of the random variable X, find the p.d.f. of $Y = X^2$.

4.48. If X_1, X_2 is a random sample from a distribution which is $n(0, 1)$, find the joint p.d.f. of $Y_1 = X_1^2 + X_2^2$ and $Y_2 = X_2$ and the marginal p.d.f. of Y_1.

4.49. If X has the p.d.f. $f(x) = \frac{1}{4}$, $-1 < x < 3$, zero elsewhere, find the p.d.f. of $Y = X^2$. *Hint.* Here $\mathscr{B} = \{y; 0 \leq y < 9\}$ and the event $Y \in B$ is the union of two mutually exclusive events if $B = \{y; 0 < y < 1\}$.

4.6 Distributions of Order Statistics

In this section the notion of an order statistic will be defined and we shall investigate some of the simpler properties of such a statistic.

These statistics have in recent times come to play an important role in statistical inference partly because some of their properties do not depend upon the distribution from which the random sample is obtained.

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution of the *continuous type* having p.d.f. $f(x)$ which is positive, provided $a < x < b$. Let Y_1 be the smallest of these X_i, Y_2 the next X_i in order of magnitude, ..., and Y_n the largest X_i. That is, $Y_1 < Y_2 < \cdots < Y_n$ represent $X_1, X_2, \ldots, X_n$ when the latter are arranged in ascending order of magnitude. Then Y_i, $i = 1, 2, \ldots, n$, is called the ith order statistic of the random sample $X_1, X_2, \ldots, X_n$. It will be shown that the joint p.d.f. of $Y_1, Y_2, \ldots, Y_n$ is given by

$$(1) \qquad g(y_1, y_2, \ldots, y_n) = (n!)f(y_1)f(y_2)\cdots f(y_n), \quad a < y_1 < y_2 < \cdots < y_n < b,$$
$$= 0 \text{ elsewhere.}$$

We shall prove this only for the case $n = 3$, but the argument is seen to be entirely general. With $n = 3$, the joint p.d.f. of X_1, X_2, X_3 is $f(x_1)f(x_2)f(x_3)$. Consider a probability such as $\Pr(a < X_1 = X_2 < b, a < X_3 < b)$. This probability is given by

$$\int_a^b \int_a^b \int_{x_2}^{x_2} f(x_1)f(x_2)f(x_3)\,dx_1\,dx_2\,dx_3 = 0,$$

since

$$\int_{x_2}^{x_2} f(x_1)\,dx_1$$

is defined in calculus to be zero. As has been pointed out, we may, without altering the distribution of X_1, X_2, X_3, define the joint p.d.f. $f(x_1)f(x_2)f(x_3)$ to be zero at all points (x_1, x_2, x_3) which have at least two of their coordinates equal. Then the set $\mathcal{A}$, where $f(x_1)f(x_2)f(x_3) > 0$, is the union of the six mutually disjoint sets:

$$A_1 = \{(x_1, x_2, x_3); a < x_1 < x_2 < x_3 < b\},$$
$$A_2 = \{(x_1, x_2, x_3); a < x_2 < x_1 < x_3 < b\},$$
$$A_3 = \{(x_1, x_2, x_3); a < x_1 < x_3 < x_2 < b\},$$
$$A_4 = \{(x_1, x_2, x_3); a < x_2 < x_3 < x_1 < b\},$$
$$A_5 = \{(x_1, x_2, x_3); a < x_3 < x_1 < x_2 < b\},$$
$$A_6 = \{(x_1, x_2, x_3); a < x_3 < x_2 < x_1 < b\}.$$

There are six of these sets because we can arrange x_1, x_2, x_3 in precisely $3! = 6$ ways. Consider the functions $y_1 =$ minimum of x_1, x_2, x_3; $y_2 =$ middle in magnitude of x_1, x_2, x_3; and $y_3 =$ maximum of x_1, x_2, x_3. These functions define one-to-one transformations which map each of

$A_1, A_2, \ldots, A_6$ onto the same set $\mathcal{B} = \{(y_1, y_2, y_3);\ a < y_1 < y_2 < y_3 < b\}$. The inverse functions are, for points in A_1, $x_1 = y_1$, $x_2 = y_2$, $x_3 = y_3$; for points in A_2, they are $x_1 = y_2$, $x_2 = y_1$, $x_3 = y_3$; and so on, for each of the remaining four sets. Then we have that

$$J_1 = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1$$

and

$$J_2 = \begin{vmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{vmatrix} = -1.$$

It is easily verified that the absolute value of each of the $3! = 6$ Jacobians is $+1$. Thus the joint p.d.f. of the three order statistics Y_1 = minimum of X_1, X_2, X_3; Y_2 = middle in magnitude of X_1, X_2, X_3; Y_3 = maximum of X_1, X_2, X_3 is

$$\begin{aligned} g(y_1, y_2, y_3) &= |J_1| f(y_1)f(y_2)f(y_3) + |J_2| f(y_2)f(y_1)f(y_3) + \cdots \\ &\quad + |J_6| f(y_3)f(y_2)f(y_1), \qquad a < y_1 < y_2 < y_3 < b, \\ &= (3!) f(y_1)f(y_2)f(y_3), \qquad a < y_1 < y_2 < y_3 < b, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

This is Equation (1) with $n = 3$.

In accordance with the natural extension of Theorem 1, Section 2.4 (p. 77) to distributions of more than two random variables, it is seen that the order statistics, unlike the items of the random sample, are stochastically dependent.

Example 1. Let X denote a random variable of the continuous type with p.d.f. $f(x)$ which is positive and continuous, provided $a < x < b$, and is zero elsewhere. The distribution function $F(x)$ of X may be written

$$F(x) = \int_a^x f(w)\, dw, \qquad a < x < b.$$

If $x \leq a$, $F(x) = 0$; and if $b \leq x$, $F(x) = 1$. Thus there is a unique median m of the distribution with $F(m) = \frac{1}{2}$. Let X_1, X_2, X_3 denote a random sample from this distribution and let $Y_1 < Y_2 < Y_3$ denote the order statistics of the sample. We shall compute the probability that $Y_2 \leq m$. The joint p.d.f. of the three order statistics is

$$\begin{aligned} g(y_1, y_2, y_3) &= 6 f(y_1)f(y_2)f(y_3), \qquad a < y_1 < y_2 < y_3 < b, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

The p.d.f. of Y_2 is then

$$\begin{aligned} h(y_2) &= 6f(y_2)\int_{y_2}^{b}\int_{a}^{y_2} f(y_1)f(y_3)\,dy_1\,dy_3, \\ &= 6f(y_2)F(y_2)[1 - F(y_2)], \qquad a < y_2 < b, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Accordingly,

$$\begin{aligned} \Pr(Y_2 \le m) &= 6\int_a^m \{F(y_2)f(y_2) - [F(y_2)]^2 f(y_2)\}\,dy_2 \\ &= 6\left\{\frac{[F(y_2)]^2}{2} - \frac{[F(y_2)]^3}{3}\right\}_a^m = \frac{1}{2}. \end{aligned}$$

The procedure used in Example 1 can be used to obtain general formulas for the marginal probability density functions of the order statistics. We shall do this now. Let X denote a random variable of the continuous type having p.d.f. $f(x)$ which is positive and continuous, provided $a < x < b$, and is zero elsewhere. Then the distribution function $F(x)$ may be written

$$\begin{aligned} F(x) &= 0, \qquad x \le a, \\ &= \int_a^x f(w)\,dw, \qquad a < x < b, \\ &= 1, \qquad b \le x. \end{aligned}$$

Accordingly, $F'(x) = f(x)$, $a < x < b$. Moreover, if $a < x < b$,

$$\begin{aligned} 1 - F(x) &= F(b) - F(x) \\ &= \int_a^b f(w)\,dw - \int_a^x f(w)\,dw \\ &= \int_x^b f(w)\,dw. \end{aligned}$$

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from this distribution, and let $Y_1, Y_2, \ldots, Y_n$ denote the order statistics of this random sample. Then the joint p.d.f. of $Y_1, Y_2, \ldots, Y_n$ is

$$\begin{aligned} g(y_1, y_2, \ldots, y_n) &= n!f(y_1)f(y_2)\cdots f(y_n), \\ &\qquad a < y_1 < y_2 < \cdots < y_n < b, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

It will first be shown how the marginal p.d.f. of Y_n may be expressed in terms of the distribution function $F(x)$ and the p.d.f. $f(x)$ of the random variable X. If $a < y_n < b$, the marginal p.d.f. of Y_n is given by

$$g_n(y_n) = \int_a^{y_n} \cdots \int_a^{y_4} \int_a^{y_3} \int_a^{y_2} n!\, f(y_1)f(y_2) \cdots f(y_n)\, dy_1\, dy_2\, dy_3 \cdots dy_{n-1}$$

$$= \int_a^{y_n} \cdots \int_a^{y_4} \int_a^{y_3} n! \left(\int_a^{y_2} f(y_1)\, dy_1\right) f(y_2) \cdots f(y_n)\, dy_2 \cdots dy_{n-1}$$

$$= \int_a^{y_n} \cdots \int_a^{y_4} \int_a^{y_3} n!\, F(y_2)f(y_2) \cdots f(y_n)\, dy_2 \cdots dy_{n-1},$$

since $F(x) = \int_a^x f(w)\, dw$. Now

$$\int_a^{y_3} F(y_2)f(y_2)\, dy_2 = \frac{[F(y_2)]^2}{2}\Bigg|_a^{y_3}$$

$$= \frac{[F(y_3)]^2}{2},$$

since $F(a) = 0$. Thus

$$g_n(y_n) = \int_a^{y_n} \cdots \int_a^{y_4} n!\, \frac{[F(y_3)]^2}{2} f(y_3) \cdots f(y_n)\, dy_3 \cdots dy_{n-1}.$$

But

$$\int_a^{y_4} \frac{[F(y_3)]^2}{2} f(y_3)\, dy_3 = \frac{[F(y_3)]^3}{2 \cdot 3}\Bigg|_a^{y_4} = \frac{[F(y_4)]^3}{2 \cdot 3},$$

so

$$g_n(y_n) = \int_a^{y_n} \cdots \int_a^{y_5} n!\, \frac{[F(y_4)]^3}{3!} f(y_4) \cdots f(y_n)\, dy_4 \cdots dy_{n-1}.$$

If the successive integrations on $y_4, \ldots, y_{n-1}$ are carried out, it is seen that

$$g_n(y_n) = n!\, \frac{[F(y_n)]^{n-1}}{(n-1)!} f(y_n)$$

$$= n[F(y_n)]^{n-1} f(y_n), \qquad a < y_n < b,$$

$$= 0 \text{ elsewhere.}$$

It will next be shown how to express the marginal p.d.f. of Y_1 in terms of $F(x)$ and $f(x)$. We have, for $a < y_1 < b$,

$$g_1(y_1) = \int_{y_1}^b \cdots \int_{y_{n-3}}^b \int_{y_{n-2}}^b \int_{y_{n-1}}^b n!\, f(y_1)f(y_2) \cdots f(y_n)\, dy_n\, dy_{n-1} \cdots dy_2$$

$$= \int_{y_1}^b \cdots \int_{y_{n-3}}^b \int_{y_{n-2}}^b n!\, f(y_1)f(y_2) \cdots$$

$$f(y_{n-1})[1 - F(y_{n-1})]\, dy_{n-1} \cdots dy_2.$$

But

$$\int_{y_{n-2}}^{b} [1 - F(y_{n-1})]f(y_{n-1})\,dy_{n-1} = -\frac{[1 - F(y_{n-1})]^2}{2}\Bigg|_{y_{n-2}}^{b}$$
$$= \frac{[1 - F(y_{n-2})]^2}{2},$$

so that

$$g_1(y_1) = \int_{y_1}^{b} \cdots \int_{y_{n-3}}^{b} n!\, f(y_1) \cdots f(y_{n-2}) \frac{[1 - F(y_{n-2})]^2}{2}\, dy_{n-2} \cdots dy_2.$$

Upon completing the integrations, it is found that

$$g_1(y_1) = n[1 - F(y_1)]^{n-1} f(y_1), \qquad a < y_1 < b,$$
$$= 0 \text{ elsewhere.}$$

Once it is observed that

$$\int_a^x [F(w)]^{\alpha-1} f(w)\,dw = \frac{[F(x)]^\alpha}{\alpha}, \qquad \alpha > 0$$

and that

$$\int_y^b [1 - F(w)]^{\beta-1} f(w)\,dw = \frac{[1 - F(y)]^\beta}{\beta}, \qquad \beta > 0,$$

it is easy to express the marginal p.d.f. of any order statistic, say Y_k, in terms of $F(x)$ and $f(x)$. This is done by evaluating the integral

$$g_k(y_k) = \int_a^{y_k} \cdots \int_a^{y_2} \int_{y_k}^{b} \cdots \int_{y_{n-1}}^{b} n!\, f(y_1) f(y_2) \cdots f(y_n)\, dy_n \cdots$$
$$dy_{k+1}\, dy_1 \cdots dy_{k-1}.$$

The result is

$$(2) \quad g_k(y_k) = \frac{n!}{(k-1)!\,(n-k)!} [F(y_k)]^{k-1} [1 - F(y_k)]^{n-k} f(y_k), \qquad a < y_k < b,$$
$$= 0 \text{ elsewhere.}$$

The student may be interested in noting that in Example 5, p. 89, the p.d.f. of Y_2, when $n = 3$, was obtained in another manner.

Finally the joint p.d.f. of any two order statistics, say $Y_i < Y_j$, is as easily expressed in terms of $F(x)$ and $f(x)$. We have

$$g_{ij}(y_i, y_j) = \int_a^{y_i} \cdots \int_a^{y_2} \int_{y_i}^{y_j} \cdots \int_{y_{j-2}}^{y_j} \int_{y_j}^{b} \cdots \int_{y_{n-1}}^{b} n!\, f(y_1) \cdots$$
$$f(y_n)\, dy_n \cdots dy_{j+1}\, dy_{j-1} \cdots dy_{i+1}\, dy_1 \cdots dy_{i-1}.$$

Since, for $\gamma > 0$,

$$\int_x^y [F(y) - F(w)]^{\gamma-1} f(w)\, dw = -\frac{[F(y) - F(w)]^{\gamma}}{\gamma}\Bigg|_x^y$$
$$= \frac{[F(y) - F(x)]^{\gamma}}{\gamma},$$

it is found that

$$(3)\quad g_{ij}(y_i, y_j) = \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!} \times [F(y_i)]^{i-1}[F(y_j) - F(y_i)]^{j-i-1}[1 - F(y_j)]^{n-j} f(y_i) f(y_j)$$

for $a < y_i < y_j < b$, and zero elsewhere.

Three illustrative examples follow.

Example 2. Let $Y_1 < Y_2 < Y_3 < Y_4$ denote the order statistics of a random sample of size 4 from a distribution having p.d.f.

$$f(x) = 2x, \qquad 0 < x < 1,$$
$$= 0 \text{ elsewhere.}$$

We shall express the p.d.f. of Y_3 in terms of $f(x)$ and $F(x)$ and then compute $\Pr(\frac{1}{2} < Y_3)$. Here $F(x) = x^2$, provided $0 < x < 1$, so that

$$g_3(y_3) = \frac{4!}{2!\,1!}(y_3^2)^2(1 - y_3^2)(2y_3), \qquad 0 < y_3 < 1,$$
$$= 0 \text{ elsewhere.}$$

Thus

$$\Pr(\tfrac{1}{2} < Y_3) = \int_{1/2}^{\infty} g_3(y_3)\, dy_3$$
$$= \int_{1/2}^{1} 24(y_3^5 - y_3^7)\, dy_3 = \tfrac{243}{256}.$$

Example 3. Let $Y_1 < Y_2 < \cdots < Y_5$ denote the order statistics of a random sample of size 5 from a distribution having p.d.f.

$$f(x) = e^{-x}, \qquad 0 < x < \infty,$$
$$= 0 \text{ elsewhere.}$$

It will be shown that the statistics $Z_1 = Y_2$ and $Z_2 = Y_4 - Y_2$ are stochastically independent. Since $F(x) = 1 - e^{-x}$, $0 < x < \infty$, the joint p.d.f. of Y_2 and Y_4 is

$$g_{24}(y_2, y_4) = \frac{5!}{1!\,1!\,1!}(1 - e^{-y_2})(e^{-y_2} - e^{-y_4})(e^{-y_4})e^{-y_2-y_4},$$
$$0 < y_2 < y_4 < \infty,$$
$$= 0 \text{ elsewhere.}$$

The transformation $z_1 = y_2$, $z_2 = y_4 - y_2$ maps the set $\{(y_2, y_4);\ 0 < y_2 < y_4 < \infty\}$ onto the set $\{(z_1, z_2);\ 0 < z_1 < \infty,\ 0 < z_2 < \infty\}$, and the Jacobian of the transformation is 1. Accordingly, the joint p.d.f. of Z_1 and Z_2 is

$$h(z_1, z_2) = 120e^{-4z_1}(1 - e^{-z_1})e^{-2z_2}(1 - e^{-z_2}), \qquad 0 < z_1 < \infty, 0 < z_2 < \infty,$$
$$= 0 \text{ elsewhere.}$$

In accordance with Theorem 1, p. 77, Z_1 and Z_2 are stochastically independent.

Certain functions of the order statistics $Y_1, Y_2, \ldots, Y_n$ are important statistics themselves. A few of these are: (a) $Y_n - Y_1$, which is called the range of the random sample; (b) $(Y_1 + Y_n)/2$, which is called the midrange of the random sample; and (c) if n is odd, $Y_{(n+1)/2}$, which is called the median of the random sample.

Example 4. Let Y_1, Y_2, Y_3 be the order statistics of a random sample of size 3 from a distribution having p.d.f.

$$f(x) = 1, \qquad 0 < x < 1,$$
$$= 0 \text{ elsewhere.}$$

We seek the p.d.f. of the sample range $Z_1 = Y_3 - Y_1$. Since $F(x) = x$, $0 < x < 1$, the joint p.d.f. of Y_1 and Y_3 is

$$g_{13}(y_1, y_3) = 6(y_3 - y_1), \qquad 0 < y_1 < y_3 < 1,$$
$$= 0 \text{ elsewhere.}$$

In addition to $Z_1 = Y_3 - Y_1$, let $Z_2 = Y_3$. Consider the functions $z_1 = y_3 - y_1$, $z_2 = y_3$, and their inverses $y_1 = z_2 - z_1$, $y_3 = z_2$, so that the corresponding Jacobian of the one-to-one transformation is

$$J = \begin{vmatrix} \frac{\partial y_1}{\partial z_1} & \frac{\partial y_1}{\partial z_2} \\ \frac{\partial y_3}{\partial z_1} & \frac{\partial y_3}{\partial z_2} \end{vmatrix} = \begin{vmatrix} -1 & 1 \\ 0 & 1 \end{vmatrix} = -1.$$

Thus the joint p.d.f. of Z_1 and Z_2 is

$$h(z_1, z_2) = |-1|6z_1 = 6z_1, \qquad 0 < z_1 < z_2 < 1.$$
$$= 0 \text{ elsewhere.}$$

Accordingly, the p.d.f. of the range $Z_1 = Y_3 - Y_1$ of the random sample of size 3 is

$$h_1(z_1) = \int_{z_1}^{1} 6z_1\, dz_2 = 6z_1(1 - z_1), \qquad 0 < z_1 < 1,$$
$$= 0 \text{ elsewhere.}$$

EXERCISES

4.50. Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size 4 from the distribution having p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Find Pr $(3 \leq Y_4)$.

4.51. Let X_1, X_2, X_3 be a random sample from a distribution of the continuous type having p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. Compute the probability that the smallest of these X_i exceeds the median of the distribution.

4.52. Let $f(x) = \frac{1}{6}$, $x = 1, 2, 3, 4, 5, 6$, zero elsewhere, be the p.d.f. of a distribution of the discrete type. Show that the p.d.f. of the smallest item of a random sample of size 5 from this distribution is

$$g_1(y_1) = \left(\frac{7 - y_1}{6}\right)^5 - \left(\frac{6 - y_1}{6}\right)^5, \qquad y_1 = 1, 2, \ldots, 6,$$

zero elsewhere. Note that in this exercise the random sample is from a distribution of the discrete type. All formulas in the text were derived under the assumption that the random sample is from a distribution of the continuous type and are not applicable. Why?

4.53. Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ denote the order statistics of a random sample of size 5 from a distribution of the continuous type having distribution function $F(x)$ such that $F'(x) = f(x)$ is positive and continuous, provided $a < x < b$. Show by integrating the joint p.d.f. of $Y_1, \ldots, Y_5$ on y_1, y_3, y_4, y_5 that the p.d.f. of Y_2 is $g_2(y_2) = 20F(y_2)[1 - F(y_2)]^3 f(y_2)$, $a < y_2 < b$, and zero elsewhere.

4.54. Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ denote the order statistics of a random sample of size 5 from a distribution having p.d.f. $f(x) = 3x^2$, $0 < x < 1$, zero elsewhere. Show that $Z_1 = Y_2/Y_4$ and $Z_2 = Y_4$ are stochastically independent.

4.55. Find the probability that the range of a random sample of size 4 from the uniform distribution having the p.d.f. $f(x) = 1$, $0 < x < 1$, zero elsewhere, is less than $\frac{1}{2}$.

4.56. Let $Y_1 < Y_2 < Y_3$ be the order statistics of a random sample of size 3 from a distribution having the p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. Show that $Z_1 = Y_1/Y_2$, $Z_2 = Y_2/Y_3$, and $Z_3 = Y_3$ are mutually stochastically independent.

4.57. If a random sample of size 2 is taken from a distribution having p.d.f. $f(x) = 2(1 - x)$, $0 < x < 1$, zero elsewhere, compute the probability that one sample item is at least twice as large as the other.

4.58. Let $Y_1 < Y_2 < Y_3$ denote the order statistics of a random sample of size 3 from a distribution with p.d.f. $f(x) = 1$, $0 < x < 1$, zero elsewhere. Let $Z = (Y_1 + Y_3)/2$ be the midrange of the sample. Find the p.d.f. of Z.

4.59. Let $Y_1 < Y_2$ denote the order statistics of a random sample of size 2 from $n(0, \sigma^2)$. Show that $E(Y_1) = -\sigma/\sqrt{\pi}$.

4.60. Let $Y_1 < Y_2$ be the order statistics of a random sample of size 2 from a distribution of the continuous type which has p.d.f. $f(x)$ such that $f(x) > 0$, provided $x \geq 0$, and $f(x) = 0$ elsewhere. Show that the stochastic independence of $Z_1 = Y_1$ and $Z_2 = Y_2 - Y_1$ characterizes the gamma p.d.f. $f(x)$, which has parameters $\alpha = 1$ and $\beta > 0$. *Hint.* Use the change of variable technique to find the joint p.d.f. of Z_1 and Z_2 from that of Y_1 and Y_2. Accept the fact that the functional equation $h(0)h(x + y) \equiv h(x)h(y)$ has the solution $h(x) = c_1 e^{c_2 x}$, where c_1 and c_2 are constants.

4.61. Let Y denote the median of a random sample of size $n = 2k + 1$, k a positive integer, from a distribution which is $n(\mu, \sigma^2)$. Prove that the graph of the p.d.f. of Y is symmetric with respect to the vertical axis through $y = \mu$ and deduce that $E(Y) = \mu$.

4.62. Let X and Y denote stochastically independent random variables with respective probability density functions $f(x) = 2x$, $0 < x < 1$, zero elsewhere, and $g(y) = 3y^2$, $0 < y < 1$, zero elsewhere. Let $U = \min(X, Y)$ and $V = \max(X, Y)$. Find the joint p.d.f. of U and V.

4.63. Let the joint p.d.f. of X and Y be $f(x, y) = \frac{12}{7}x(x + y)$, $0 < x < 1$, $0 < y < 1$, zero elsewhere. Let $U = \min(X, Y)$ and $V = \max(X, Y)$. Find the joint p.d.f. of U and V.

4.7 The Moment-Generating Function Technique

The change of variable procedure has been seen, in certain cases, to be an effective method of finding the distribution of a function of several random variables. An alternative procedure, built around the concept of the moment-generating function of a distribution, will be presented in this section. This procedure is particularly effective in certain instances. We should recall that a moment-generating function, when it exists, is unique and that it uniquely determines the distribution of a probability.

Let $\varphi(x_1, x_2, \ldots, x_n)$ denote the joint p.d.f. of the n random variables $X_1, X_2, \ldots, X_n$. These random variables may or may not be the items of a random sample from some distribution that has a given p.d.f. $f(x)$. Let $Y_1 = u_1(X_1, X_2, \ldots, X_n)$. We seek $g(y_1)$, the p.d.f. of the random variable Y_1. Consider the moment-generating function of Y_1. If it exists, it is given by

$$M(t) = E(e^{tY_1}) = \int_{-\infty}^{\infty} e^{ty_1} g(y_1)\, dy_1$$

in the continuous case. It would seem that we need to know $g(y_1)$ before we can compute $M(t)$. That this is not the case is a fundamental fact. To see this, consider

$$(1) \qquad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\,[tu_1(x_1, \ldots, x_n)]\varphi(x_1, \ldots, x_n)\, dx_1 \cdots dx_n,$$

which we assume to exist for $-h < t < h$. We shall introduce n new variables of integration. They are $y_1 = u_1(x_1, x_2, \ldots, x_n), \ldots, y_n = u_n(x_1, x_2, \ldots, x_n)$. Momentarily we assume that these functions define a one-to-one transformation. Let $x_i = w_i(y_1, y_2, \ldots, y_n), i = 1, 2, \ldots, n$, denote the inverse functions and let J denote the Jacobian. Under this transformation Display (1) becomes

$$(2) \qquad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{ty_1}|J|\varphi(w_1, \ldots, w_n)\, dy_2 \cdots dy_n\, dy_1.$$

In accordance with Section 4.5,

$$|J|\varphi[w_1(y_1, y_2, \ldots, y_n), \ldots, w_n(y_1, y_2, \ldots, y_n)]$$

is the joint p.d.f. of $Y_1, Y_2, \ldots, Y_n$. The marginal p.d.f. $g(y_1)$ of Y_1 is obtained by integrating this joint p.d.f. on $y_2, \ldots, y_n$. Since the factor e^{ty_1} does not involve the variables $y_2, \ldots, y_n$, Display (2) may be written as

$$(3) \qquad \int_{-\infty}^{\infty} e^{ty_1} g(y_1)\, dy_1.$$

But this is by definition the moment-generating function $M(t)$ of the distribution of Y_1. That is to say, we can compute $E[\exp\,(tu_1(X_1, \ldots, X_n)]$ and have the value of $E(e^{tY_1})$, where $Y_1 = u_1(X_1, \ldots, X_n)$. This fact provides another technique to help us find the p.d.f. of a function of several random variables. For if the moment-generating function of Y_1 is seen to be that of a certain kind of distribution, the uniqueness property makes it certain that Y_1 has that kind of distribution. When the p.d.f. of Y_1 is obtained in this manner, we say that we use the *moment-generating function technique.*

The student will observe that we have assumed the transformation to be one-to-one. We did this for simplicity of presentation. If the transformation is not one-to-one, let

$$x_j = w_{ji}(y_1, \ldots, y_n), \qquad j = 1, 2, \ldots, n, \quad i = 1, 2, \ldots, k,$$

denote the k groups of n inverse functions each. Let $J_i, i = 1, 2, \ldots, k$, denote the k Jacobians. Then

$$(4) \qquad \sum_{i=1}^{k} |J_i|\varphi[w_{1i}(y_1, \ldots, y_n), \ldots, w_{ni}(y_1, \ldots, y_n)]$$

is the joint p.d.f. of $Y_1, \ldots, Y_n$. Then Display (1) becomes Display (2) with $|J|\varphi(w_1, \ldots, w_n)$ replaced by Display (4). Hence our result is valid if the transformation is not one-to-one. It seems evident that we can treat the discrete case in an analogous manner with the same result.

It should be noted that the expectation, subject to its existence, of any function of Y_1 can be computed in like manner. That is, if $w(y_1)$ is a function of y_1, then

$$E[w(Y_1)] = \int_{-\infty}^{\infty} w(y_1)g(y_1)\,dy_1$$
$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} w[u_1(x_1, \ldots, x_n)]\varphi(x_1, \ldots, x_n)\,dx_1 \cdots dx_n.$$

We shall now give some examples and prove some theorems where we use the moment-generating function technique. In the first example, to emphasize the nature of the problem, we find the distribution of a rather simple statistic both by a direct probabilistic argument and by the moment-generating function technique.

Example 1. Let the stochastically independent random variables X_1 and X_2 have the same p.d.f.

$$f(x) = \frac{x}{6}, \qquad x = 1, 2, 3,$$
$$= 0 \text{ elsewhere;}$$

that is, the p.d.f. of X_1 is $f(x_1)$ and that of X_2 is $f(x_2)$; and so the joint p.d.f. of X_1 and X_2 is

$$f(x_1)f(x_2) = \frac{x_1x_2}{36}, \qquad x_1 = 1, 2, 3, x_2 = 1, 2, 3,$$
$$= 0 \text{ elsewhere.}$$

A probability, such as $\Pr(X_1 = 2, X_2 = 3)$, can be seen immediately to be $(2)(3)/36 = \frac{1}{6}$. However, consider a probability such as $\Pr(X_1 + X_2 = 3)$. The computation can be made by first observing that the event $X_1 + X_2 = 3$ is the union, exclusive of the events with probability zero, of the two mutually exclusive events $(X_1 = 1, X_2 = 2)$ and $(X_1 = 2, X_2 = 1)$. Thus,

$$\Pr(X_1 + X_2 = 3) = \Pr(X_1 = 1, X_2 = 2) + \Pr(X_1 = 2, X_2 = 1)$$
$$= \frac{(1)(2)}{36} + \frac{(2)(1)}{36} = \frac{4}{36}.$$

More generally, let y represent any of the numbers 2, 3, 4, 5, 6. The probability of each of the events $X_1 + X_2 = y$, $y = 2, 3, 4, 5, 6$, can be computed as in the case $y = 3$. Let $g(y) = \Pr(X_1 + X_2 = y)$. Then the table

y	2	3	4	5	6
$g(y)$	$\frac{1}{36}$	$\frac{4}{36}$	$\frac{10}{36}$	$\frac{12}{36}$	$\frac{9}{36}$

gives the values of $g(y)$ for $y = 2, 3, 4, 5, 6$. For all other values of y, $g(y) = 0$. What we have actually done is to define a new random variable Y by $Y = X_1 + X_2$, and we have found the p.d.f. $g(y)$ of this random variable Y. We shall now solve the same problem, but by the moment-generating function technique.

Now the moment-generating function of Y is

$$\begin{aligned} M(t) &= E(e^{t(X_1+X_2)}) \\ &= E(e^{tX_1}e^{tX_2}) \\ &= E(e^{tX_1})E(e^{tX_2}) \end{aligned}$$

since X_1 and X_2 are stochastically independent. In this example X_1 and X_2 have the same distribution, so they have the same moment-generating function; that is,

$$E(e^{tX_1}) = E(e^{tX_2}) = \tfrac{1}{6}e^{t} + \tfrac{2}{6}e^{2t} + \tfrac{3}{6}e^{3t}.$$

Thus,

$$\begin{aligned} M(t) &= (\tfrac{1}{6}e^{t} + \tfrac{2}{6}e^{2t} + \tfrac{3}{6}e^{3t})^2 \\ &= \tfrac{1}{36}e^{2t} + \tfrac{4}{36}e^{3t} + \tfrac{10}{36}e^{4t} + \tfrac{12}{36}e^{5t} + \tfrac{9}{36}e^{6t}. \end{aligned}$$

This form of $M(t)$ tells us immediately that the p.d.f. $g(y)$ of Y is zero except at $y = 2, 3, 4, 5, 6$, and that $g(y)$ assumes the values $\frac{1}{36}, \frac{4}{36}, \frac{10}{36}, \frac{12}{36}, \frac{9}{36}$, respectively, at these points where $g(y) > 0$. This is, of course, the same result that was obtained in the first solution. There appears here to be little, if any, preference for one solution over the other. But in more complicated situations, and particularly with random variables of the continuous type, the moment-generating function technique can prove very powerful.

Example 2. Let X_1 and X_2 be stochastically independent with normal distributions $n(\mu_1, \sigma_1^2)$ and $n(\mu_2, \sigma_2^2)$, respectively. Define the random variable Y by $Y = X_1 - X_2$. The problem is to find $g(y)$, the p.d.f. of Y. This will be done by first finding the moment-generating function of Y. It is

$$\begin{aligned} M(t) &= E(e^{t(X_1-X_2)}) \\ &= E(e^{tX_1}e^{-tX_2}) \\ &= E(e^{tX_1})E(e^{-tX_2}), \end{aligned}$$

since X_1 and X_2 are stochastically independent. It is known that

$$E(e^{tX_1}) = \exp\left(\mu_1 t + \frac{\sigma_1^2 t^2}{2}\right)$$

and that

$$E(e^{tX_2}) = \exp\left(\mu_2 t + \frac{\sigma_2^2 t^2}{2}\right)$$

for all real t. Then $E(e^{-tX_2})$ can be obtained from $E(e^{tX_2})$ by replacing t by $-t$. That is,

$$E(e^{-tX_2}) = \exp\left(-\mu_2 t + \frac{\sigma_2^2 t^2}{2}\right).$$

Finally then,

$$\begin{aligned} M(t) &= \exp\left(\mu_1 t + \frac{\sigma_1^2 t^2}{2}\right) \exp\left(-\mu_2 t + \frac{\sigma_2^2 t^2}{2}\right) \\ &= \exp\left((\mu_1 - \mu_2)t + \frac{(\sigma_1^2 + \sigma_2^2)t^2}{2}\right). \end{aligned}$$

The distribution of Y is completely determined by its moment-generating function $M(t)$, and it is seen that Y has the p.d.f. $g(y)$, which is $n(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$. That is, the difference between two stochastically independent, normally distributed, random variables is itself a random variable which is normally distributed with mean equal to the difference of the means (in the order indicated) and the variance equal to the sum of the variances.

The following theorem, which is a generalization of Example 2, is very important in distribution theory.

Theorem 1. *Let $X_1, X_2, \ldots, X_n$ be mutually stochastically independent random variables having, respectively, the normal distributions $n(\mu_1, \sigma_1^2), n(\mu_2, \sigma_2^2), \ldots,$ and $n(\mu_n, \sigma_n^2)$. The random variable $Y = k_1X_1 + k_2X_2 + \cdots + k_nX_n$, where $k_1, k_2, \ldots, k_n$ are real constants, is normally distributed with mean $k_1\mu_1 + \cdots + k_n\mu_n$ and variance $k_1^2\sigma_1^2 + \cdots + k_n^2\sigma_n^2$. That is, Y is $n\left(\sum_1^n k_i\mu_i, \sum_1^n k_i^2\sigma_i^2\right)$.*

Proof. Because $X_1, X_2, \ldots, X_n$ are mutually stochastically independent, the moment-generating function of Y is given by

$$\begin{aligned} M(t) &= E\{\exp[t(k_1X_1 + k_2X_2 + \cdots + k_nX_n)]\} \\ &= E(e^{tk_1X_1})E(e^{tk_2X_2}) \cdots E(e^{tk_nX_n}). \end{aligned}$$

Now

$$E(e^{tX_i}) = \exp\left(\mu_i t + \frac{\sigma_i^2 t^2}{2}\right),$$

for all real t, $i = 1, 2, \ldots, n$. Hence we have

$$E(e^{tk_iX_i}) = \exp\left(\mu_i(k_i t) + \frac{\sigma_i^2(k_i t)^2}{2}\right).$$

That is, the moment-generating function of Y is

$$M(t) = \prod_{i=1}^{n} \exp\left[(k_i\mu_i)t + \frac{(k_i^2\sigma_i^2)t^2}{2}\right]$$

$$= \exp\left[\left(\sum_1^n k_i\mu_i\right)t + \frac{\left(\sum_1^n k_i^2\sigma_i^2\right)t^2}{2}\right].$$

But this is the moment-generating function of a distribution which is $n\left(\sum_1^n k_i\mu_i, \sum_1^n k_i^2\sigma_i^2\right)$. This is the desired result.

If, in Theorem 1, we set each $k_i = 1$, we see that the sum of n mutually stochastically independent normally distributed variables has a normal distribution. The next theorem proves a similar result for chi-square variables.

Theorem 2. *Let $X_1, X_2, \ldots, X_n$ be mutually stochastically independent variables that have, respectively, the chi-square distributions $\chi^2(r_1)$, $\chi^2(r_2), \ldots,$ and $\chi^2(r_n)$. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has a chi-square distribution with $r_1 + \cdots + r_n$ degrees of freedom; that is, Y is $\chi^2(r_1 + \cdots + r_n)$.*

Proof. The moment-generating function of Y is

$$M(t) = E\{\exp[t(X_1 + X_2 + \cdots + X_n)]\}$$
$$= E(e^{tX_1})E(e^{tX_2}) \cdots E(e^{tX_n})$$

because $X_1, X_2, \ldots, X_n$ are mutually stochastically independent. Since

$$E(e^{tX_i}) = (1 - 2t)^{-r_i/2}, \qquad t < \tfrac{1}{2}, \ i = 1, 2, \ldots, n,$$

we have

$$M(t) = (1 - 2t)^{-(r_1 + r_2 + \cdots + r_n)/2}, \qquad t < \tfrac{1}{2}.$$

But this is the moment-generating function of a distribution that is $\chi^2(r_1 + r_2 + \cdots + r_n)$. Accordingly, Y has this chi-square distribution.

Next, let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution which is $n(\mu, \sigma^2)$. In accordance with Theorem 2 of Section 3.4, each of the random variables $(X_i - \mu)^2/\sigma^2$, $i = 1, 2, \ldots, n$, is $\chi^2(1)$. Moreover, these n random variables are mutually stochastically independent. Accordingly, by Theorem 2, the random variable $Y = \sum_1^n [(X_i - \mu)/\sigma]^2$ is $\chi^2(n)$. This proves the following theorem:

Theorem 3. *Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution which is $n(\mu, \sigma^2)$. The random variable*

$$Y = \sum_1^n \left(\frac{X_i - \mu}{\sigma}\right)^2$$

has a chi-square distribution with n degrees of freedom.

Not always do we sample from a distribution of one random variable. Let the random variables X and Y have the joint p.d.f. $f(x, y)$ and let the $2n$ random variables $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ have the joint p.d.f.

$$f(x_1, y_1)f(x_2, y_2)\cdots f(x_n, y_n).$$

The n random pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are then mutually stochastically independent and are said to constitute a random sample of size n from the distribution of X and Y. In the next paragraph we shall take $f(x, y)$ to be the normal bivariate p.d.f., and we shall solve a problem in sampling theory when we are sampling from this two-variable distribution.

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ denote a random sample of size n from a bivariate normal distribution with p.d.f. $f(x, y)$ and parameters μ_1, μ_2, σ_1^2, σ_2^2, and ρ. We wish to find the joint p.d.f. of the two statistics $\bar{X} = \sum_1^n X_i/n$ and $\bar{Y} = \sum_1^n Y_i/n$. We call $\bar{X}$ the mean of $X_1, \ldots, X_n$ and $\bar{Y}$ the mean of $Y_1, \ldots, Y_n$. Since the joint p.d.f. of the $2n$ random variables (X_i, Y_i), $i = 1, 2, \ldots, n$, is given by

$$\varphi = f(x_1, y_1)f(x_2, y_2)\cdots f(x_n, y_n),$$

the moment-generating function of the two means $\bar{X}$ and $\bar{Y}$ is given by

$$\begin{aligned} M(t_1, t_2) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \exp\left(\frac{t_1\sum_1^n x_i}{n} + \frac{t_2\sum_1^n y_i}{n}\right)\varphi\, dx_1\cdots dy_n \\ &= \prod_{i=1}^{n}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(\frac{t_1x_i}{n} + \frac{t_2y_i}{n}\right)f(x_i, y_i)\, dx_i\, dy_i\right]. \end{aligned}$$

The justification of the form of the right-hand member of the second equality is that each pair (X_i, Y_i) has the same p.d.f., and that these n pairs are mutually stochastically independent. The twofold integral in the brackets in the last equality is the moment-generating function of X_i and Y_i (see p. 113) with t_1 replaced by t_1/n and t_2 replaced by t_2/n. Accordingly,

$$\begin{aligned} M(t_1, t_2) &= \prod_{i=1}^{n}\exp\left[\frac{t_1\mu_1}{n} + \frac{t_2\mu_2}{n} \right. \\ &\qquad \left. + \frac{\sigma_1^2(t_1/n)^2 + 2\rho\sigma_1\sigma_2(t_1/n)(t_2/n) + \sigma_2^2(t_2/n)^2}{2}\right] \\ &= \exp\left[t_1\mu_1 + t_2\mu_2 + \frac{(\sigma_1^2/n)t_1^2 + 2\rho(\sigma_1\sigma_2/n)t_1t_2 + (\sigma_2^2/n)t_2^2}{2}\right]. \end{aligned}$$

But this is the moment-generating function of a bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2/n and σ_2^2/n, and correlation coefficient ρ; therefore, $\bar{X}$ and $\bar{Y}$ have this joint distribution.

EXERCISES

4.64. Let the stochastically independent random variables X_1 and X_2 have the same p.d.f. $f(x) = \frac{1}{6}$, $x = 1, 2, 3, 4, 5, 6$, zero elsewhere. Find the p.d.f. of $Y = X_1 + X_2$. Note, under appropriate assumptions, that Y may be interpreted as the sum of the spots which appear when two dice are cast.

4.65. Let X_1 and X_2 be stochastically independent with normal distributions $n(6, 1)$ and $n(7, 1)$, respectively. Find $\Pr(X_1 > X_2)$. *Hint.* Write $\Pr(X_1 > X_2) = \Pr(X_1 - X_2 > 0)$ and determine the distribution of $X_1 - X_2$.

4.66. Let X_1 and X_2 be stochastically independent random variables. Let X_1 and $Y = X_1 + X_2$ have chi-square distributions with r_1 and r degrees of freedom, respectively. Here $r_1 < r$. Show that X_2 has a chi-square distribution with $r - r_1$ degrees of freedom. *Hint.* Write $M(t) = E(e^{t(X_1 + X_2)})$ and make use of the stochastic independence of X_1 and X_2.

4.67. Let the stochastically independent random variables X_1 and X_2 have binomial distributions with parameters n_1, $p_1 = \frac{1}{2}$ and n_2, $p_2 = \frac{1}{2}$, respectively. Show that $Y = X_1 - X_2 + n_2$ has a binomial distribution with parameters $n = n_1 + n_2$, $p = \frac{1}{2}$.

4.68. Let the stochastically independent random variables X_1 and X_2 have chi-square distributions with r_1 and r_2 degrees of freedom, respectively. Define the random variable Y by $Y = k_1X_1 + k_2X_2$, where k_1 and k_2 are real constants not equal to zero. When, and only when, will Y have a chi-square distribution?

4.69. Let X be $n(0, 1)$. Use the moment-generating function technique to show that $Y = X^2$ is $\chi^2(1)$. *Hint.* Evaluate the integral which represents $E(e^{tX^2})$ by writing $w = x\sqrt{1 - 2t}$, $t < \frac{1}{2}$.

4.70. Let $X_1, X_2, \ldots, X_n$ denote n mutually stochastically independent random variables with the moment-generating functions $M_1(t), M_2(t), \ldots, M_n(t)$, respectively. (a) Show that $Y = k_1X_1 + k_2X_2 + \cdots + k_nX_n$, where $k_1, k_2, \ldots, k_n$ are real constants, has the moment-generating function $M(t) = \prod_1^n M_i(k_it)$. (b) If each $k_i = 1$ and if X_i is Poisson with mean μ_i, $i = 1, 2, \ldots, n$, prove that Y is Poisson with mean $\mu_1 + \cdots + \mu_n$.

4.71. If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with moment-generating function $M(t)$, show that the moment-generating functions of $\sum_1^n X_i$ and $\sum_1^n X_i/n$ are, respectively, $[M(t)]^n$ and $[M(t/n)]^n$.

4.72. Let Y be the sum of three items of a random sample from a distribution that has p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Compute $\Pr(Y \leq 2)$.

4.73. If X and Y have a bivariate normal distribution with parameters μ_1, μ_2, σ_1^2, σ_2^2, and ρ, show that $Z = aX + bY + c$ is $n(a\mu_1 + b\mu_2 + c, a^2\sigma_1^2 + 2ab\rho\sigma_1\sigma_2 + b^2\sigma_2^2)$, where a, b, and c are constants. *Hint.* Use the moment-generating function $M(t_1, t_2)$ of X and Y to find the moment-generating function of Z.

4.74. Let X and Y have a bivariate normal distribution with parameters $\mu_1 = 25$, $\mu_2 = 35$, $\sigma_1^2 = 4$, $\sigma_2^2 = 16$, and $\rho = \frac{1}{2}$. If $Z = 3X - 2Y$, find $\Pr(-2 < Z < 19)$.

4.75. Let X_1, X_2 be a random sample from the normal distribution $n(0, 1)$. Let $Y = X_1 + X_2$ and $Z = X_1^2 + X_2^2$. Show that the moment-generating function of the joint distribution of Y and Z is

$$E\{\exp[t_1(X_1 + X_2) + t_2(X_1^2 + X_2^2)]\} = \frac{\exp[t_1^2/(1 - 2t_2)]}{1 - 2t_2}$$

for $-\infty < t_1 < \infty$, $-\infty < t_2 < \frac{1}{2}$. Find the correlation coefficient of Y and Z.

4.76. Let U and V be stochastically independent random variables, each having a normal distribution with mean zero and variance one. Show that the moment-generating function $E(e^{t(UV)})$ of the product UV is $(1 - t^2)^{-1/2}$, $-1 < t < 1$.

4.77. Let X and Y have a bivariate normal distribution with the parameters μ_1, μ_2, σ_1^2, σ_2^2, and ρ. Show that $W = X - \mu_1$ and $Z = (Y - \mu_2) - \rho(\sigma_2/\sigma_1)(X - \mu_1)$ are stochastically independent normal variables.

4.78. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the normal distribution $n(\mu, \sigma^2)$. Find the joint distribution of $Y = \sum_1^n a_i X_i$ and $Z = \sum_1^n b_i X_i$, where the a_i and b_i are real constants. When, and only when, are Y and Z stochastically independent?

4.79. Let X_1, X_2 be a random sample of size 2 from a distribution with positive variance and moment-generating function $M(t)$. If $Y = X_1 + X_2$ and $Z = X_1 - X_2$ are stochastically independent, prove that the distribution from which the sample is taken is a normal distribution. *Hint.* Show that $m(t_1, t_2) = E\{\exp[t_1(X_1 + X_2) + t_2(X_1 - X_2)]\} = M(t_1 + t_2)M(t_1 - t_2)$. Express each member of $m(t_1, t_2) \doteq m(t_1, 0)m(0, t_2)$ in terms of M; differentiate twice with respect to t_2; set $t_2 = 0$; and solve the resulting differential equation in M.

4.8 The Distributions of $\bar{X}$ and nS^2/σ^2

Let $X_1, X_2, \ldots, X_n$, denote a random sample of size $n \geq 2$ from a distribution which is $n(\mu, \sigma^2)$. In this section, we shall investigate the distributions of the mean and the variance of this random sample, that is, the distributions of the two statistics $\bar{X} = \sum_1^n X_i/n$ and $S^2 = \sum_1^n (X_i - \bar{X})^2/n$.

The problem of the distribution of $\bar{X}$, the mean of the sample, is solved by the use of Theorem 1 of Section 4.7. We have here, in the notation of the statement of that theorem, $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$, and $k_1 = k_2 = \cdots = k_n = 1/n$. Accordingly, $Y = \bar{X}$ has a normal distribution with mean and variance given by

$$\sum_1^n \left(\frac{1}{n}\mu\right) = \mu, \qquad \sum_1^n \left[\left(\frac{1}{n}\right)^2 \sigma^2\right] = \frac{\sigma^2}{n},$$

respectively. That is, $\bar{X}$ is $n(\mu, \sigma^2/n)$.

Example 1. Let $\bar{X}$ be the mean of a random sample of size 25 from a distribution which is $n(75, 100)$. Thus $\bar{X}$ is $n(75, 4)$. Then, for instance,

$$\begin{aligned}\Pr(71 < \bar{X} < 79) &= N\left(\frac{79 - 75}{2}\right) - N\left(\frac{71 - 75}{2}\right) \\ &= N(2) - N(-2) = 0.954.\end{aligned}$$

We now take up the problem of the distribution of S^2, the variance of a random sample $X_1, \ldots, X_n$ from a distribution which is $n(\mu, \sigma^2)$. It turns out that the random variable nS^2/σ^2 is somewhat more tractable. There is a very elementary proof of the nature of the distribution of nS^2/σ^2 if we will but establish the stochastic independence of $\bar{X}$, the mean of the sample, and the variance S^2. We shall first show that $\bar{X}$ is stochastically independent of $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$. It then follows that $\bar{X}$ is stochastically independent of $nS^2 = \sum_1^n (X_i - \bar{X})^2$. Let $M(t, t_1, \ldots, t_n)$ be the moment-generating function of $\bar{X}, X_1 - \bar{X}, \ldots, X_n - \bar{X}$. Thus

$$\begin{aligned}&M(t, t_1, \ldots, t_n) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[t\bar{x} + t_1(x_1 - \bar{x}) + \cdots + t_n(x_n - \bar{x})\right. \\ &\qquad \left. - \frac{\sum (x_i - \mu)^2}{2\sigma^2}\right] dx_1 \cdots dx_n.\end{aligned}$$

Consider the integral on x_i, $i = 1, 2, \ldots, n$. We have

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{[t + nt_i - (t_1 + \cdots + t_n)]\frac{x_i}{n} - \frac{(x_i - \mu)^2}{2\sigma^2}\right\} dx_i.$$

This integral exists for all real values of $t, t_1, \ldots, t_n$ and it may be interpreted as the moment-generating function of a distribution which is $n(\mu, \sigma^2)$, with the more familiar symbol t replaced by

$$\frac{1}{n}[t + nt_i - \sum t_j] = \frac{1}{n}[t + n(t_i - \bar{t})], \qquad \text{where } \bar{t} = \frac{\sum t_j}{n}.$$

Thus the integral is equal to

$$\exp\left\{\frac{\mu}{n}[t + n(t_i - \bar{t})] + \frac{\sigma^2[t + n(t_i - \bar{t})]^2}{2n^2}\right\},$$

and hence $M(t, t_1, \ldots, t_n)$ is equal to

$$\exp\left(\sum_{i=1}^{n}\left\{\frac{\mu}{n}t + \mu(t_i - \bar{t}) + \frac{\sigma^2[t + n(t_i - \bar{t})]^2}{2n^2}\right\}\right).$$

Since $\sum_1^n (t_i - \bar{t}) = 0$, we have

$$\begin{aligned} M(t, t_1, \ldots, t_n) &= \exp\left\{\mu t + \frac{\sigma^2 t^2}{2n} + \frac{\sigma^2 \sum (t_i - \bar{t})^2}{2}\right\} \\ &= \exp\left[\mu t + \frac{\sigma^2 t^2}{2n}\right] \exp\left\{\frac{\sigma^2 \sum (t_i - \bar{t})^2}{2}\right\}. \end{aligned}$$

In the right-hand member of the preceding equation, the first factor is the moment-generating function of $\bar{X}$ and the second factor does not depend upon t. In accordance with a generalization of Theorem 4, p. 80, $\bar{X}$ and the random variables $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$ are stochastically independent. Accordingly, $\bar{X}$ and $nS^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2$ are stochastically independent.

Remark. The second factor of $M(t, t_1, \ldots, t_n)$ is the moment-generating function of the joint distribution of $X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}$. At this time we do not know what distribution is implied by this moment-generating function, but that has nothing to do with the stochastic independence of $\bar{X}$ and $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$. Actually the implied distribution is called a *singular multivariate* normal distribution. *Nonsingular* multivariate normal distributions are discussed in Chapter 12.

We shall now prove that nS^2/σ^2 is $\chi^2(n-1)$. We have

$$\sum_1^n (X_i - \mu)^2 = \sum_1^n (X_i - \bar{X} + \bar{X} - \mu)^2$$
$$= \sum_1^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2,$$

because $2(\bar{X} - \mu)\sum_1^n (X_i - \bar{X}) = 0$. That is,

$$\frac{\sum_1^n (X_i - \mu)^2}{\sigma^2} = \frac{n(\bar{X} - \mu)^2}{\sigma^2} + \frac{nS^2}{\sigma^2}.$$

Theorem 3 of Section 4.7 states that $\sum_1^n (X_i - \mu)^2/\sigma^2$ is $\chi^2(n)$. We know that $\bar{X}$ is $n(\mu, \sigma^2/n)$ so that $\sqrt{n}(\bar{X} - \mu)/\sigma$ is $n(0, 1)$. Hence $n(\bar{X} - \mu)^2/\sigma^2$ is $\chi^2(1)$. Because $\bar{X}$ and S^2 are stochastically independent, $n(\bar{X} - \mu)^2/\sigma^2$ and nS^2/σ^2 are also. This stochastic independence permits us to write

$$E\left\{\exp\left[t\frac{\sum_1^n (X_i - \mu)^2}{\sigma^2}\right]\right\} = E\left\{\exp\left[t\frac{n(\bar{X} - \mu)^2}{\sigma^2} + t\frac{nS^2}{\sigma^2}\right]\right\}$$
$$= E\left\{\exp\left[t\frac{n(\bar{X} - \mu)^2}{\sigma^2}\right]\right\}E(e^{t(nS^2/\sigma^2)}).$$

That is,

$$(1 - 2t)^{-n/2} = (1 - 2t)^{-1/2}E(e^{t(nS^2/\sigma^2)}), \qquad t < \tfrac{1}{2}.$$

Consequently, the moment-generating function of nS^2/σ^2 is

$$E(e^{t(nS^2/\sigma^2)}) = (1 - 2t)^{-(n-1)/2}, \qquad t < \tfrac{1}{2}.$$

But this is the moment-generating function of a chi-square distribution with $n - 1$ degrees of freedom, so nS^2/σ^2 is $\chi^2(n-1)$. The determination of the p.d.f. of S^2 is left as an exercise.

EXERCISES

4.80. Let $\bar{X}$ be the mean of a random sample of size 5 from a normal distribution with $\mu = 0$ and $\sigma^2 = 125$. Determine c so that $\Pr(\bar{X} < c) = 0.90$.

4.81. If $\bar{X}$ is the mean of a random sample of size n from a normal distribution with mean μ and variance 100, find n so that $\Pr(\mu - 5 < \bar{X} < \mu + 5) = 0.954$.

4.82. Let X_1 and X_2 be stochastically independent random variables. Let X_1 and $Y = X_1 + X_2$ have Poisson distributions with parameters μ_1 and μ, respectively. Here $\mu_1 < \mu$. Show that X_2 has a Poisson distribution with parameter $\mu - \mu_1$.

4.83. Let $X_1, X_2, \ldots, X_{25}$ and $Y_1, Y_2, \ldots, Y_{25}$ be two random samples from two independent normal distributions $n(0, 16)$ and $n(1, 9)$, respectively. Let $\bar{X}$ and $\bar{Y}$ denote the corresponding sample means. Compute $\Pr(\bar{X} > \bar{Y})$.

4.84. Find the mean and variance of $S^2 = \sum_1^n (X_i - \bar{X})^2/n$, where $X_1, X_2, \ldots, X_n$ is a random sample from $n(\mu, \sigma^2)$. *Hint.* Find the mean and variance of nS^2/σ^2.

4.85. Let S^2 be the variance of a random sample of size 6 from the normal distribution $n(\mu, 12)$. Find $\Pr(2.30 < S^2 < 22.2)$.

4.86. Find the p.d.f. of the sample variance S^2 provided the distribution from which the sample arises is $n(\mu, \sigma^2)$.

4.87. Let $\bar{X}$ and S^2 be the mean and the variance of a random sample of size 25 from a distribution which is $n(3, 100)$. Evaluate $\Pr(0 < \bar{X} < 6, 55.2 < S^2 < 145.6)$.

4.9 Expectations of Functions of Random Variables

Let $X_1, X_2, \ldots, X_n$ denote random variables that have the joint p.d.f. $f(x_1, x_2, \ldots, x_n)$. Let the random variable Y be defined by $Y = u(X_1, X_2, \ldots, X_n)$. We found, in Section 4.7, that we can compute expectations of functions of Y without first finding the p.d.f. of Y. Indeed, this fact was the basis of the moment-generating function procedure for finding the p.d.f. of Y. We can take advantage of this fact in a number of other instances. Some illustrative examples will be given.

Example 1. Given that W is $n(0, 1)$, that V is $\chi^2(r)$ with $r \geq 2$, and let W and V be stochastically independent. The mean of the random variable $T = W\sqrt{r/V}$ exists and is zero because the graph of the p.d.f. of T (see Section 4.4) is symmetric about the vertical axis through $t = 0$. The variance of T, when it exists, could be computed by integrating the product of t^2 and the p.d.f. of T. But it seems much simpler to compute

$$\sigma_T^2 = E(T^2) = E\left[W^2 \frac{r}{V}\right] = E(W^2)E\left[\frac{r}{V}\right].$$

Now W^2 is $\chi^2(1)$, so $E(W^2) = 1$. Furthermore,

$$E\left(\frac{r}{V}\right) = \int_0^\infty \frac{r}{v} \frac{1}{2^{r/2}\Gamma(r/2)} v^{r/2-1} e^{-v/2}\, dv$$

exists if $r > 2$ and is given by

$$\frac{r\Gamma[(r-2)/2]}{2\Gamma(r/2)} = \frac{r\Gamma[(r-2)/2]}{2[(r-2)/2]\Gamma[(r-2)/2]} = \frac{r}{r-2}.$$

Thus, $\sigma_T^2 = r/(r-2)$, $r > 2$.

Example 2. Let X_i denote a random variable with mean μ_i and variance σ_i^2, $i = 1, 2, \ldots, n$. Let $X_1, X_2, \ldots, X_n$ be mutually stochastically independent and let $k_1, k_2, \ldots, k_n$ denote real constants. We shall compute the mean and variance of the linear function $Y = k_1X_1 + k_2X_2 + \cdots + k_nX_n$. Because E is a linear operator, the mean of Y is given by

$$\begin{aligned}\mu_Y &= E(k_1X_1 + k_2X_2 + \cdots + k_nX_n)\\ &= k_1E(X_1) + k_2E(X_2) + \cdots + k_nE(X_n)\\ &= k_1\mu_1 + k_2\mu_2 + \cdots + k_n\mu_n = \sum_1^n k_i\mu_i.\end{aligned}$$

The variance of Y is given by

$$\begin{aligned}\sigma_Y^2 &= E\{[(k_1X_1 + \cdots + k_nX_n) - (k_1\mu_1 + \cdots + k_n\mu_n)]^2\}\\ &= E\{[k_1(X_1 - \mu_1) + \cdots + k_n(X_n - \mu_n)]^2\}\\ &= E\Big\{\sum_{i=1}^n k_i^2(X_i - \mu_i)^2 + 2\sum\sum_{i<j} k_ik_j(X_i - \mu_i)(X_j - \mu_j)\Big\}\\ &= \sum_{i=1}^n k_i^2E[(X_i - \mu_i)^2] + 2\sum\sum_{i<j} k_ik_jE[(X_i - \mu_i)(X_j - \mu_j)].\end{aligned}$$

Consider $E[(X_i - \mu_i)(X_j - \mu_j)]$, $i < j$. Because X_i and X_j are stochastically independent, we have

$$E[(X_i - \mu_i)(X_j - \mu_j)] = E(X_i - \mu_i)E(X_j - \mu_j) = 0.$$

Finally, then,

$$\sigma_Y^2 = \sum_{i=1}^n k_i^2E[(X_i - \mu_i)^2] = \sum_{i=1}^n k_i^2\sigma_i^2.$$

We can obtain a more general result if, in Example 2, we remove the hypothesis of mutual stochastic independence of $X_1, X_2, \ldots, X_n$. We shall do this and we shall let ρ_{ij} denote the correlation coefficient of X_i and X_j. Thus, for easy reference to Example 2, we write

$$E[(X_i - \mu_i)(X_j - \mu_j)] = \rho_{ij}\sigma_i\sigma_j, \qquad i < j.$$

If we refer to Example 2, we see that again $\mu_Y = \sum_1^n k_i\mu_i$. But now

$$\sigma_Y^2 = \sum_1^n k_i^2\sigma_i^2 + 2\sum\sum_{i<j} k_ik_j\rho_{ij}\sigma_i\sigma_j.$$

Thus we have the following theorem.

Theorem 4. *Let $X_1, \ldots, X_n$ denote random variables that have means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$. Let ρ_{ij}, $i \neq j$, denote the correlation coefficient of X_i and X_j and let $k_1, \ldots, k_n$ denote real constants. The mean and the variance of the linear function*

$$Y = \sum_1^n k_i X_i$$

are, respectively,

$$\mu_Y = \sum_1^n k_i \mu_i$$

and

$$\sigma_Y^2 = \sum_1^n k_i^2 \sigma_i^2 + 2 \sum\sum_{i<j} k_i k_j \rho_{ij} \sigma_i \sigma_j.$$

The following corollary of this theorem is quite useful.

Corollary. *Let $X_1, \ldots, X_n$ denote the items of a random sample of size n from a distribution that has mean μ and variance σ^2. The mean and the variance of $Y = \sum_1^n k_i X_i$ are, respectively, $\mu_Y = \left(\sum_1^n k_i\right)\mu$ and $\sigma_Y^2 = \left(\sum_1^n k_i^2\right)\sigma^2$.*

Example 3. Let $\bar{X} = \sum_1^n X_i/n$ denote the mean of a random sample of size n from a distribution that has mean μ and variance σ^2. In accordance with the Corollary, we have $\mu_{\bar{X}} = \mu \sum_1^n (1/n) = \mu$ and $\sigma_{\bar{X}}^2 = \sigma^2 \sum_1^n (1/n)^2 = \sigma^2/n$. We have seen, in Section 4.8, that if our sample is from a distribution that is $n(\mu, \sigma^2)$, then $\bar{X}$ is $n(\mu, \sigma^2/n)$. It is interesting that $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}}^2 = \sigma^2/n$ whether the sample is or is not from a normal distribution.

EXERCISES

4.88. Let X_1, X_2, X_3, X_4 be four mutually stochastically independent random variables having the same p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. Find the mean and variance of the sum Y of these four random variables.

4.89. Let X_1 and X_2 be two stochastically independent random variables so that the variances of X_1 and X_2 are $\sigma_1^2 = k$ and $\sigma_2^2 = 2$, respectively. Given that the variance of $Y = 3X_2 - X_1$ is 25, find k.

4.90. If the stochastically independent variables X_1 and X_2 have means μ_1, μ_2 and variances σ_1^2, σ_2^2, respectively, show that the mean and variance of the product $Y = X_1 X_2$ are $\mu_1\mu_2$ and $\sigma_1^2\sigma_2^2 + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2$, respectively.

4.91. Find the mean and variance of the sum Y of the items of a random sample of size 5 from the distribution having p.d.f. $f(x) = 6x(1 - x)$, $0 < x < 1$, zero elsewhere.

4.92. Let Y_2 denote the median of a random sample X_1, X_2, X_3 of size 3 from a continuous-type uniform distribution on the interval $(0, \theta)$. Compare $E(Y_2)$ and $E(\bar{X})$, where $\bar{X}$ is the mean of the sample.

4.93. Let Y_2 denote the median of a random sample X_1, X_2, X_3 of size 3 from a distribution with p.d.f. $f(x) = 2x$, $0 < x < 1$, zero elsewhere. Compare $E(Y_2)$ and $E(\bar{X})$, where $\bar{X}$ is the mean of the sample.

4.94. Determine the mean and variance of the mean $\bar{X}$ of a random sample of size 9 from a distribution having p.d.f. $f(x) = 4x^3$, $0 < x < 1$, zero elsewhere.

4.95. Let X and Y be random variables with $\mu_1 = 1$, $\mu_2 = 4$, $\sigma_1^2 = 4$, $\sigma_2^2 = 6$, $\rho = \frac{1}{2}$. Find the mean and variance of $Z = 3X - 2Y$.

4.96. Let X and Y be stochastically independent random variables with means μ_1, μ_2 and variances σ_1^2, σ_2^2. Determine the correlation coefficient of X and $Z = X - Y$ in terms of μ_1, μ_2, σ_1^2, σ_2^2.

4.97. Let X and Y be random variables with means μ_1, μ_2; variances σ_1^2, σ_2^2; and correlation coefficient ρ. Show that the correlation coefficient of $W = aX + b$, $a > 0$, and $Z = cY + d$, $c > 0$, is ρ.

4.98. Let μ and σ^2 denote the mean and variance of the random variable X. Let $Y = c + bX$, where b and c are real constants. Show that the mean and the variance of Y are, respectively, $c + b\mu$ and $b^2\sigma^2$.

4.99. A person rolls a die, tosses a coin, and draws a card from an ordinary deck. He receives three dollars for each point up on the die, ten dollars for a head, zero dollars for a tail, and one dollar for each spot on the card (jack = 11, queen = 12, king = 13). If we assume that the three random variables involved are mutually stochastically independent and uniformly distributed, compute the mean and variance of the amount to be received.

4.100. Let U and V be two stochastically independent chi-square variables with r_1 and r_2 degrees of freedom, respectively. Find the mean and variance of $F = (r_2U)/(r_1V)$. What restriction is needed on the parameters r_1 and r_2 in order to ensure the existence of both the mean and the variance of F?

4.101. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution with mean μ and variance σ^2. Show that $E(S^2) = (n - 1)\sigma^2/n$, where S^2 is the variance of the random sample.

4.102. Let X_1 and X_2 be stochastically independent random variables with nonzero variances. Find the correlation coefficient of $Y = X_1X_2$ and X_1 in terms of the means and variances of X_1 and X_2.

4.103. Let X_1 and X_2 have a joint distribution with parameters μ_1, μ_2, σ_1^2, σ_2^2, and ρ. Find the correlation coefficient of the linear functions $Y = a_1X_1 + a_2X_2$ and $Z = b_1X_1 + b_2X_2$ in terms of the real constants a_1, a_2, b_1, b_2, and the parameters of the distribution.

4.104. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution which has mean μ and variance σ^2. Use Chebyshev's inequality to show, for every $\epsilon > 0$, that $\lim_{n \to \infty} \Pr(|\overline{X} - \mu| < \epsilon) = 1$; this is another form of the law of large numbers.

4.105. Let X denote a random variable of the continuous type whose mean is μ. Let $Y = v(X)$, let $E(Y)$ exist, and let $v(x)$ be given by $v(x) = v(\mu) + v'(\mu)(x - \mu) + v''(\xi)(x - \mu)^2/2$, where ξ is between μ and x. If $v''(x) \geq 0$ for all x, show that $E[v(X)] \geq v(\mu)$. If $v''(x) \leq 0$ for all x, show that $E[v(X)] \leq v(\mu)$.

4.106. Let X_1, X_2, and X_3 be random variables with equal variances but with correlation coefficients $\rho_{12} = 0.3$, $\rho_{13} = 0.5$, and $\rho_{23} = 0.2$. Find the correlation coefficient of the linear functions $Y = X_1 + X_2$ and $Z = X_2 + X_3$.

4.107. Find the variance of the sum of 10 random variables if each has variance 5 and if each pair has correlation coefficient 0.5.

4.108. Let $X_1, \ldots, X_n$ be random variables which have means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$. Let ρ_{ij}, $i \neq j$, denote the correlation coefficient of X_i and X_j. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be real constants. Show that the covariance of $Y = \sum_{i=1}^{n} a_iX_i$ and $Z = \sum_{j=1}^{n} b_jX_j$ is $\sum_{j=1}^{n} \sum_{i=1}^{n} a_ib_j\sigma_i\sigma_j\rho_{ij}$, where $\rho_{ii} = 1$, $i = 1, 2, \ldots, n$.

Chapter 5

Limiting Distributions

5.1 Limiting Distributions

In some of the preceding chapters it has been demonstrated by example that the distribution of a random variable (perhaps a statistic) often depends upon a positive integer n. For example, if the random variable X is $b(n, p)$, the distribution of X depends upon n. If $\bar{X}$ is the mean of a random sample of size n from a distribution which is $n(\mu, \sigma^2)$, then $\bar{X}$ is itself $n(\mu, \sigma^2/n)$ and the distribution of $\bar{X}$ depends upon n. If S^2 is the variance of this random sample from the normal distribution to which we have just referred, the random variable nS^2/σ^2 is $\chi^2(n - 1)$, and so the distribution of this random variable depends upon n.

We know from experience that the determination of the p.d.f. of a random variable can, upon occasion, present rather formidable computational difficulties. For example, if $\bar{X}$ is the mean of a random sample $X_1, X_2, \ldots, X_n$ from a distribution that has the p.d.f.

$$\begin{aligned} f(x) &= 1, \qquad 0 < x < 1, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

then (p. 161) the moment-generating function of $\bar{X}$ is given by $[M(t/n)]^n$, where here

$$\begin{aligned} M(t) &= \int_0^1 e^{tx}\, dx = \frac{e^t - 1}{t}, \qquad t \neq 0, \\ &= 1, \qquad t = 0. \end{aligned}$$

Hence

$$E(e^{t\bar{X}}) = \left(\frac{e^{t/n} - 1}{t/n}\right)^n, \qquad t \neq 0,$$
$$= 1, \qquad t = 0.$$

Since the moment-generating function of $\bar{X}$ depends upon n, the distribution of $\bar{X}$ depends upon n. It is true that various mathematical techniques can be used to determine the p.d.f. of $\bar{X}$ for a fixed, but arbitrarily fixed, positive integer n. But the p.d.f. is so complicated that few, if any, of us would be interested in using it to compute probabilities about $\bar{X}$. One of the purposes of this chapter is to provide ways of approximating, for large values of n, some of these complicated probability density functions.

Consider a distribution that depends upon the positive integer n. Clearly, the distribution function F of that distribution will also depend upon n. Throughout this chapter, we denote this fact by writing the distribution function as F_n and the corresponding p.d.f. as f_n. Moreover, to emphasize the fact that we are working with sequences of distribution functions, we place a subscript n on the random variables. For example, we shall write

$$F_n(\bar{x}) = \int_{-\infty}^{x} \frac{1}{\sqrt{1/n}\sqrt{2\pi}} e^{-nw^2/2}\, dw$$

for the distribution function of the mean $\bar{X}_n$ of a random sample of size n from a normal distribution with mean zero and variance one.

We now define a limiting distribution of a random variable whose distribution depends upon n.

Definition 1. Let the distribution function $F_n(y)$ of the random variable Y_n depend upon n, a positive integer. If $F(y)$ is a distribution function and if $\lim_{n\to\infty} F_n(y) = F(y)$ for every point y at which $F(y)$ is continuous, then the random variable Y_n is said to have a *limiting distribution* with distribution function $F(y)$.

The following examples are illustrative of random variables that have limiting distributions.

Example 1. Let Y_n denote the nth order statistic of a random sample $X_1, X_2, \ldots, X_n$ from a distribution having p.d.f.

$$f(x) = \frac{1}{\theta}, \qquad 0 < x < \theta, 0 < \theta < \infty,$$
$$= 0 \text{ elsewhere.}$$

The p.d.f. of Y_n is

$$g_n(y) = \frac{ny^{n-1}}{\theta^n}, \qquad 0 < y < \theta,$$
$$= 0 \text{ elsewhere,}$$

and the distribution function of Y_n is

$$F_n(y) = 0, \qquad y < 0,$$
$$= \int_0^y \frac{nz^{n-1}}{\theta^n}\,dz = \left(\frac{y}{\theta}\right)^n, \qquad 0 \le y < \theta,$$
$$= 1, \qquad \theta \le y < \infty.$$

Then

$$\lim_{n\to\infty} F_n(y) = 0, \qquad -\infty < y < \theta,$$
$$= 1, \qquad \theta \le y < \infty.$$

Now

$$F(y) = 0, \qquad -\infty < y < \theta,$$
$$= 1, \qquad \theta \le y < \infty,$$

is a distribution function. Moreover, $\lim_{n\to\infty} F_n(y) = F(y)$ at each point of continuity of $F(y)$. In accordance with the definition of a limiting distribution, the random variable Y_n has a limiting distribution with distribution function $F(y)$. Recall that a distribution of the discrete type which has a probability of one at a single point has been called a *degenerate* distribution. Thus, in this example, the limiting distribution of Y_n is degenerate. Sometimes this is the case, sometimes a limiting distribution is not degenerate, and sometimes there is no limiting distribution at all.

Example 2. Let $\bar{X}_n$ have the distribution function

$$F_n(\bar{x}) = \int_{-\infty}^{\bar{x}} \frac{1}{\sqrt{1/n}\sqrt{2\pi}} e^{-nw^2/2}\,dw.$$

If the change of variable $v = \sqrt{n}w$ is made, we have

$$F_n(\bar{x}) = \int_{-\infty}^{\sqrt{n}\bar{x}} \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\,dv.$$

It is clear that

$$\lim_{n\to\infty} F_n(\bar{x}) = 0, \qquad \bar{x} < 0,$$
$$= \tfrac{1}{2}, \qquad \bar{x} = 0,$$
$$= 1, \qquad \bar{x} > 0.$$

Now the function

$$F(\bar{x}) = 0, \qquad \bar{x} < 0,$$
$$= 1, \qquad \bar{x} \ge 0,$$

is a distribution function and $\lim_{n\to\infty} F_n(\bar{x}) = F(\bar{x})$ at every point of continuity of $F(\bar{x})$. To be sure, $\lim_{n\to\infty} F_n(0) \neq F(0)$, but $F(\bar{x})$ is not continuous at $\bar{x} = 0$. Accordingly, the random variable $\bar{X}_n$ has a limiting distribution with distribution function $F(\bar{x})$. Again this limiting distribution is degenerate and has all the probability at the one point $\bar{x} = 0$.

Example 3. The fact that limiting distributions, if they exist, cannot in general be determined by taking the limit of the p.d.f. will now be illustrated. Let X_n have the p.d.f.

$$\begin{aligned} f_n(x) &= 1, \qquad x = 2 + \frac{1}{n}, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Clearly, $\lim_{n\to\infty} f_n(x) = 0$ for all values of x. This may suggest that X_n has no limiting distribution. However, the distribution function of X_n is

$$\begin{aligned} F_n(x) &= 0, \qquad x < 2 + \frac{1}{n}, \\ &= 1, \qquad x \geq 2 + \frac{1}{n}, \end{aligned}$$

and

$$\begin{aligned} \lim_{n\to\infty} F_n(x) &= 0, \qquad x \leq 2, \\ &= 1, \qquad x > 2. \end{aligned}$$

Since

$$\begin{aligned} F(x) &= 0, \qquad x < 2, \\ &= 1, \qquad x \geq 2, \end{aligned}$$

is a distribution function, and since $\lim_{n\to\infty} F_n(x) = F(x)$ at all points of continuity of $F(x)$, there is a limiting distribution of X_n with distribution function $F(x)$.

Example 4. Let Y_n denote the nth order statistic of a random sample from the uniform distribution of Example 1. Let $Z_n = n(\theta - Y_n)$. The p.d.f. of Z_n is

$$\begin{aligned} h_n(z) &= \frac{(\theta - z/n)^{n-1}}{\theta^n}, \qquad 0 < z < n\theta, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

and the distribution function of Z_n is

$$\begin{aligned} G_n(z) &= 0, \qquad z < 0, \\ &= \int_0^z \frac{(\theta - w/n)^{n-1}}{\theta^n}\, dw = 1 - \left(1 - \frac{z}{n\theta}\right)^n, \qquad 0 \leq z < n\theta, \\ &= 1, \qquad n\theta \leq z. \end{aligned}$$

Hence

$$\lim_{n\to\infty} G_n(z) = 0, \qquad z \leq 0$$
$$= 1 - e^{-z/\theta}, \qquad 0 < z < \infty.$$

Now

$$G(z) = 0, \qquad z < 0,$$
$$= 1 - e^{-z/\theta}, \qquad 0 \leq z,$$

is a distribution function which is everywhere continuous and $\lim_{n\to\infty} G_n(z) = G(z)$ at all points. Thus Z_n has a limiting distribution with distribution function $G(z)$. This affords us an example of a limiting distribution which is not degenerate.

EXERCISES

5.1. Let $\bar{X}_n$ denote the mean of a random sample of size n from a distribution which is $n(\mu, \sigma^2)$. Find the limiting distribution of $\bar{X}_n$.

5.2. Let Y_1 denote the first order statistic of a random sample of size n from a distribution that has the p.d.f. $f(x) = e^{-(x-\theta)}$, $\theta < x < \infty$, zero elsewhere. Let $Z_n = n(Y_1 - \theta)$. Investigate the limiting distribution of Z_n.

5.3. Let Y_n denote the nth order statistic of a random sample from a distribution of the continuous type that has distribution function $F(x)$ and p.d.f. $f(x) = F'(x)$. Find the limiting distribution of $Z_n = n[1 - F(Y_n)]$.

5.4. Let the p.d.f. of Y_n be $f_n(y) = 1$, $y = n$, zero elsewhere. Show that Y_n does not have a limiting distribution. (The student should accept, at this time, the fact that a distribution function F has at most a countable number of discontinuities.)

5.5. Let Y_2 denote the second order statistic of a random sample of size n from a distribution of the continuous type that has distribution function $F(x)$ and p.d.f. $f(x) = F'(x)$. Find the limiting distribution of $W_n = nF(Y_2)$.

5.6. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution which is $n(\mu, \sigma^2)$, where $\mu > 0$. Show that the sum $Z_n = \sum_1^n X_i$ does not have a limiting distribution.

5.2 Stochastic Convergence

When the limiting distribution of a random variable is degenerate, the random variable is said to *converge stochastically* to the constant that has a probability of one. Thus Examples 1 to 3 of Section 5.1 illustrate not only the notion of a limiting distribution but also the concept of stochastic convergence. In Example 1, the nth order

statistic Y_n converges stochastically to θ; in Example 2, the statistic $\bar{X}_n$ converges stochastically to zero, the mean of the normal distribution from which the sample was taken; and in Example 3, the random variable X_n converges stochastically to 2. We shall show that in some instances the inequality of Chebyshev can be used to advantage in proving stochastic convergence. But first we shall prove the following theorem.

Theorem 1. *Let $F_n(y)$ denote the distribution function of a random variable Y_n whose distribution depends upon the positive integer n. Let c denote a constant which does not depend upon n. The random variable Y_n converges stochastically to the constant c if and only if, for every $\epsilon > 0$, the*

$$\lim_{n\to\infty} \Pr(|Y_n - c| < \epsilon) = 1.$$

Proof. First, assume that the $\lim_{n\to\infty} \Pr(|Y_n - c| < \epsilon) = 1$ for every $\epsilon > 0$. We are to prove that the random variable Y_n converges stochastically to the constant c. This means we must prove that

$$\begin{aligned} \lim_{n\to\infty} F_n(y) &= 0, \qquad y < c, \\ &= 1, \qquad y > c. \end{aligned}$$

Note that we do not need to know anything about the $\lim_{n\to\infty} F_n(c)$. For if the limit of $F_n(y)$ is as indicated, then Y_n has a limiting distribution with distribution function

$$\begin{aligned} F(y) &= 0, \qquad y < c, \\ &= 1, \qquad y \geq c. \end{aligned}$$

Now

$$\Pr(|Y_n - c| < \epsilon) = F_n[(c + \epsilon)-] - F_n(c - \epsilon),$$

where $F_n[(c + \epsilon)-]$ is the left-hand limit of $F_n(y)$ at $y = c + \epsilon$. Thus we have

$$1 = \lim_{n\to\infty} \Pr(|Y_n - c| < \epsilon) = \lim_{n\to\infty} F_n[(c + \epsilon)-] - \lim_{n\to\infty} F_n(c - \epsilon).$$

Because $0 \leq F_n(y) \leq 1$ for all values of y and for every positive integer n, it must be that

$$\lim_{n\to\infty} F_n(c - \epsilon) = 0, \qquad \lim_{n\to\infty} F_n[(c + \epsilon)-] = 1.$$

Since this is true for every $\epsilon > 0$, we have

$$\begin{aligned} \lim_{n\to\infty} F_n(y) &= 0, \qquad y < c, \\ &= 1, \qquad y > c, \end{aligned}$$

as we were required to show.

To complete the proof of Theorem 1, we assume that

$$\lim_{n\to\infty} F_n(y) = 0, \qquad y < c,$$
$$= 1, \qquad y > c.$$

We are to prove that $\lim_{n\to\infty} \Pr(|Y_n - c| < \epsilon) = 1$ for every $\epsilon > 0$. Because

$$\Pr(|Y_n - c| < \epsilon) = F_n[(c + \epsilon)-] - F_n(c - \epsilon),$$

and because it is given that

$$\lim_{n\to\infty} F_n[(c + \epsilon)-] = 1,$$
$$\lim_{n\to\infty} F_n(c - \epsilon) = 0,$$

for every $\epsilon > 0$, we have the desired result. This completes the proof of the theorem.

We should like to point out a simple but useful fact. Clearly,

$$\Pr(|Y_n - c| < \epsilon) + \Pr(|Y_n - c| \geq \epsilon) = 1.$$

Thus the limit of $\Pr(|Y_n - c| < \epsilon)$ is equal to one when and only when

$$\lim_{n\to\infty} \Pr(|Y_n - c| \geq \epsilon) = 0.$$

That is, this last limit is also a necessary and sufficient condition for the stochastic convergence of the random variable Y_n to the constant c.

Example 1. Let $\overline{X}_n$ denote the mean of a random sample of size n from a distribution which has mean μ and positive variance σ^2. Then the mean and variance of $\overline{X}_n$ are μ and σ^2/n. Consider, for every fixed $\epsilon > 0$, the probability

$$\Pr(|\overline{X}_n - \mu| \geq \epsilon) = \Pr\left(|\overline{X}_n - \mu| \geq \frac{k\sigma}{\sqrt{n}}\right),$$

where $k = \epsilon\sqrt{n}/\sigma$. In accordance with the inequality of Chebyshev, this probability is less than or equal to $1/k^2 = \sigma^2/n\epsilon^2$. So, for every fixed $\epsilon > 0$, we have

$$\lim_{n\to\infty} \Pr(|\overline{X}_n - \mu| \geq \epsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0.$$

Hence $\overline{X}_n$ converges stochastically to μ if σ^2 is finite. In a more advanced course, the student will learn that μ finite is sufficient to ensure this stochastic convergence.

Remark. The condition $\lim_{n\to\infty} \Pr(|Y_n - c| < \epsilon) = 1$ is often used as the definition of convergence in probability and one says that Y_n converges

to c *in probability*. Thus, stochastic convergence and convergence in probability are equivalent. A stronger type of convergence is given by $\Pr\left(\lim_{n\to\infty} |Y_n - c| < \epsilon\right) = 1$, for every $\epsilon > 0$; in this case we say that Y_n converges to c *with probability one*. Although we do not consider this type of convergence, it is known that the mean $\bar{X}_n$ of a random sample converges, with probability one, to the mean μ of the distribution, provided the latter exists. This is one form of the *strong law of large numbers*.

EXERCISES

5.7. Let the random variable Y_n have a distribution which is $b(n, p)$. (a) Prove that Y_n/n converges stochastically to p. This result is one form of the weak law of large numbers. (b) Prove that $1 - Y_n/n$ converges stochastically to $1 - p$.

5.8. Let S_n^2 denote the variance of a random sample of size n from a distribution which is $n(\mu, \sigma^2)$. Prove that $nS_n^2/(n-1)$ converges stochastically to σ^2.

5.9. Let W_n denote a random variable with mean μ and variance b/n^p, where $p > 0$, μ, and b are constants (not functions of n). Prove that W_n converges stochastically to μ.

5.10. Let Y_n denote the nth order statistic of a random sample of size n from a uniform distribution on the interval $(0, \theta)$, as in Example 1 of Section 5.1. Prove that $Z_n = \sqrt{Y_n}$ converges stochastically to $\sqrt{\theta}$.

5.11. Let X have a distribution with mean μ and let $E[(X-\mu)^k] = \mu_k$ exist for $k = 2, 3, 4$. Let S_n^2 denote the variance of a random sample of size n from this distribution. Then the mean of S_n^2 is given by $(n-1)\mu_2/n = (n-1)\sigma^2/n$ and the variance of S_n^2 is given by $(\mu_4 - \mu_2^2)/n - 2(\mu_4 - 2\mu_2^2)/n^2 + (\mu_4 - 3\mu_2^2)/n^3$. Prove that $nS_n^2/(n-1)$ converges stochastically to σ^2.

5.3 Limiting Moment-Generating Functions

To find the limiting distribution function of a random variable Y_n by use of the definition of limiting distribution function obviously requires that we know $F_n(y)$ for each positive integer n. But, as indicated in the introductory remarks of Section 5.1, this is precisely the problem we should like to avoid. If it exists, the moment-generating function that corresponds to the distribution function $F_n(y)$ often provides a convenient method of determining the limiting distribution function. To emphasize that the distribution of a random variable Y_n depends upon the positive integer n, in this chapter we shall write the moment-generating function of Y_n in the form $M(t; n)$.

The following theorem, which is essentially Curtiss' modification of a theorem of Lévy and Cramér, explains how the moment-generating function may be used in problems of limiting distributions. A proof of the theorem requires a knowledge of that same facet of analysis that permitted us to assert that a moment-generating function, when it exists, uniquely determines a distribution. Accordingly, no proof of the theorem will be given.

Theorem 2. *Let the random variable Y_n have the distribution function $F_n(y)$ and the moment-generating function $M(t; n)$ that exists for $-h < t < h$ for all n. If there exists a distribution function $F(y)$, with corresponding moment-generating function $M(t)$, defined for $|t| \leq h_1 < h$, such that $\lim_{n\to\infty} M(t; n) = M(t)$, then Y_n has a limiting distribution with distribution function $F(y)$.*

In this and the subsequent section are several illustrations of the use of Theorem 2. In some of these examples it is convenient to use a certain limit that is established in some courses in advanced calculus. We refer to a limit of the form

$$\lim_{n\to\infty}\left[1 + \frac{b}{n} + \frac{\psi(n)}{n}\right]^{cn},$$

where b and c do not depend upon n and where $\lim_{n\to\infty} \psi(n) = 0$. Then

$$\lim_{n\to\infty}\left[1 + \frac{b}{n} + \frac{\psi(n)}{n}\right]^{cn} = \lim_{n\to\infty}\left[1 + \frac{b}{n}\right]^{cn} = e^{bc}.$$

For example,

$$\lim_{n\to\infty}\left[1 - \frac{t^2}{n} + \frac{t^3}{n^{3/2}}\right]^{-n/2} = \lim_{n\to\infty}\left[1 - \frac{t^2}{n} + \frac{t^3/\sqrt{n}}{n}\right]^{-n/2}.$$

Here $b = -t^2$, $c = -\frac{1}{2}$, and $\psi(n) = t^3/\sqrt{n}$. Accordingly, for every fixed value of t, the limit is $e^{t^2/2}$.

Example 1. Let Y_n have a distribution which is $b(n, p)$. Suppose that the mean $\mu = np$ is the same for every n; that is, $p = \mu/n$, where μ is a constant. We shall find the limiting distribution of the binomial distribution, when $p = \mu/n$, by finding the limit of $M(t; n)$. Now

$$M(t; n) = E(e^{tY_n}) = [(1 - p) + pe^t]^n = \left[1 + \frac{\mu(e^t - 1)}{n}\right]^n$$

for all real values of t. Hence we have

$$\lim_{n\to\infty} M(t; n) = e^{\mu(e^t - 1)}$$

for all real values of t. Since there exists a distribution, namely, the Poisson distribution with mean μ, that has this moment-generating function $e^{\mu(e^t-1)}$, then in accordance with the theorem and under the conditions stated, it is seen that Y_n has a limiting Poisson distribution with mean μ.

Whenever a random variable has a limiting distribution, we may, if we wish, use the limiting distribution as an approximation to the exact distribution function. The result of this example enables us to use the Poisson distribution as an approximation to the binomial distribution when n is large and p is small. This is clearly an advantage, for it is easy to provide tables for the one-parameter Poisson distribution. On the other hand, the binomial distribution has two parameters, and tables for this distribution are very ungainly. To illustrate the use of the approximation, let Y have a binomial distribution with $n = 50$ and $p = \frac{1}{25}$. Then

$$\Pr(Y \leq 1) = (\tfrac{24}{25})^{50} + 50(\tfrac{1}{25})(\tfrac{24}{25})^{49} = 0.400,$$

approximately. Since $\mu = np = 2$, the Poisson approximation to this probability is

$$e^{-2} + 2e^{-2} = 0.406.$$

Example 2. Let Z_n be $\chi^2(n)$. Then the moment-generating function of Z_n is $(1 - 2t)^{-n/2}$, $t < \frac{1}{2}$. The mean and the variance of Z_n are, respectively, n and $2n$. The limiting distribution of the random variable $Y_n = (Z_n - n)/\sqrt{2n}$ will be investigated. Now the moment-generating function of Y_n is

$$\begin{aligned} M(t; n) &= E\left\{\exp\left[t\left(\frac{Z_n - n}{\sqrt{2n}}\right)\right]\right\} \\ &= e^{-tn/\sqrt{2n}} E(e^{tZ_n/\sqrt{2n}}) \\ &= \exp\left[-\left(t\sqrt{\frac{2}{n}}\right)\left(\frac{n}{2}\right)\right]\left(1 - 2\frac{t}{\sqrt{2n}}\right)^{-n/2}, \qquad t < \frac{\sqrt{2n}}{2}. \end{aligned}$$

This may be written in the form

$$M(t; n) = \left(e^{t\sqrt{2/n}} - t\sqrt{\frac{2}{n}}\, e^{t\sqrt{2/n}}\right)^{-n/2}, \qquad t < \sqrt{\frac{n}{2}}.$$

In accordance with Taylor's formula, there exists a number $\xi(n)$, between 0 and $t\sqrt{2/n}$, such that

$$e^{t\sqrt{2/n}} = 1 + t\sqrt{\frac{2}{n}} + \frac{1}{2}\left(t\sqrt{\frac{2}{n}}\right)^2 + \frac{e^{\xi(n)}}{6}\left(t\sqrt{\frac{2}{n}}\right)^3.$$

If this sum is substituted for $e^{t\sqrt{2/n}}$ in the last expression for $M(t; n)$, it is seen that

$$M(t; n) = \left(1 - \frac{t^2}{n} + \frac{\psi(n)}{n}\right)^{-n/2},$$

where

$$\psi(n) = \frac{\sqrt{2}t^3 e^{\xi(n)}}{3\sqrt{n}} - \frac{\sqrt{2}t^3}{\sqrt{n}} - \frac{2t^4 e^{\xi(n)}}{3n}.$$

Since $\xi(n) \to 0$ as $n \to \infty$, then $\lim_{n\to\infty} \psi(n) = 0$ for every fixed value of t. In accordance with the limit proposition cited earlier in this section, we have

$$\lim_{n\to\infty} M(t; n) = e^{t^2/2}$$

for all real values of t. That is, the random variable $Y_n = (Z_n - n)/\sqrt{2n}$ has a limiting normal distribution with mean zero and variance one.

EXERCISES

5.12. Let X_n have a gamma distribution with parameter $\alpha = n$ and β, where β is not a function of n. Let $Y_n = X_n/n$. Find the limiting distribution of Y_n.

5.13. Let Z_n be $\chi^2(n)$ and let $W_n = Z_n/n^2$. Find the limiting distribution of W_n.

5.14. Let X be $\chi^2(50)$. Approximate Pr $(40 < X < 60)$.

5.15. Let $p = 0.95$ be the probability that a man, in a certain age group, lives at least five years. (a) If we are to observe 60 such men and if we assume independence, find the probability that at least 56 of them live five or more years. (b) Find an approximation to the result of part (a) by using the Poisson distribution. *Hint.* Redefine p to be 0.05 and $1 - p = 0.95$.

5.16. Let the random variable Z_n have a Poisson distribution with parameter $\mu = n$. Show that the limiting distribution of the random variable $Y_n = (Z_n - n)/\sqrt{n}$ is normal with mean zero and variance one.

5.17. Let S_n^2 denote the variance of a random sample of size n from a distribution which is $n(\mu, \sigma^2)$. It has been proved that $nS_n^2/(n-1)$ converges stochastically to σ^2. Prove that S_n^2 converges stochastically to σ^2.

5.18. Let X_n and Y_n have a bivariate normal distribution with parameters μ_1, μ_2, σ_1^2, σ_2^2 (free of n) but $\rho = 1 - 1/n$. Consider the conditional distribution of Y_n, given $X_n = x$. Investigate the limit of this conditional distribution as $n \to \infty$. What is the limiting distribution if $\rho = -1 + 1/n$? Reference to these facts was made in the Remark, Section 2.3 (p. 70).

5.19. Let $\bar{X}_n$ denote the mean of a random sample of size n from a Poisson distribution with parameter $\mu = 1$. (a) Show that the moment-generating function of $Y_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma = \sqrt{n}(\bar{X}_n - 1)$ is given by $\exp[-t\sqrt{n} + n(e^{t/\sqrt{n}} - 1)]$. (b) Investigate the limiting distribution of Y_n as $n \to \infty$.

5.20. Let $\bar{X}_n$ denote the mean of a random sample of size n from a distribution which has p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. (a) Show that the moment-generating function $M(t; n)$ of $Y_n = \sqrt{n}(\bar{X}_n - 1)$ is equal to $[e^{t/\sqrt{n}} - (t/\sqrt{n})e^{t/\sqrt{n}}]^{-n}$, $t < \sqrt{n}$. (b) Find the limiting distribution of Y_n as $n \to \infty$. This exercise and the immediately preceding one are special instances of an important theorem that will be proved in the next section.

5.4 The Central Limit Theorem

It was seen (p. 163) that, if $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2, the random variable

$$\frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

is, for every positive integer n, normally distributed with zero mean and unit variance. In probability theory there is a very elegant theorem called the *central limit theorem*. A special case of this theorem asserts the remarkable and important fact that if $X_1, X_2, \ldots, X_n$ denote the items of a random sample of size n from any distribution having positive variance σ^2 (and hence finite mean μ), then the random variable $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting normal distribution with zero mean and unit variance. If this fact can be established, it will imply, whenever the conditions of the theorem are satisfied, that (for fixed n) the random variable $\sqrt{n}(\bar{X} - \mu)/\sigma$ has an approximate normal distribution with mean zero and variance one. It will then be possible to use this approximate normal distribution to compute approximate probabilities concerning $\bar{X}$.

Only a modified form of this special case of the central limit theorem will be proved here. The modification consists in requiring the existence of a moment-generating function for the distribution from which the sample is to be taken. This is, of course, much more stringent than merely requiring a positive variance of that distribution. The modified form of the theorem follows.

Theorem 3. *Let $X_1, X_2, \ldots, X_n$ denote the items of a random sample from a distribution which has mean μ, positive variance σ^2, and moment-generating function $M(t)$ that exists for $-h < t < h$. Then the random variable $Y_n = \left(\sum_1^n X_i - n\mu\right)/\sqrt{n}\sigma = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting distribution which is normal with mean zero and variance one.*

Proof. Since $M(t) = E(e^{tX})$ exists for $-h < t < h$, the function

$$m(t) = E(e^{t(X-\mu)}) = e^{-\mu t}M(t)$$

exists for $-h < t < h$. It will first be shown that

$$m(t) = 1 + \frac{\sigma^2 t^2}{2} + \frac{[m''(\xi) - \sigma^2]t^2}{2},$$

where ξ is between 0 and t. To see this, note that

$$m'(t) = e^{-\mu t}M'(t) - \mu e^{-\mu t}M(t)$$

and

$$m''(t) = e^{-\mu t}M''(t) - 2\mu e^{-\mu t}M'(t) + \mu^2 e^{-\mu t}M(t).$$

Now $M(0) = 1$, $M'(0) = \mu$, and $M''(0) = \sigma^2 + \mu^2$. Hence $m(0) = 1$, $m'(0) = 0$, and $m''(0) = \sigma^2$. By Taylor's formula there exists a number ξ between 0 and t such that

$$\begin{aligned} m(t) &= m(0) + m'(0)t + \frac{m''(\xi)t^2}{2} \\ &= 1 + \frac{m''(\xi)t^2}{2}. \end{aligned}$$

If $\sigma^2 t^2/2$ is added and subtracted, then

$$m(t) = 1 + \frac{\sigma^2 t^2}{2} + \frac{[m''(\xi) - \sigma^2]t^2}{2},$$

as was to be shown.

Next consider $M(t; n)$, where

$$\begin{aligned} M(t; n) &= E\left[\exp\left(t\frac{\sum X_i - n\mu}{\sigma\sqrt{n}}\right)\right] \\ &= E\left[\exp\left(t\frac{X_1 - \mu}{\sigma\sqrt{n}}\right)\exp\left(t\frac{X_2 - \mu}{\sigma\sqrt{n}}\right)\cdots\exp\left(t\frac{X_n - \mu}{\sigma\sqrt{n}}\right)\right] \\ &= E\left[\exp\left(t\frac{X_1 - \mu}{\sigma\sqrt{n}}\right)\right]\cdots E\left[\exp\left(t\frac{X_n - \mu}{\sigma\sqrt{n}}\right)\right] \\ &= \left\{E\left[\exp\left(t\frac{X - \mu}{\sigma\sqrt{n}}\right)\right]\right\}^n \\ &= \left[m\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n, \qquad -h < \frac{t}{\sigma\sqrt{n}} < h. \end{aligned}$$

In $m(t)$, replace t by $t/\sigma\sqrt{n}$ to obtain

$$m\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + \frac{[m''(\xi) - \sigma^2]t^2}{2n\sigma^2},$$

where now ξ is between 0 and $t/\sigma\sqrt{n}$ with $-h\sigma\sqrt{n} < t < h\sigma\sqrt{n}$. Accordingly,

$$M(t; n) = \left\{1 + \frac{t^2}{2n} + \frac{[m''(\xi) - \sigma^2]t^2}{2n\sigma^2}\right\}^n.$$

Since $m''(t)$ is continuous at $t = 0$ and since $\xi \to 0$ as $n \to \infty$, we have

$$\lim_{n\to\infty} [m''(\xi) - \sigma^2] = 0.$$

The limit proposition cited on p. 179 shows that

$$\lim_{n\to\infty} M(t; n) = e^{t^2/2}$$

for all real values of t. This proves that the random variable $Y_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting normal distribution with mean zero and variance one.

We interpret this theorem as saying, with n a fixed positive integer, that the random variable $\sqrt{n}(\bar{X} - \mu)/\sigma$ has an approximate normal distribution with mean zero and variance one; and in applications we use the approximate normal p.d.f. as though it were the exact p.d.f. of $\sqrt{n}(\bar{X} - \mu)/\sigma$.

Some illustrative examples, here and later, will help show the importance of this version of the central limit theorem.

Example 1. Let $\bar{X}$ denote the mean of a random sample of size 75 from the distribution which has the p.d.f.

$$f(x) = 1, \qquad 0 < x < 1,$$
$$= 0 \text{ elsewhere.}$$

It was stated in Section 5.1 that the exact p.d.f. of $\bar{X}$, say $g(\bar{x})$, is rather complicated. It can be shown that $g(\bar{x})$ has a graph at points of positive probability density which is composed of arcs of 75 different polynomials of degree 74. The computation of such a probability as $\Pr(0.45 < \bar{X} < 0.55)$ would be extremely laborious. The conditions of the theorem are satisfied since $M(t)$ exists for all real values of t. Moreover, $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{12}$, so we have approximately

$$\Pr(0.45 < \bar{X} < 0.55) = \Pr\left[\frac{\sqrt{n}(0.45 - \mu)}{\sigma} < \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} < \frac{\sqrt{n}(0.55 - \mu)}{\sigma}\right]$$
$$= \Pr[-1.5 < 30(\bar{X} - 0.5) < 1.5]$$
$$= 0.866,$$

from Table III in the Appendix.

Example 2. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution which is $b(1, p)$. Here $\mu = p$, $\sigma^2 = p(1 - p)$, and $M(t)$ exists for all real values of t. If $Y_n = X_1 + \cdots + X_n$, it is known that Y_n is $b(n, p)$. Calculation of probabilities concerning Y_n, when we do not use the Poisson approximation, can be greatly simplified by making use of the fact that $(Y_n - np)/\sqrt{np(1-p)} = \sqrt{n}(\bar{X}_n - p)/\sqrt{p(1-p)} = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting distribution which is normal with mean zero and variance one. Let $n = 100$ and $p = \frac{1}{2}$, and suppose we wish to compute Pr $(Y = 48, 49, 50, 51, 52)$. Since Y is a random variable of the discrete type, the events $Y = 48, 49, 50, 51, 52$ and $47.5 < Y < 52.5$ are equivalent. That is, Pr $(Y = 48, 49, 50, 51, 52)$ = Pr $(47.5 < Y < 52.5)$. Since $np = 50$ and $np(1 - p) = 25$, the latter probability may be written

$$\Pr(47.5 < Y < 52.5) = \Pr\left(\frac{47.5 - 50}{5} < \frac{Y - 50}{5} < \frac{52.5 - 50}{5}\right)$$

$$= \Pr\left(-0.5 < \frac{Y - 50}{5} < 0.5\right).$$

Since $(Y - 50)/5$ has an approximate normal distribution with mean zero and variance one, Table III shows this probability to be approximately 0.382.

The convention of selecting the event $47.5 < Y < 52.5$, instead of, say, $47.8 < Y < 52.3$, as the event equivalent to the event $Y = 48, 49, 50, 51, 52$ seems to have originated in the following manner: The probability, Pr $(Y = 48, 49, 50, 51, 52)$, can be interpreted as the sum of five rectangular areas where the rectangles have bases one but the heights are, respectively, Pr $(Y = 48), \ldots,$ Pr $(Y = 52)$. If these rectangles are so located that the midpoints of their bases are, respectively, at the points $48, 49, \ldots, 52$ on a horizontal axis, then in approximating the sum of these areas by an area bounded by the horizontal axis, the graph of a normal p.d.f., and two ordinates, it seems reasonable to take the two ordinates at the points 47.5 and 52.5.

EXERCISES

5.21. Let $\bar{X}$ denote the mean of a random sample of size 100 from a distribution which is $\chi^2(50)$. Compute an approximate value of Pr $(49 < \bar{X} < 51)$.

5.22. Let $\bar{X}$ denote the mean of a random sample of size 128 from a gamma distribution with $\alpha = 2$ and $\beta = 4$. Approximate Pr $(7 < \bar{X} < 9)$.

5.23. Let Y be $b(72, \frac{1}{3})$. Approximate Pr $(22 \leq Y \leq 28)$.

5.24. Compute an approximate probability that the mean of a random sample of size 15 from a distribution having p.d.f. $f(x) = 3x^2$, $0 < x < 1$, zero elsewhere, is between $\frac{3}{5}$ and $\frac{4}{5}$.

5.25. Let Y denote the sum of the items of a random sample of size 12 from a distribution having p.d.f. $f(x) = \frac{1}{6}$, $x = 1, 2, 3, 4, 5, 6$, zero elsewhere. Compute an approximate value of $\Pr(36 \leq Y \leq 48)$.

5.26. Let Y be $b(400, \frac{1}{5})$. Compute an approximate value of $\Pr(0.25 < Y/n)$.

5.27. If Y is $b(100, \frac{1}{2})$, approximate the value of $\Pr(Y = 50)$.

5.28. Let Y be $b(n, 0.55)$. Find the smallest value of n so that (approximately) $\Pr(Y/n > \frac{1}{2}) \geq 0.95$.

5.29. Let $f(x) = 1/x^2$, $1 < x < \infty$, zero elsewhere, be the p.d.f. of a random variable X. Consider a random sample of size 72 from the distribution having this p.d.f. Compute approximately the probability that more than 50 of the items of the random sample are less than 3.

5.30. Forty-eight measurements are recorded to several decimal places. Each of these 48 numbers is rounded off to the nearest integer. The sum of the original 48 numbers is approximated by the sum of these integers. If we assume that the errors made by rounding off are stochastically independent and have uniform distributions over the interval $(-\frac{1}{2}, \frac{1}{2})$, compute approximately the probability that the sum of the integers is within 2 units of the true sum.

5.5 Some Theorems on Limiting Distributions

In this section we shall present some theorems that can often be used to simplify the study of certain limiting distributions.

Theorem 4. *Let $F_n(u)$ denote the distribution function of a random variable U_n whose distribution depends upon the positive integer n. Let U_n converge stochastically to the constant $c \neq 0$. The random variable U_n/c converges stochastically to one.*

The proof of this theorem is very easy and is left as an exercise.

Theorem 5. *Let $F_n(u)$ denote the distribution function of a random variable U_n whose distribution depends upon the positive integer n. Further, let U_n converge stochastically to the positive constant c and let $\Pr(U_n < 0) = 0$ for every n. The random variable $\sqrt{U_n}$ converges stochastically to $\sqrt{c}$.*

Proof. We are given that the $\lim_{n\to\infty} \Pr(|U_n - c| \geq \epsilon) = 0$ for every $\epsilon > 0$. We are to prove that the $\lim_{n\to\infty} \Pr(|\sqrt{U_n} - \sqrt{c}| \geq \epsilon') = 0$ for every $\epsilon' > 0$. Now the probability

$$\Pr(|U_n - c| \geq \epsilon) = \Pr[|(\sqrt{U_n} - \sqrt{c})(\sqrt{U_n} + \sqrt{c})| \geq \epsilon]$$

$$= \Pr\left(|\sqrt{U_n} - \sqrt{c}| \geq \frac{\epsilon}{\sqrt{U_n} + \sqrt{c}}\right)$$

$$\geq \Pr\left(|\sqrt{U_n} - \sqrt{c}| \geq \frac{\epsilon}{\sqrt{c}}\right) \geq 0.$$

If we let $\epsilon' = \epsilon/\sqrt{c}$, and if we take the limit, as n becomes infinite, we have

$$0 = \lim_{n \to \infty} \Pr(|U_n - c| \geq \epsilon) \geq \lim_{n \to \infty} \Pr(|\sqrt{U_n} - \sqrt{c}| \geq \epsilon') = 0$$

for every $\epsilon' > 0$. This completes the proof.

The conclusions of Theorems 4 and 5 are very natural ones and they certainly appeal to our intuition. There are many other theorems of this flavor in probability theory. As exercises, it is to be shown that if the random variables U_n and V_n converge stochastically to the respective constants c and d, then U_nV_n converges stochastically to the constant cd, and U_n/V_n converges stochastically to the constant c/d, provided $d \neq 0$. However, we will accept, without proof, the following theorem.

Theorem 6. *Let $F_n(u)$ denote the distribution function of a random variable U_n whose distribution depends upon the positive integer n. Let U_n have a limiting distribution with distribution function $F(u)$. Let $H_n(v)$ denote the distribution function of a random variable V_n whose distribution depends upon the positive integer n. Let V_n converge stochastically to one. The limiting distribution of the random variable $W_n = U_n/V_n$ is the same as that of U_n; that is, W_n has a limiting distribution with distribution function $F(w)$.*

Example 1. Let Y_n denote a random variable which is $b(n, p)$, $0 < p < 1$. We know that

$$U_n = \frac{Y_n - np}{\sqrt{np(1-p)}}$$

has a limiting distribution which is $n(0, 1)$. Moreover, it has been proved that Y_n/n and $1 - Y_n/n$ converge stochastically to p and $1 - p$, respectively; thus $(Y_n/n)(1 - Y_n/n)$ converges stochastically to $p(1 - p)$. Then, by Theorem 4, $(Y_n/n)(1 - Y_n/n)/[p(1 - p)]$ converges stochastically to one, and Theorem 5 asserts that the following does also:

$$V_n = \left[\frac{(Y_n/n)(1 - Y_n/n)}{p(1-p)}\right]^{1/2}.$$

Thus, in accordance with Theorem 6, the ratio $W_n = U_n/V_n$, namely,

$$\frac{Y_n - np}{\sqrt{n(Y_n/n)(1 - Y_n/n)}},$$

has a limiting distribution which is $n(0, 1)$. This fact enables us to write (with n a fixed positive integer)

$$\Pr\left[-2 < \frac{Y - np}{\sqrt{n(Y/n)(1 - Y/n)}} < 2\right] = 0.954,$$

approximately.

Example 2. Let $\bar{X}_n$ and S_n^2 denote, respectively, the mean and the variance of a random sample of size n from a distribution which is $n(\mu, \sigma^2)$, $\sigma^2 > 0$. It has been proved that $\bar{X}_n$ converges stochastically to μ and that S_n^2 converges stochastically to σ^2. Theorem 5 asserts that S_n converges stochastically to σ and Theorem 4 tells us that S_n/σ converges stochastically to one. In accordance with Theorem 6, the random variable $W_n = \sigma\bar{X}_n/S_n$ has the same limiting distribution as does $\bar{X}_n$. That is, $\sigma\bar{X}_n/S_n$ converges stochastically to μ.

EXERCISES

5.31. Prove Theorem 4. *Hint.* Note that $\Pr(|U_n/c - 1| < \epsilon) = \Pr(|U_n - c| < \epsilon|c|)$, for every $\epsilon > 0$. Then take $\epsilon' = \epsilon|c|$.

5.32. Let $\bar{X}_n$ denote the mean of a random sample of size n from a gamma distribution with parameters $\alpha = \mu > 0$ and $\beta = 1$. Show that the limiting distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sqrt{\bar{X}_n}$ is $n(0, 1)$.

5.33. Let $T_n = (\bar{X}_n - \mu)/\sqrt{S_n^2/(n-1)}$, where $\bar{X}_n$ and S_n^2 represent, respectively, the mean and the variance of a random sample of size n from a distribution which is $n(\mu, \sigma^2)$. Prove that the limiting distribution of T_n is $n(0, 1)$.

5.34. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be the items of two independent random samples, each of size n, from the distributions which have the respective means μ_1 and μ_2 and the common variance σ^2. Find the limiting distribution of

$$\frac{(\bar{X}_n - \bar{Y}_n) - (\mu_1 - \mu_2)}{\sigma\sqrt{2/n}},$$

where $\bar{X}_n$ and $\bar{Y}_n$ are the respective means of the samples. *Hint.* Let $\bar{Z}_n = \sum_1^n Z_i/n$, where $Z_i = X_i - Y_i$.

5.35. Let U_n and V_n converge stochastically to c and d, respectively. Prove the following. (a) The sum $U_n + V_n$ converges stochastically to $c + d$. *Hint.* Show that $\Pr(|U_n + V_n - c - d| \geq \epsilon) \leq \Pr(|U_n - c| + |V_n - d| \geq \epsilon) \leq \Pr(|U_n - c| \geq \epsilon/2 \text{ or } |V_n - d| \geq \epsilon/2) \leq \Pr(|U_n - c| \geq \epsilon/2) + \Pr(|V_n - d| \geq \epsilon/2)$. (b) The product U_nV_n converges stochastically to cd. (c) If $d \neq 0$, the ratio U_n/V_n converges stochastically to c/d.

Chapter 6

Interval Estimation

6.1 Random Intervals

An interval, at least one of whose end points is a random variable, will be called a *random interval*. Let X denote a random variable and consider the event $1 < X < 2$. This event is equivalent to $2 < 2X$ and $X < 2$ or to $X < 2 < 2X$. If p is the probability $\Pr(1 < X < 2)$, then $p = \Pr(X < 2 < 2X)$. The interval $(X, 2X)$ is a random interval, and we may read the probability $\Pr(X < 2 < 2X) = p$ as follows: The probability is p that the random interval $(X, 2X)$ includes the point 2. To be sure, the probability that the random interval $(X, 2X)$ includes the point 2 is precisely the probability that the random variable X satisfies $1 < X < 2$. We shall now compute some probabilities and discuss some problems that involve random intervals.

Example 1. Let X be $\chi^2(16)$. What is the probability that the random interval $(X, 3.3X)$ contains the point $x = 26.3$? And what is the expected length of the interval? Now $X < 26.3 < 3.3X$ when and only when $X < 26.3$ and $X > 26.3/3.3 = 7.97$, approximately; that is, when $7.97 < X < 26.3$. Accordingly, the probability that the random interval $(X, 3.3X)$ includes the point $x = 26.3$ is given by

$$\Pr(7.97 < X < 26.3),$$

where X is $\chi^2(16)$. We see from Table II in the Appendix that this probability is almost 0.90. The length of the random interval is $3.3X - X = 2.3X$. Since X is $\chi^2(16)$, we have $E(X) = 16$. Accordingly, the expected value of the length of the random interval is $2.3E(X) = 36.8$.

Example 2. Let X be $n(0, \sigma^2)$, $\sigma^2 > 0$. What is the probability that the random interval $(|X|, |10X|)$ includes the point σ? And what is the expected value of the length of this random interval? Now $|X| < \sigma < 10|X|$ when and only when $\sigma/10 < |X| < \sigma$. The event $\sigma/10 < |X| < \sigma$ is the union of the two mutually exclusive events $\sigma/10 < X < \sigma$ and $-\sigma < X < -\sigma/10$. Since X is $n(0, \sigma^2)$, these mutually exclusive events have the same probability. Accordingly, the probability that the random interval $(|X|, |10X|)$ includes the point σ is

$$\begin{aligned}\Pr\left(\frac{\sigma}{10} < |X| < \sigma\right) &= 2\Pr\left(\frac{1}{10} < \frac{X}{\sigma} < 1\right)\\ &= 2[N(1) - N(0.1)]\\ &= 0.60, \text{ approximately.}\end{aligned}$$

The length of the random interval is $9|X|$. Now

$$\begin{aligned}E(|X|) &= 2\int_0^\infty \frac{x}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}\, dx\\ &= \sigma\sqrt{\frac{2}{\pi}}.\end{aligned}$$

Hence the expected value of the length of the random interval is $9\sigma\sqrt{2/\pi} = 7.2\sigma$, approximately.

Example 3. Let $\bar{X}$ denote the mean of a random sample of size n from a distribution which is $n(\mu, \sigma^2)$, $\sigma^2 > 0$. Then $\bar{X}$ is $n(\mu, \sigma^2/n)$. What is the probability that the random interval $(\bar{X} - 2\sigma/\sqrt{n}, \bar{X} + 2\sigma/\sqrt{n})$ includes the point μ, whatever be the values of μ, $\sigma^2 > 0$, and the positive integer n? The event

$$\bar{X} - \frac{2\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{2\sigma}{\sqrt{n}}$$

occurs when and only when the event

$$-2 < \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} < 2$$

occurs. Thus these events have the same probability. But $\sqrt{n}(\bar{X} - \mu)/\sigma$ is $n(0, 1)$. Accordingly, the probability that the random interval $(\bar{X} - 2\sigma/\sqrt{n}, \bar{X} + 2\sigma/\sqrt{n})$ includes the point μ is $N(2) - N(-2) = 0.954$. It is seen that this probability in no manner depends upon the values of μ, $\sigma^2 > 0$, and the positive integer n. Consider next the length of the random interval. The length is $4\sigma/\sqrt{n}$, a constant function of $\bar{X}$; so the expected value of the length of the random interval is $4\sigma/\sqrt{n}$. Note that we do not know this length

(or expected length) until both σ and n are known; and note further that, for σ assigned, we can make this length as short as we please by taking n sufficiently large.

Example 4. Let $X_1, \ldots, X_{10}$ denote a random sample of size 10 from a distribution which is $n(\mu, \sigma^2)$. Let $Y = \sum_1^{10} (X_i - \mu)^2$. What is the probability that the random interval $(Y/20.5, Y/3.25)$ includes the point σ^2? We know that Y/σ^2 is $\chi^2(10)$. Moreover, the events $Y/20.5 < \sigma^2 < Y/3.25$ and $3.25 < Y/\sigma^2 < 20.5$ are equivalent. Accordingly, the probability that our random interval includes the point σ^2 is

$$\Pr\left(3.25 < \frac{Y}{\sigma^2} < 20.5\right),$$

where Y/σ^2 is $\chi^2(10)$. From Table II in the Appendix, this probability is found to be 0.95. The length of the random interval is $Y(1/3.25 - 1/20.5) = 0.26Y$, approximately. Since $E(Y/\sigma^2) = 10$, we have $E(Y) = 10\sigma^2$. Accordingly, the expected value of the length of the random interval is $2.6\sigma^2$.

EXERCISES

6.1. Let the random variable X have the p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Compute the probability that the random interval $(X, 3X)$ includes the point $x = 3$. What is the expected value of the length of this random interval?

6.2. Let X_1, X_2 denote a random sample from a distribution with p.d.f. $f(x) = \frac{1}{4}$, $x = 1, 2, 3, 4$, zero elsewhere. Compute the probability that the random interval $(1, X_1 + X_2 - \frac{1}{2})$ contains the point 3.

6.3. Let X_1, X_2 denote a random sample of size 2 from a continuous-type uniform distribution on the interval (0, 1). Compute the probability that the random interval $(X_1/3X_2, 2X_1/X_2)$ includes the point $\frac{2}{3}$.

6.4. Let $\overline{X}$ denote the mean of a random sample of size n from a distribution which is $n(\mu, \sigma^2)$. What is the probability that the random interval $[n(\overline{X} - \mu)^2/5.02, n(\overline{X} - \mu)^2/0.001]$ includes the point σ^2? What is the expected value of the length of this random interval? Note that this expected length is the same for all values of n. Compare your result with that of Example 4 of this section.

6.5. Let S^2 denote the variance of a random sample of size n from a distribution which is $n(\mu, \sigma^2)$, $\sigma^2 > 0$. Thus nS^2/σ^2 is $\chi^2(n - 1)$. If $n = 10$, what is the probability that the random interval $(S^2/1.9, S^2/0.27)$ includes the point σ^2? What is the expected value of the length of this random interval? Compare your result with those of Example 4 and Exercise 6.4.

6.2 Confidence Intervals for Means

The first five chapters of this book deal with certain concepts and problems of probability theory. Throughout we have carefully distinguished between a sample space $\mathscr{C}$ of outcomes and the space $\mathscr{A}$ of one or more random variables defined on $\mathscr{C}$. With this chapter we begin a study of some problems in statistics and here we are more interested in the number (or numbers) by which an outcome is represented than we are in the outcome itself. Accordingly, we shall adopt a frequently used convention: We shall refer to a random variable X as the outcome of a random experiment and we shall refer to the space of X as the sample space. Were it not so awkward, we would call X the numerical outcome. Once the experiment has been performed and it is found that $X = x$, we shall call x the experimental value of X for that performance of the experiment.

This convenient terminology can be used to advantage in more general situations. To illustrate this, let a random experiment be repeated n independent times and under identical conditions. Then the random variables $X_1, X_2, \ldots, X_n$ (each of which assigns a numerical value to an outcome) constitute (p. 118) the items of a random sample. If we are more concerned with the numerical representations of the outcomes than with the outcomes themselves, it seems natural to refer to $X_1, X_2, \ldots, X_n$ as the outcomes. And what more appropriate name can we give to the space of a random sample than the sample space? Once the experiment has been performed the indicated number of times and it is found that $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, we shall refer to $x_1, x_2, \ldots, x_n$ as the experimental values of $X_1, X_2, \ldots, X_n$ or as the sample data.

We shall use the terminology of the two preceding paragraphs, and in this section we shall give some examples of *statistical inference*. These examples will be built around the notion of a *confidence interval* for an unknown parameter in a p.d.f.

Suppose we are willing to accept as a fact that the (numerical) outcome X of a random experiment is a random variable that has a normal distribution with known variance σ^2 but unknown mean μ. That is, μ is some constant, but its value is unknown. To elicit some information about μ, we decide to repeat the random experiment n independent times, n being a fixed positive integer, and under identical conditions. Let the random variables $X_1, X_2, \ldots, X_n$ denote, respectively, the outcomes to be obtained on these n repetitions of the experiment. If our assumptions are fulfilled, we then have under consideration a random

sample $X_1, X_2, \ldots, X_n$ from a distribution which is $n(\mu, \sigma^2)$, σ^2 known. Consider the statistic $\bar{X} = \left(\sum_1^n X_i\right)/n$. In Example 3, p. 191, it was found that the probability is 0.954 that the random interval $(\bar{X} - 2\sigma/\sqrt{n}, \bar{X} + 2\sigma/\sqrt{n})$ contains the fixed (but unknown) mean μ. That is,

$$\Pr\left(\bar{X} - \frac{2\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.954.$$

Since σ is a known number, each of the random variables $\bar{X} - 2\sigma/\sqrt{n}$ and $\bar{X} + 2\sigma/\sqrt{n}$ is a statistic. The interval $(\bar{X} - 2\sigma/\sqrt{n}, \bar{X} + 2\sigma/\sqrt{n})$ is a random interval. In this case, both end points of the interval are statistics. The immediately preceding probability statement can be read: Prior to the repeated independent performances of the random experiment, the probability is 0.954 that the random interval $(\bar{X} - 2\sigma/\sqrt{n}, \bar{X} + 2\sigma/\sqrt{n})$ includes the unknown fixed point (parameter) μ.

Up to this point, only probability has been involved; the determination of the p.d.f. of $\bar{X}$ and the determination of the random interval were problems of probability. Now the problem becomes statistical. Suppose the experiment yields $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. Then the sample value of $\bar{X}$ is $\bar{x} = (x_1 + x_2 + \cdots + x_n)/n$, a known number. Moreover, since σ is known, the interval $(\bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n})$ has known end points. Obviously we cannot say that 0.954 is the probability that the particular interval $(\bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n})$ includes the parameter μ, for μ, although unknown, is some constant, and this particular interval either does or does not include μ. However, the fact that we had such a high degree of probability, prior to the performance of the experiment, that the random interval $(\bar{X} - 2\sigma/\sqrt{n}, \bar{X} + 2\sigma/\sqrt{n})$ includes the fixed point (parameter) μ leads us to have some reliance on the particular interval $(\bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n})$. This reliance is reflected by calling the known interval $(\bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n})$ a 95.4 per cent confidence interval for μ. The number 0.954 is called the *confidence coefficient*. The confidence coefficient is equal to the probability that the random interval includes the parameter. One may, of course, obtain a 90 per cent or a 99 per cent confidence interval for μ. A statistical inference of this sort is an example of *interval estimation* of a parameter.

If σ were not known, the end points of the random interval would not be statistics. Although the probability statement about the random

interval remains valid, the sample data would not yield an interval with known end points.

Example 1. If in the preceding discussion $n = 40$, $\sigma^2 = 10$, and $\bar{x} = 7.164$, then $(7.164 - 1.282\sqrt{\frac{10}{40}},\ 7.164 + 1.282\sqrt{\frac{10}{40}})$, or $(6.523, 7.805)$, is an 80 per cent confidence interval for μ. Thus, we have an interval estimate of μ.

In the next example, we shall show how the central limit theorem may be used to help us find an approximate confidence interval for μ when our sample arises from a distribution which is not normal.

Example 2. Let $\bar{X}$ denote the mean of a random sample of size 25 from a distribution which has a moment-generating function, variance $\sigma^2 = 100$, and mean μ. Since $\sigma/\sqrt{n} = 2$, then approximately

$$\Pr\left(-1.96 < \frac{\bar{X} - \mu}{2} < 1.96\right) = 0.95,$$

or

$$\Pr(\bar{X} - 3.92 < \mu < \bar{X} + 3.92) = 0.95.$$

Let the observed mean of the sample be $\bar{x} = 67.53$. Accordingly, the interval from $\bar{x} - 3.92 = 63.61$ to $\bar{x} + 3.92 = 71.45$ is an approximate 95 per cent confidence interval for the mean μ.

Let us now turn to the problem of finding a confidence interval for the mean μ of a normal distribution when we are not so fortunate as to know the variance σ^2. In Section 4.8 we found that nS^2/σ^2, where S^2 is the variance of a random sample of size n from a distribution which is $n(\mu, \sigma^2)$, is $\chi^2(n - 1)$. Thus we have $\sqrt{n}(\bar{X} - \mu)/\sigma$ to be $n(0, 1)$, nS^2/σ^2 to be $\chi^2(n - 1)$, and the two to be stochastically independent. In Section 4.4, the random variable T was defined in terms of two such random variables as these. In accordance with that section and the foregoing results, we know that

$$T = \frac{[\sqrt{n}(\bar{X} - \mu)]/\sigma}{\sqrt{nS^2/[\sigma^2(n - 1)]}} = \frac{\bar{X} - \mu}{S/\sqrt{n - 1}}$$

has a t distribution with $n - 1$ degrees of freedom, whatever the value of $\sigma^2 > 0$. For a given positive integer n and a probability of 0.95, say, we can find numbers $a < b$ from Table IV in the Appendix, such that

$$\Pr\left[a < \frac{\bar{X} - \mu}{S/\sqrt{n - 1}} < b\right] = 0.95.$$

Since the graph of the p.d.f. of the random variable T is symmetric about the vertical axis through the origin, we would doubtless take

$a = -b$, $b > 0$. If the probability of this event is written (with $a = -b$) in the form

$$\Pr\left(\bar{X} - \frac{bS}{\sqrt{n-1}} < \mu < \bar{X} + \frac{bS}{\sqrt{n-1}}\right) = 0.95,$$

then the interval $[\bar{X} - (bS/\sqrt{n-1}), \bar{X} + (bS/\sqrt{n-1})]$ is a random interval having probability 0.95 of including the unknown fixed point (parameter) μ. If the experimental values of $X_1, X_2, \ldots, X_n$ are $x_1, x_2, \ldots, x_n$ with $s^2 = \sum_1^n (x_i - \bar{x})^2/n$, where $\bar{x} = \sum_1^n x_i/n$, then the interval $[\bar{x} - (bs/\sqrt{n-1}), \bar{x} + (bs/\sqrt{n-1})]$ is a 95 per cent confidence interval for μ for every $\sigma^2 > 0$.

Example 3. If in the preceding discussion $n = 10$, $\bar{x} = 3.22$, and $s = 1.17$, then the interval $[3.22 - (2.262)(1.17)/\sqrt{9},\ 3.22 + (2.262)(1.17)/\sqrt{9}]$ or (2.34, 4.10) is a 95 per cent confidence interval for μ.

Remark. If one wishes to find a confidence interval for μ and if the variance σ^2 of the nonnormal distribution is unknown (unlike Example 2 of this section), he may with large samples proceed as follows. If the conditions of Exercise 5.11 are satisfied, then S^2, the variance of a random sample of size $n \geq 2$, converges stochastically to σ^2 Then in

$$\frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{nS^2/(n-1)\sigma^2}} = \frac{\sqrt{n-1}(\bar{X} - \mu)}{S}$$

the numerator of the left-hand member has a limiting distribution which is $n(0, 1)$ and the denominator of that member converges stochastically to one. Thus, $\sqrt{n-1}(\bar{X} - \mu)/S$ has a limiting distribution which is $n(0, 1)$. This fact enables us to find approximate confidence intervals for μ when our conditions are satisfied. A similar procedure can be followed in the next section when seeking confidence intervals for the difference of the means of two independent nonnormal distributions.

We shall now consider the problem of determining a confidence interval for the unknown parameter p of a binomial distribution when the parameter n is known. Let Y be $b(n, p)$, where $0 < p < 1$ and n is known. Then p is the mean of Y/n. We shall use a result of Example 1, p. 187, to find an approximate 95.4 per cent confidence interval for the mean p. There we found that

$$\Pr\left[-2 < \frac{Y - np}{\sqrt{n(Y/n)(1 - Y/n)}} < 2\right] = 0.954,$$

approximately. Since

$$\frac{Y - np}{\sqrt{n(Y/n)(1 - Y/n)}} = \frac{(Y/n) - p}{\sqrt{(Y/n)(1 - Y/n)/n}},$$

the above probability statement can easily be written in the form

$$\Pr\left[\frac{Y}{n} - 2\sqrt{\frac{(Y/n)(1 - Y/n)}{n}} < p < \frac{Y}{n} + 2\sqrt{\frac{(Y/n)(1 - Y/n)}{n}}\right] = 0.954,$$

approximately. Thus, for large n, if the experimental value of Y is y, the interval

$$\left(\frac{y}{n} - 2\sqrt{\frac{(y/n)(1 - y/n)}{n}}, \frac{y}{n} + 2\sqrt{\frac{(y/n)(1 - y/n)}{n}}\right)$$

provides an approximate 95.4 per cent confidence interval for p.

A more complicated approximate 95.4 per cent confidence interval can be obtained from the fact that $Z = (Y - np)/\sqrt{np(1 - p)}$ has a limiting distribution which is $n(0, 1)$, and the fact that the event $-2 < Z < 2$ is equivalent to the event

$$(1) \quad \frac{Y + 2 - 2\sqrt{[Y(n - Y)/n] + 1}}{n + 4} < p < \frac{Y + 2 + 2\sqrt{[Y(n - Y)/n] + 1}}{n + 4}.$$

The first of these facts was established in Example 2, p. 185; the proof of Inequalities (1) is left as an exercise. Thus an experimental value y of Y may be used in Inequalities (1) to determine an approximate 95.4 per cent confidence interval for p.

If one wishes a 95 per cent confidence interval for p which does not depend upon limiting distribution theory, he may use the following approach. (This approach is quite general and can be used in other instances.) Determine two *increasing* functions of p, say $c_1(p)$ and $c_2(p)$, such that for each value of p we have, at least approximately,

$$\Pr[c_1(p) < Y < c_2(p)] = 0.95.$$

The reason that this may be approximate is due to the fact that Y has a distribution of the discrete type and thus it is, in general, impossible to achieve the probability 0.95 exactly. With $c_1(p)$ and $c_2(p)$ increasing functions, they have single-valued inverses, say $d_1(y)$ and $d_2(y)$, respectively. Thus the events $c_1(p) < Y < c_2(p)$ and $d_2(Y) < p < d_1(Y)$ are equivalent and we have, at least approximately,

$$\Pr[d_2(Y) < p < d_1(Y)] = 0.95.$$

In the case of the binomial distribution, the functions $c_1(p)$, $c_2(p)$, $d_2(y)$, and $d_1(y)$ cannot be found explicitly, but a number of books provide tables of $d_2(y)$ and $d_1(y)$ for various values of n.

Example 4. If, in the preceding discussion, we take $n = 100$ and $y = 20$, the first approximate 95.4 per cent confidence interval is given by $(0.2 - 2\sqrt{(0.2)(0.8)/100}, 0.2 + 2\sqrt{(0.2)(0.8)/100})$ or (0.12, 0.28). The approximate 95.4 per cent confidence interval which is provided by Inequalities (1) is

$$\left(\frac{22 - 2\sqrt{(1600/100) + 1}}{104}, \frac{22 + 2\sqrt{(1600/100) + 1}}{104}\right)$$

or (0.13, 0.29). By referring to the appropriate tables found elsewhere, we find that an approximate 95 per cent confidence interval has the limits $d_2(20) = 0.13$ and $d_1(20) = 0.29$. Thus, in this example, we see that all three methods yield results that are in substantial agreement.

EXERCISES

6.6. Let the observed value of the mean $\bar{X}$ of a random sample of size 20 from a distribution which is $n(\mu, 80)$ be 81.2. Find a 95 per cent confidence interval for μ.

6.7. Let $\bar{X}$ be the mean of a random sample of size n from a distribution which is $n(\mu, 9)$. Find n such that $\Pr(\bar{X} - 1 < \mu < \bar{X} + 1) = 0.90$, approximately.

6.8. Let a random sample of size 17 from the normal distribution $n(\mu, \sigma^2)$ yield $\bar{x} = 4\,7$ and $s^2 = 5.76$. Determine a 90 per cent confidence interval for μ.

6.9. Let $\bar{X}$ denote the mean of a random sample of size n from a distribution which has mean μ, variance $\sigma^2 = 10$, and a moment-generating function. Find n so that the probability is approximately 0.954 that the random interval $(\bar{X} - \frac{1}{2}, \bar{X} + \frac{1}{2})$ includes μ.

6.10. Let $X_1, X_2, \ldots, X_9$ be a random sample of size 9 from a distribution which is $n(\mu, \sigma^2)$. (a) If σ is known, find the length of a 95 per cent confidence interval for μ if this interval is based on the random variable $\sqrt{9}(\bar{X} - \mu)/\sigma$. (b) If σ is unknown, find the expected value of the length of a 95 per cent confidence interval for μ if this interval is based on the random variable $\sqrt{8}(\bar{X} - \mu)/S$. (c) Compare these two answers.

6.11. Let $X_1, X_2, \ldots, X_n, X_{n+1}$ be a random sample of size $n + 1$, $n > 1$, from a distribution which is $n(\mu, \sigma^2)$. Let $\bar{X} = \sum_1^n X_i/n$, and $S^2 = \sum_1^n (X_i - \bar{X})^2/n$. Find the constant c so that the statistic $c(\bar{X} - X_{n+1})/S$

has a t distribution. If $n = 8$, determine k such that $\Pr(\bar{X} - kS < X_9 < \bar{X} + kS) = 0.80$.

6.12. Let Y be $b(300, p)$. If the observed value of Y is $y = 75$, find an approximate 90 per cent confidence interval for p.

6.13. Let $\bar{X}$ be the mean of a random sample of size n from a distribution which is $n(\mu, \sigma^2)$, where the positive variance σ^2 is known. Use the fact that $N(2) - N(-2) = 0.954$ to find, for each μ, $c_1(\mu)$ and $c_2(\mu)$ such that $\Pr[c_1(\mu) < \bar{X} < c_2(\mu)] = 0.954$. Note that $c_1(\mu)$ and $c_2(\mu)$ are increasing functions of μ. Solve for the respective functions $d_1(\bar{x})$ and $d_2(\bar{x})$; thus we also have that $\Pr[d_2(\bar{X}) < \mu < d_1(\bar{X})] = 0.954$. Compare this with the answer obtained previously in the text.

6.14. In the notation of the discussion of the confidence interval for p, show that the event $-2 < Z < 2$ is equivalent to Inequalities (1). *Hint.* First observe that $-2 < Z < 2$ is equivalent to $Z^2 < 4$, which can be written as an inequality involving a quadratic expression in p.

6.15. Let $\bar{X}$ denote the mean of a random sample of size 25 from a gamma-type distribution with $\alpha = 4$ and $\beta > 0$. Use the central limit theorem to find an approximate 0.954 confidence interval for μ, the mean of the gamma distribution.

6.3 Confidence Intervals for Differences of Means

The random variable T may also be used to obtain a confidence interval for the difference $\mu_1 - \mu_2$ between the means of two independent normal distributions, say $n(\mu_1, \sigma^2)$ and $n(\mu_2, \sigma^2)$, when the distributions have the same, but unknown, variance σ^2.

Remark. Let X have a normal distribution with unknown parameters μ_1 and σ^2. A modification can be made in conducting the experiment so that the variance of the distribution will remain the same but the mean of the distribution will be changed; say, increased. After the modification has been effected, let the random variable be denoted by Y, and let Y have a normal distribution with unknown parameters μ_2 and σ^2. Naturally it is hoped that μ_2 is greater than μ_1, that is, that $\mu_1 - \mu_2 < 0$. Accordingly, one seeks a confidence interval for $\mu_1 - \mu_2$ in order to make a statistical inference.

A confidence interval for $\mu_1 - \mu_2$ may be obtained as follows: Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ denote, respectively, independent random samples from the two independent distributions having, respectively, the probability density functions $n(\mu_1, \sigma^2)$ and $n(\mu_2, \sigma^2)$. Denote the means of the samples by $\bar{X}$ and $\bar{Y}$ and the variances of the samples by S_1^2 and S_2^2, respectively. It should be noted that these four

statistics are mutually stochastically independent. The stochastic independence of $\bar{X}$ and S_1^2 (and, inferentially that of $\bar{Y}$ and S_2^2) was established in Section 4.8; the assumption that X and Y have independent distributions accounts for the stochastic independence of the others. Thus $\bar{X}$ and $\bar{Y}$ are normally and stochastically independently distributed with means μ_1 and μ_2 and variances σ^2/n and σ^2/m, respectively. In accordance with p. 158, their difference $\bar{X} - \bar{Y}$ is normally distributed with mean $\mu_1 - \mu_2$ and variance $\sigma^2/n + \sigma^2/m$. Then the random variable

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2/n + \sigma^2/m}}$$

is normally distributed with zero mean and unit variance. This random variable may serve as the numerator of a T random variable. Further, nS_1^2/σ^2 and mS_2^2/σ^2 have stochastically independent chi-square distributions with $n - 1$ and $m - 1$ degrees of freedom, respectively, so that their sum $(nS_1^2 + mS_2^2)/\sigma^2$ has a chi-square distribution with $n + m - 2$ degrees of freedom, provided $m + n - 2 > 0$. Because of the mutual stochastic independence of $\bar{X}$, $\bar{Y}$, S_1^2, and S_2^2, it is seen that

$$\sqrt{\frac{nS_1^2 + mS_2^2}{\sigma^2(n + m - 2)}}$$

may serve as the denominator of a T random variable. That is, the random variable

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{nS_1^2 + mS_2^2}{n + m - 2}\left(\dfrac{1}{n} + \dfrac{1}{m}\right)}}$$

has a t distribution with $n + m - 2$ degrees of freedom. As in the previous section, we can (once n and m are specified positive integers with $n + m - 2 > 0$) find a positive number b from Table IV such that

$$\Pr(-b < T < b) = 0.95.$$

If we set

$$R = \sqrt{\frac{nS_1^2 + mS_2^2}{n + m - 2}\left(\frac{1}{n} + \frac{1}{m}\right)},$$

this probability may be written in the form

$$\Pr[(\bar{X} - \bar{Y}) - bR < \mu_1 - \mu_2 < (\bar{X} - \bar{Y}) + bR] = 0.95.$$

It follows that the random interval

$$\left[(\bar{X} - \bar{Y}) - b\sqrt{\frac{nS_1^2 + mS_2^2}{n + m - 2}\left(\frac{1}{n} + \frac{1}{m}\right)},\right.$$
$$\left.(\bar{X} - \bar{Y}) + b\sqrt{\frac{nS_1^2 + mS_2^2}{n + m - 2}\left(\frac{1}{n} + \frac{1}{m}\right)}\right]$$

has probability 0.95 of including the unknown fixed point $(\mu_1 - \mu_2)$. As usual, the experimental values of $\bar{X}$, $\bar{Y}$, S_1^2, and S_2^2, namely, $\bar{x}$, $\bar{y}$, s_1^2, and s_2^2, will provide a 95 per cent confidence interval for $\mu_1 - \mu_2$ when the variances of the two independent normal distributions are unknown but equal. A consideration of the difficulty encountered when the unknown variances of the two independent normal distributions are not equal is assigned to one of the exercises.

Example 1. It may be verified that if in the preceding discussion $n = 10$, $m = 7$, $\bar{x} = 4.2$, $\bar{y} = 3.4$, $s_1^2 = 49$, $s_2^2 = 32$, then the interval $(-5.16, 6.76)$ is a 90 per cent confidence interval for $\mu_1 - \mu_2$.

Let Y_1 and Y_2 be two stochastically independent random variables with binomial distributions $b(n_1, p_1)$ and $b(n_2, p_2)$, respectively. Let us now turn to the problem of finding a confidence interval for the difference $p_1 - p_2$ of the means of Y_1/n_1 and Y_2/n_2 when n_1 and n_2 are known. Since the mean and the variance of $Y_1/n_1 - Y_2/n_2$ are, respectively, $p_1 - p_2$ and $p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2$, then the random variable given by the ratio

$$\frac{(Y_1/n_1 - Y_2/n_2) - (p_1 - p_2)}{\sqrt{p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2}}$$

has mean zero and variance one for all positive integers n_1 and n_2. Moreover, since both Y_1 and Y_2 have approximate normal distributions for large n_1 and n_2, one suspects that the ratio has an approximate normal distribution. This is actually the case but it will not be proved here. Moreover, if $n_1/n_2 = c$, where c is a fixed positive constant, the result of Exercise 6.20 shows that the random variable

$$(1) \qquad \frac{(Y_1/n_1)(1 - Y_1/n_1)/n_1 + (Y_2/n_2)(1 - Y_2/n_2)/n_2}{p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2}$$

converges stochastically to one as $n_2 \to \infty$ (and thus $n_1 \to \infty$ since $n_1/n_2 = c$, $c > 0$). In accordance with Theorem 6, p. 187, the random variable

$$W = [(Y_1/n_1 - Y_2/n_2) - (p_1 - p_2)]/U,$$

where

$$U = \sqrt{(Y_1/n_1)(1 - Y_1/n_1)/n_1 + (Y_2/n_2)(1 - Y_2/n_2)/n_2},$$

has a limiting distribution which is $n(0, 1)$. The event $-2 < W < 2$, the probability of which is approximately equal to 0.954, is equivalent to the event

$$Y_1/n_1 - Y_2/n_2 - 2U < p_1 - p_2 < Y_1/n_1 - Y_2/n_2 + 2U.$$

Accordingly, the experimental values y_1 and y_2 of Y_1 and Y_2, respectively, will provide an approximate 95.4 per cent confidence interval for $p_1 - p_2$.

Example 2. If, in the preceding discussion, we take $n_1 = 100$, $n_2 = 400$, $y_1 = 30$, $y_2 = 80$, then the experimental values of $Y_1/n_1 - Y_2/n_2$ and U are 0.1 and $\sqrt{(0.3)(0.7)/100 + (0.2)(0.8)/400} = 0.05$, respectively. Thus the interval (0, 0.2) is an approximate 95.4 per cent confidence interval for $p_1 - p_2$.

EXERCISES

6.16. Let two independent random samples, each of size 10, from two independent normal distributions $n(\mu_1, \sigma^2)$ and $n(\mu_2, \sigma^2)$ yield $\bar{x} = 4.8$, $s_1^2 = 8.64$, $\bar{y} = 5.6$, $s_2^2 = 7.88$. Find a 95 per cent confidence interval for $\mu_1 - \mu_2$.

6.17. Discuss the problem of finding a confidence interval for the difference $\mu_1 - \mu_2$ between the two means of two independent normal distributions if the variances σ_1^2 and σ_2^2 are known but not necessarily equal.

6.18. Discuss Exercise 6.17 when it is assumed that the variances are unknown and unequal. This is a very difficult problem, and the discussion should point out exactly where the difficulty lies. If, however, the variances are unknown but their ratio σ_1^2/σ_2^2 is a known constant k, then a statistic which is a T random variable can again be used. Why?

6.19. Let $\overline{X}$ and $\overline{Y}$ be the means of two independent random samples, each of size n, from the respective distributions $n(\mu_1, \sigma^2)$ and $n(\mu_2, \sigma^2)$, where the common variance is known. Find n such that $\Pr(\overline{X} - \overline{Y} - \sigma/5 < \mu_1 - \mu_2 < \overline{X} - \overline{Y} + \sigma/5) = 0.90$.

6.20. Under the conditions given, show that the random variable defined by Ratio (1) of the text converges stochastically to one.

6.4 Confidence Intervals for Variances

Let the random variable X be $n(\mu, \sigma^2)$. We shall discuss the problem of finding a confidence interval for σ^2. Our discussion will consist of two parts. First, when μ is a known number, and second, when μ is unknown.

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution which is $n(\mu, \sigma^2)$, where μ is a known number. The random variable $Y = \sum_1^n (X_i - \mu)^2/\sigma^2$ is $\chi^2(n)$. Let us select a probability, say 0.95, and for the fixed positive integer n determine values of a and b, $a < b$, from Table II, such that

$$\Pr(a < Y < b) = 0.95.$$

Thus,

$$\Pr\left[a < \frac{\sum_1^n (X_i - \mu)^2}{\sigma^2} < b\right] = 0.95,$$

or

$$\Pr\left[\frac{\sum_1^n (X_i - \mu)^2}{b} < \sigma^2 < \frac{\sum_1^n (X_i - \mu)^2}{a}\right] = 0.95.$$

Since μ, a, and b are known constants, each of $\sum_1^n (X_i - \mu)^2/b$ and $\sum_1^n (X_i - \mu)^2/a$ is a statistic. Moreover, the interval

$$\left[\frac{\sum_1^n (X_i - \mu)^2}{b}, \frac{\sum_1^n (X_i - \mu)^2}{a}\right]$$

is a random interval having probability of 0.95 that it includes the unknown fixed point (parameter) σ^2. Once the random experiment has been performed, and it is found that $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, then the particular interval

$$\left[\frac{\sum_1^n (x_i - \mu)^2}{b}, \frac{\sum_1^n (x_i - \mu)^2}{a}\right]$$

is a 95 per cent confidence interval for σ^2.

The student will immediately observe that there are no unique numbers $a < b$ such that $\Pr(a < Y < b) = 0.95$. A common method of procedure is to find a and b such that $\Pr(Y < a) = 0.025$ and $\Pr(b < Y) = 0.025$. That procedure will be followed in this book.

Example 1. If in the preceding discussion $\mu = 0$, $n = 10$, and $\sum_1^{10} x_i^2 = 106.6$, then the interval (106.6/20.5, 106.6/3.25), or (5.2, 32.8), is a 95 per cent confidence interval for the variance σ^2, since $\Pr(Y < 3.25) = 0.025$ and

$\Pr(20.5 < Y) = 0.025$, provided Y has a chi-square distribution with 10 degrees of freedom.

We now turn to the case in which μ is not known. This case can be handled by making use of the fact that nS^2/σ^2 is $\chi^2(n-1)$. For a fixed positive integer $n \geq 2$, we can find, from Table II, values of a and b, $a < b$, such that

$$\Pr\left(a < \frac{nS^2}{\sigma^2} < b\right) = 0.95.$$

Here, of course, we would find a and b by using a chi-square distribution with $n-1$ degrees of freedom. In accordance with the convention previously adopted, we would select a and b so that

$$\Pr\left(\frac{nS^2}{\sigma^2} < a\right) = 0.025 \qquad \text{and} \qquad \Pr\left(\frac{nS^2}{\sigma^2} > b\right) = 0.025.$$

We then have

$$\Pr\left(\frac{nS^2}{b} < \sigma^2 < \frac{nS^2}{a}\right) = 0.95,$$

so that $(nS^2/b, nS^2/a)$ is a random interval having probability 0.95 of including the fixed but unknown point (parameter) σ^2. After the random experiment has been performed and we find, say, $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, with $s^2 = \sum_1^n (x_i - \bar{x})^2/n$, we have, as a 95 per cent confidence interval for σ^2, the interval $(ns^2/b, ns^2/a)$.

Example 2. If, in the preceding discussion, we have $n = 9$, $s^2 = 7.63$, then the interval $[9(7.63)/17.5, 9(7.63)/2.18]$ or $(3.92, 31.50)$ is a 95 per cent confidence interval for the variance σ^2.

Next let X and Y denote stochastically independent random variables that are $n(\mu_1, \sigma_1^2)$ and $n(\mu_2, \sigma_2^2)$, respectively. We shall determine a confidence interval for the ratio σ_2^2/σ_1^2 when μ_1 and μ_2 are unknown.

Remark. Consider a situation in which a random variable X has a normal distribution with variance σ_1^2. Although σ_1^2 is not known, it is found that the experimental values of X are quite widely dispersed, so that σ_1^2 must be fairly large. It is believed that a certain modification in conducting the experiment may reduce the variance. After the modification has been effected, let the random variable be denoted by Y, and let Y have a normal distribution with variance σ_2^2. Naturally it is hoped that σ_2^2 is less than σ_1^2, that is, that $\sigma_2^2/\sigma_1^2 < 1$. In order to make a statistical inference, we find a confidence interval for the ratio σ_2^2/σ_1^2.

Consider a random sample $X_1, X_2, \ldots, X_n$ of size $n \geq 2$ from the distribution of X and a random sample $Y_1, Y_2, \ldots, Y_m$ of size $m \geq 2$ from the independent distribution of Y. Here n and m may or may not be equal. Let the means of the two samples be denoted by $\bar{X}$ and $\bar{Y}$, and the variances of the two samples by $S_1^2 = \sum_1^n (X_i - \bar{X})^2/n$ and $S_2^2 = \sum_1^m (Y_i - \bar{Y})^2/m$, respectively. The stochastically independent random variables nS_1^2/σ_1^2 and mS_2^2/σ_2^2 have chi-square distributions with $n - 1$ and $m - 1$ degrees of freedom, respectively. On p. 137 a random variable called F was defined, and through the change of variable technique the p.d.f. of F was obtained. If nS_1^2/σ_1^2 is divided by $n - 1$, the number of degrees of freedom, and if mS_2^2/σ_2^2 is divided by $m - 1$, then, by definition of an F random variable, we have that

$$F = \frac{nS_1^2/[\sigma_1^2(n-1)]}{mS_2^2/[\sigma_2^2(m-1)]}$$

has an F distribution with parameters $n - 1$ and $m - 1$. For numerically given values of n and m and with a preassigned probability, say 0.95, we can determine from Table V, in accordance with our convention, numbers $0 < a < b$ such that

$$\Pr\left[a < \frac{nS_1^2/[\sigma_1^2(n-1)]}{mS_2^2/[\sigma_2^2(m-1)]} < b\right] = 0.95.$$

If the probability of this event is written in the form

$$\Pr\left[a\,\frac{mS_2^2/(m-1)}{nS_1^2/(n-1)} < \frac{\sigma_2^2}{\sigma_1^2} < b\,\frac{mS_2^2/(m-1)}{nS_1^2/(n-1)}\right] = 0.95,$$

it is seen that the interval

$$\left[a\,\frac{mS_2^2/(m-1)}{nS_1^2/(n-1)}, b\,\frac{mS_2^2/(m-1)}{nS_1^2/(n-1)}\right]$$

is a random interval having probability 0.95 of including the fixed but unknown point σ_2^2/σ_1^2. If the experimental values of $X_1, X_2, \ldots, X_n$ and of $Y_1, Y_2, \ldots, Y_m$ are denoted by $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$, respectively, and if $ns_1^2 = \sum_1^n (x_i - \bar{x})^2$, $ms_2^2 = \sum_1^m (y_i - \bar{y})^2$, then the interval with known end points, namely,

$$\left[a\,\frac{ms_2^2/(m-1)}{ns_1^2/(n-1)}, b\,\frac{ms_2^2/(m-1)}{ns_1^2/(n-1)}\right],$$

is a 95 per cent confidence interval for the ratio σ_2^2/σ_1^2 of the two unknown variances.

Example 3. If in the preceding discussion $n = 10$, $m = 5$, $s_1^2 = 20.0$, $s_2^2 = 35.6$, then the interval

$$\left[\left(\frac{1}{4.72}\right)\frac{5(35.6)/4}{10(20.0)/9}, (8.90)\frac{5(35.6)/4}{10(20.0)/9}\right]$$

or (0.4, 17.8) is a 95 per cent confidence interval for σ_2^2/σ_1^2.

EXERCISES

6.21. If 8.6, 7.9, 8.3, 6.4, 8.4, 9.8, 7.2, 7.8, 7.5 are the observed values of a random sample of size 9 from a distribution that is $n(8, \sigma^2)$, construct a 90 per cent confidence interval for σ^2.

6.22. Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution $n(\mu, \sigma^2)$. Let $0 < a < b$. Show that the mathematical expectation of the length of the random interval $\left(\sum_1^n (X_i - \mu)^2/b, \sum_1^n (X_i - \mu)^2/a\right)$ is $(b - a) \times (n\sigma^2/ab)$.

6.23. A random sample of size 15 from the normal distribution $n(\mu, \sigma^2)$ yields $\bar{x} = 3.2$ and $s^2 = 4.24$. Determine a 90 per cent confidence interval for σ^2.

6.24. Let two independent random samples of sizes $n = 16$ and $m = 10$, taken from two independent normal distributions $n(\mu_1, \sigma_1^2)$ and $n(\mu_2, \sigma_2^2)$, respectively, yield $\bar{x} = 3.6$, $s_1^2 = 4.14$, $\bar{y} = 13.6$, $s_2^2 = 7.26$. Find a 90 per cent confidence interval for σ_2^2/σ_1^2 when μ_1 and μ_2 are unknown.

6.25. Discuss the problem of finding a confidence interval for the ratio σ_2^2/σ_1^2 of the two unknown variances of two independent normal distributions if the means μ_1 and μ_2 are known.

6.26. Let $X_1, X_2, \ldots, X_6$ be a random sample of size 6 from a gamma distribution with parameters $\alpha = 1$ and unknown $\beta > 0$. Discuss the construction of a 98 per cent confidence interval for β. *Hint.* What is the distribution of $2\sum_1^6 X_i/\beta$?

6.27. Let S_1^2 and S_2^2 denote, respectively, the variances of random samples, of sizes n and m, from two independent distributions which are $n(\mu_1, \sigma^2)$ and $n(\mu_2, \sigma^2)$. Use the fact that $(nS_1^2 + mS_2^2)/\sigma^2$ is $\chi^2(n + m - 2)$ to find a confidence interval for the common unknown variance σ^2.

6.28. Let Y_4 be the nth order statistic of a random sample, n = 4, from a continuous-type uniform distribution on the interval $(0, \theta)$. Let $0 < c_1 < c_2 \leq 1$ be selected so that $\Pr(c_1\theta < Y_4 < c_2\theta) = 0.95$. Verify that $c_1 = \sqrt[4]{0.05}$ and $c_2 = 1$ satisfy these conditions. What then is a 95 per cent confidence interval for θ?

6.5 Bayesian Interval Estimates

In Sections 6.2, 6.3, and 6.4 we constructed two statistics, say U and V, $U < V$, such that we have a preassigned probability p that the random interval (U, V) contains a fixed but unknown point (parameter). We then *adopted* this *principle*: Use the experimental results to compute the values of U and V, say u and v; then call the interval (u, v) a $100p$ per cent confidence interval for the parameter. Adoption of this principle provided us with one method of interval estimation. This method of interval estimation is widely used in statistical literature and in the applications. But it is important for us to understand that other principles can be adopted. The student should constantly keep in mind that as long as he is working with probability, he is in the realm of mathematics; but once he begins to make inferences or to draw conclusions about a random experiment, which inferences are based upon experimental data, he is in the field of statistics.

There are various reasons why some prefer to adopt different principles. One of them is illustrated by the following example.

Example 1. Let $X_1, X_2, \ldots, X_n$ denote a random sample from the uniform distribution that has the p.d.f.

$$\begin{aligned} f(x) &= \tfrac{1}{2}, \qquad \theta - 1 < x < \theta + 1, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

where $\theta > 0$. Let Y_1 and Y_n be the first and the nth order statistics. Let us compute the probability that the random interval (Y_1, Y_n) includes the point θ, the mean of the uniform distribution. We could compute this probability by using the joint p.d.f. of Y_1 and Y_n. But the following useful fact makes that unnecessary. The complement of the event $Y_1 < \theta < Y_n$ is the union of the mutually exclusive events $Y_n < \theta$ and $\theta < Y_1$. That is, the only ways in which the interval (Y_1, Y_n) does not include θ are those in which all the items of the sample are below θ or those in which all the items of the sample are above θ. Thus,

$$\begin{aligned} \Pr(Y_1 < \theta < Y_n) &= 1 - \Pr(Y_n < \theta) - \Pr(\theta < Y_1) \\ &= 1 - \Pr(X_i < \theta, i = 1, \ldots, n) \\ &\quad - \Pr(\theta < X_i, i = 1, \ldots, n). \end{aligned}$$

Since $\Pr(X_i < \theta) = \Pr(\theta < X_i) = \frac{1}{2}$, $i = 1, 2, \ldots, n$, and since $X_1, X_2, \ldots, X_n$ are mutually stochastically independent, we have

$$\Pr(Y_1 < \theta < Y_n) = 1 - (\tfrac{1}{2})^n - (\tfrac{1}{2})^n = 1 - (\tfrac{1}{2})^{n-1}.$$

If we take $n = 5$, this probability is approximately 0.94. Let the experiment yield the observed values $Y_1 = y_1$ and $Y_n = y_n$. The interval (y_1, y_n) is then

a 94 per cent confidence interval for θ. Now it is quite possible, and highly probable, that the difference $y_n - y_1$ exceeds 1. In such a case we know definitely that the interval (y_1, y_n) includes the point θ because the difference $y_n - y_1$ exceeds one half of the difference $(\theta + 1) - (\theta - 1) = 2$. Thus, when $y_n - y_1 > 1$, it may seem odd to call the interval (y_1, y_n) a 94 per cent confidence interval for the unknown parameter θ. That is, there is no difference of opinion about the probabilistic statement that $\Pr(Y_1 < \theta < Y_n) = 0.94$ approximately when $n = 5$, but there can be a difference of opinion about the desirability of calling the interval (y_1, y_n), with $y_n - y_1 > 1$ a 94 per cent confidence interval for θ.

We shall now describe another approach to the problem of interval estimation. This approach takes into account any prior knowledge of the experiment that the statistician has and it is one application of a principle of statistical inference that may be called *Bayesian* statistics. Consider a random variable X that has a distribution of probability that depends upon the symbol θ, where θ is an element of a well-defined set Ω. For example, if the symbol θ is the mean of a normal distribution, then Ω may be the real line. We have previously looked upon θ as being some constant, although an unknown constant. Let us now introduce a random variable Θ that has a distribution of probability over the set Ω; and, just as we look upon x as a possible value of the random variable X, we now look upon θ as a possible value of the random variable Θ. Thus the distribution of X depends upon θ, a random determination of the random variable Θ. We shall denote the p.d.f. of Θ by $h(\theta)$ and we take $h(\theta) = 0$ when θ is not an element of Ω. Let $X_1, X_2, \ldots, X_n$ denote a random sample from this distribution of X, and let Y denote a statistic which is a function of $X_1, X_2, \ldots, X_n$. We can find the p.d.f. of Y for every given θ; that is, we can find the conditional p.d.f. of Y, given $\Theta = \theta$, which we denote by $g(y|\theta)$. Thus the joint p.d.f. of Y and Θ is given by

$$k(\theta, y) = h(\theta)g(y|\theta).$$

If Θ is a random variable of the continuous type, the marginal p.d.f. of Y is given by

$$k_1(y) = \int_{-\infty}^{\infty} h(\theta)g(y|\theta)\, d\theta.$$

If Θ is a random variable of the discrete type, integration would be replaced by summation. In either case the conditional p.d.f. of Θ, given $Y = y$, is

$$k(\theta|y) = \frac{k(y, \theta)}{k_1(y)} = \frac{h(\theta)g(y|\theta)}{k_1(y)}, \qquad k_1(y) > 0.$$

This relationship is one form of Bayes' formula (see Exercise 2.7, p. 61). We can now find two functions $u(y)$ and $v(y)$ so that the conditional probability

$$\Pr[u(y) < \Theta < v(y) | Y = y] = \int_{u(y)}^{v(y)} k(\theta|y)\, d\theta,$$

is large, say 0.95. The experimental values of $X_1, X_2, \ldots, X_n$, say $x_1, x_2, \ldots, x_n$, provide us with an experimental value of Y, say y. Then the interval $u(y)$ to $v(y)$ is an interval estimate of θ in the sense that the conditional probability of Θ belonging to that interval is equal to 0.95.

In Bayesian statistics, the p.d.f. $h(\theta)$ is called the *prior p.d.f.* of Θ, and the conditional p.d.f. $k(\theta|y)$ is called the *posterior p.d.f.* of Θ. This is because $h(\theta)$ is the p.d.f. of Θ prior to the observation of Y, whereas $k(\theta|y)$ is the p.d.f. of Θ after the observation of Y has been made. In many instances, $h(\theta)$ is not known; yet the choice of $h(\theta)$ affects the interval $u(y)$ to $v(y)$. In these instances, the statistician takes into account all prior knowledge of the experiment and he *assigns* the prior p.d.f. $h(\theta)$. This, of course, injects the problem of *personal* or *subjective probability*.

Remark. Many persons who have contributed to the development of probability theory and statistics have thought of probability as a rational measure of belief. If this approach to probability is taken, certainly any past information about the problem under consideration should be used to help assign the various probabilities. In particular, information about the relative frequency of the occurrence of an event would undoubtedly influence the assignment of probability. However, it is important to note that *personal* or *subjective probability* can be extended to situations beyond those which can be repeated under identical conditions. In addition, if a person who assigns subjective probabilities adopts certain principles to ensure consistent assignments, it is possible to prove that this probability measure satisfies the requirements of Definition 7, p. 11. Since the mathematics of probability depends upon these requirements, the subsequent development does not depend upon the particular approach to probability. As a matter of fact, an individual statistician may very well find it appropriate to vary his approach from one situation to the next.

The following example will show that if the statistician's prior notions are not too sharp, the influence of the choice of $h(\theta)$ may be small. Specifically, if the variance of the prior distribution of Θ is large, relative to the variance of the statistic Y, given $\Theta = \theta$, the length of the interval $u(y)$ to $v(y)$ may not be changed too much by selecting somewhat different functions for $h(\theta)$.

Example 2. $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution which is $n(\theta, \sigma^2)$, where the positive variance is known. The statistician makes use of prior knowledge, and he takes $h(\theta)$, the prior p.d.f. of Θ, to be $n(\mu, \tau^2)$, where μ and τ^2 are known. Let $Y = \bar{X}$, the mean of the random sample. Then $g(y|\theta)$, the conditional p.d.f. of Y, given θ, is $n(\theta, \sigma^2/n)$. Accordingly, the joint p.d.f. $k(\theta, y)$ of Θ and Y is given by

$$k(\theta, y) = \frac{1}{(2\pi)\tau\sigma/\sqrt{n}} \exp\left[-\frac{(\theta - \mu)^2}{2\tau^2} - \frac{(y - \theta)^2}{2\sigma^2/n}\right],$$
$$-\infty < y < \infty, \quad -\infty < \theta < \infty.$$

Now

$$(y - \theta)^2 = (y - \mu + \mu - \theta)^2 = (\theta - \mu)^2 - 2(\theta - \mu)(y - \mu) + (y - \mu)^2,$$

so that the exponent in the expression for $k(\theta, y)$ can be written

$$-\left(\frac{1}{2}\right)\left[\left(\frac{1}{\tau^2} + \frac{1}{\sigma^2/n}\right)(\theta - \mu)^2 - \left(\frac{2}{\sigma^2/n}\right)(\theta - \mu)(y - \mu) + \left(\frac{1}{\sigma^2/n}\right)(y - \mu)^2\right].$$

It is easy to verify that Θ and Y have a bivariate normal distribution with common mean μ, respective variances τ^2 and $\tau^2 + \sigma^2/n$, and correlation coefficient $\tau/\sqrt{\tau^2 + \sigma^2/n}$. Thus, $k_1(y)$ is $n(\mu, \tau^2 + \sigma^2/n)$ and $k(\theta|y)$ is the normal p.d.f. with mean

$$\mu + \left(\frac{\tau}{\sqrt{\tau^2 + \sigma^2/n}}\right)\left(\frac{\tau}{\sqrt{\tau^2 + \sigma^2/n}}\right)(y - \mu) = \frac{\mu\sigma^2/n + y\tau^2}{\tau^2 + \sigma^2/n}$$

and variance

$$\tau^2\left(1 - \frac{\tau^2}{\tau^2 + \sigma^2/n}\right) = \frac{\tau^2\sigma^2/n}{\tau^2 + \sigma^2/n}.$$

So, for illustration, given $Y = y$, the conditional probability of the event

$$\frac{\mu\sigma^2/n + y\tau^2}{\tau^2 + \sigma^2/n} - \frac{2\tau\sigma/\sqrt{n}}{\sqrt{\tau^2 + \sigma^2/n}} < \Theta < \frac{\mu\sigma^2/n + y\tau^2}{\tau^2 + \sigma^2/n} + \frac{2\tau\sigma/\sqrt{n}}{\sqrt{\tau^2 + \sigma^2/n}}$$

is 0.954, approximately. The interval described by this inequality then serves as an interval estimate of θ in the sense implied by the conditional probability. Table 1 is illustrative of the comment that the distribution $n(\mu, \tau^2)$ of Θ may not influence the interval too much if the variance τ^2 of Θ is large relative to the variance σ^2/n of $Y = \bar{X}$. During the study of the table, it is important to recall that the unconditional mean of $Y = \bar{X}$ is μ.

It is interesting to compare these results with the 95.4 per cent confidence interval for θ, namely,

$$(y - 2\sigma/\sqrt{n}, y + 2\sigma/\sqrt{n}) = (y - 1, y + 1).$$

Thus, if the sample size n is large enough so that $y = \bar{x}$ approximates the mean μ and so that σ^2/n is much smaller than τ^2, then the confidence interval estimate of θ and the Bayesian interval estimate of θ are not too different.

TABLE I

Bayesian Interval Estimates of θ

($y = \bar{x}$ and the conditional probability is 0.954)

n	σ^2	τ^2	Approximate Intervals
100	25	$\frac{200}{121}$	$(0.03\mu + 0.97y - 0.98, 0.03\mu + 0.97y + 0.98)$
100	25	$\frac{400}{81}$	$(0.05\mu + 0.95y - 0.98, 0.05\mu + 0.95y + 0.98)$
100	25	$\frac{36}{25}$	$(0.15\mu + 0.85y - 0.92, 0.15\mu + 0.85y + 0.92)$
100	25	$\frac{9}{64}$	$(0.64\mu + 0.36y - 0.60, 0.64\mu + 0.36y + 0.60)$

EXERCISES

6.29. Let Y denote the sum of the items of a random sample of size n from a distribution which is $b(1, \theta)$. Assume that the unknown θ is a value of a random variable Θ which has a beta distribution with parameters α and β. (a) Display the joint p.d.f. of Θ and Y. Note that Θ is a random variable of the continuous type and Y is a random variable of the discrete type. (b) Determine the marginal p.d.f. of Y. (c) Show that the conditional distribution of Θ, given $Y = y$, is a beta distribution. Find the mean and variance of this distribution. (d) Explain how to find a Bayesian interval estimate of θ subject to the availability of suitable tables of integrals.

6.30. Let Y denote the sum of the items of a random sample of size n from a Poisson distribution with unknown mean θ. Work the previous exercise with the word "beta" replaced by the word "gamma."

6.31. Refer to Example 2 of this section and verify that $k(\theta, y)$ is the bivariate normal p.d.f. that it was stated to be.

6.32. Let X be $n(0, 1/\theta)$. Assume that the unknown θ is a value of a random variable Θ which has a gamma distribution with parameters $\alpha = r/2$ and $\beta = 2/r$, where r is a positive integer. Show that X has a marginal t distribution with r degrees of freedom. This procedure is called one of *compounding*, and it may be used by a Bayesian statistician as a way of first presenting the t distribution, as well as other distributions.

6.33. Let X have a Poisson distribution with parameter θ. Assume that the unknown θ is a value of a random variable Θ that has a gamma distribution with parameters $\alpha = r$ and $\beta = (1 - p)/p$, where r is a positive integer and $0 < p < 1$. Show, by the procedure of compounding, that X has a marginal distribution which is negative binomial, a distribution that was introduced earlier (p. 90) under very different assumptions.

Chapter 7

Point Estimation and Sufficient Statistics

7.1 The Problem of Point Estimation

Let a random variable X have a p.d.f. which is of known functional form but the p.d.f. depends upon an unknown parameter θ that may have any value in a set Ω. This will be denoted by writing the p.d.f. in the form $f(x; \theta)$, $\theta \in \Omega$. The set Ω will be called the *parameter space.* Thus we are confronted, not with one distribution of probability, but with a *family* of distributions. To each value of θ, $\theta \in \Omega$, there corresponds one member of the family. A family of probability density functions will be denoted by the symbol $\{f(x; \theta); \theta \in \Omega\}$. Any member of this family of probability density functions will be denoted by the symbol $f(x; \theta)$, $\theta \in \Omega$. We shall continue to use the special symbols that have been adopted for the normal, the chi-square, and the binomial distributions. We may, for instance, have the family $\{n(\theta, 1); \theta \in \Omega\}$, where Ω is the set $-\infty < \theta < \infty$. One member of this family of distributions is the distribution that is $n(0, 1)$. Any arbitrary member is $n(\theta, 1)$, $-\infty < \theta < \infty$.

Consider a family of probability density functions $\{f(x; \theta); \theta \in \Omega\}$. The student will immediately perceive that *interval* estimation of the parameter θ can be looked upon as an effort on the part of the statistician (who uses information derived from the performance of a random experiment, perhaps blended with prior knowledge) to select a subset of the family of probability density functions in such a way that any member of the subset can plausibly be regarded as being the p.d.f. of the random variable. In certain instances, however, this may not be

adequate for the purpose at hand; it may be that the experimenter needs to select precisely *one* member of the family as being the p.d.f. of his random variable. That is, he needs a *point* estimate of θ as opposed to an interval estimate of θ. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that has a p.d.f. which is one member (but which member we do not know) of the family $\{f(x; \theta); \theta \in \Omega\}$ of probability density functions. That is, our sample arises from a distribution that has the p.d.f. $f(x; \theta)$; $\theta \in \Omega$. Our problem is that of defining a statistic $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ so that if $x_1, x_2, \ldots, x_n$ are the observed experimental values of $X_1, X_2, \ldots, X_n$, then the number $y_1 = u_1(x_1, x_2, \ldots, x_n)$ will be a good point estimate of θ. In this connection, we need the following definition.

Definition 1. Any statistic whose mathematical expectation is equal to a parameter θ is called an *unbiased* statistic for the parameter θ. Otherwise the statistic is said to be *biased*.

Now it would seem that if $y_1 = u_1(x_1, x_2, \ldots, x_n)$ is to qualify as a good point estimate of θ, there should be a great probability that the statistic $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ will be close to θ; that is, θ should be a sort of rallying point for the numbers $y_1 = u_1(x_1, x_2, \ldots, x_n)$. This can be achieved in one way by selecting $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ in such a way that not only is Y_1 an unbiased statistic for θ but also the variance of Y_1 is as small as it can be made. We do this because the variance of Y_1 is a measure of the intensity of the concentration of the probability for Y_1 in the neighborhood of the point $\theta = E(Y_1)$. Accordingly, we define a "best" statistic for the parameter θ in the following manner.

Definition 2. For a given positive integer n, $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ will be called a *best* statistic for the parameter θ if Y_1 is unbiased, $E(Y_1) = \theta$, and if the variance of Y_1 is less than or equal to the variance of every other unbiased statistic for θ.

Remarks. Definition 2 is wholly arbitrary. There are many ways of defining a "best" statistic for a parameter. Our definition adopts the principles of unbiasedness and minimum variance as being reasonable. One of our purposes in adopting these principles is to motivate, in a somewhat natural way, the study of an important class of statistics called *sufficient statistics*.

It should also be pointed out that a statistic, which is to be used to estimate a parameter, is often called an *estimator* of that parameter. Thus, in this terminology, our best statistic is an unbiased, minimum variance

estimator. The experimentally determined value of the estimator is then referred to as a (point) *estimate* of the parameter.

For illustration, let $X_1, X_2, \ldots, X_9$ denote a random sample from a distribution which is $n(\theta, 1)$, $-\infty < \theta < \infty$. Since the statistic $\bar{X} = (X_1 + X_2 + \cdots + X_9)/9$ is $n(\theta, \frac{1}{9})$, $\bar{X}$ is an unbiased statistic for θ. The statistic X_1 is $n(\theta, 1)$; so X_1 is also an unbiased statistic for θ. Although the variance $\frac{1}{9}$ of $\bar{X}$ is less than the variance one of X_1, we cannot say, with $n = 9$, that $\bar{X}$ is the best statistic for θ; the definition of a best statistic requires that the comparison be made with every unbiased statistic for θ. To be sure, it is quite impossible to tabulate all other unbiased statistics for this parameter θ, so other methods must be developed for making the comparisons of the variances. A beginning on this problem will be made in Section 7.2.

EXERCISES

7.1. Show that the mean $\bar{X}$ of a random sample of size n from a distribution having p.d.f. $f(x; \theta) = (1/\theta)e^{-(x/\theta)}$, $0 < x < \infty$, $0 < \theta < \infty$, zero elsewhere, is an unbiased statistic for θ and has variance θ^2/n.

7.2. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a normal distribution with mean zero and variance θ, $0 < \theta < \infty$. Show that $\sum_1^n X_i^2/n$ is an unbiased statistic for θ and has variance $2\theta^2/n$.

7.3. Let $Y_1 < Y_2 < Y_3$ be the order statistics of a random sample of size 3 from the uniform distribution having p.d.f. $f(x; \theta) = 1/\theta$, $0 < x < \theta$, $0 < \theta < \infty$, zero elsewhere. Show that $4Y_1$, $2Y_2$, and $\frac{4}{3}Y_3$ are all unbiased statistics for θ. Find the variance of each of these unbiased statistics.

7.4. Let Y_1 and Y_2 be two stochastically independent unbiased statistics for θ. Say the variance of Y_1 is twice the variance of Y_2. Find the constants k_1 and k_2 so that $k_1Y_1 + k_2Y_2$ is an unbiased statistic with smallest possible variance for such a linear combination.

7.5. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution of the continuous type with p.d.f. $f(x)$, where $f(\theta - z) = f(\theta + z)$ for all z. Let $u(X_1, \ldots, X_n)$ denote a statistic for which $u(x_1 + h, \ldots, x_n + h) = u(x_1, \ldots, x_n) + h$, for all h, and $u(-x_1, \ldots, -x_n) = -u(x_1, \ldots, x_n)$. Prove that $u(X_1, \ldots, X_n)$ is an unbiased statistic for θ, subject to the existence of $E[u(X_1, \ldots, X_n)]$. *Hint.* Display $E[u(X_1, \ldots, X_n) - \theta]$ as an n-fold integral. Make the change of variables $x_i = y_i + \theta$, $i = 1, 2, \ldots, n$, and denote the resulting integral by K. In K make the change of variables $y_i = -z_i$, $i = 1, 2, \ldots, n$, to show that $K = -K$.

7.2 A Sufficient Statistic for a Parameter

Let n denote a fixed positive integer and let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution that has p.d.f. $f(x; \theta)$, $\theta \in \Omega$. The joint p.d.f. of $X_1, X_2, \ldots, X_n$ is then

$$f(x_1; \theta)f(x_2; \theta)\cdots f(x_n; \theta).$$

Let $Y_1 = u_1(X_1, X_2, \ldots, X_n)$, $Y_2 = u_2(X_1, X_2, \ldots, X_n), \ldots, Y_n = u_n(X_1, X_2, \ldots, X_n)$ be n statistics. If the associated transformation is one-to-one, the joint p.d.f. of $Y_1, Y_2, \ldots, Y_n$ is given (in a notation which has been used previously) by

$$g(y_1, y_2, \ldots, y_n; \theta) = |J|f[w_1(y_1, \ldots, y_n); \theta] \times f[w_2(y_1, \ldots, y_n); \theta]\cdots f[w_n(y_1, \ldots, y_n); \theta],$$

if the random variables are of the continuous type, and by

$$g(y_1, y_2, \ldots, y_n; \theta) = f[w_1(y_1, \ldots, y_n); \theta] \times f[w_2(y_1, \ldots, y_n); \theta]\cdots f[w_n(y_1, \ldots, y_n); \theta]$$

if the random variables are of the discrete type. If the transformation is not one-to-one, the right-hand member in each case is the sum of, say, k terms of this form. Whether or not the transformation is one-to-one, the marginal p.d.f. of Y_1 is given by

$$g_1(y_1; \theta) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(y_1, y_2, \ldots, y_n; \theta)\, dy_2\cdots dy_n$$

in the continuous case and by

$$g_1(y_1; \theta) = \sum_{y_n}\cdots\sum_{y_2} g(y_1, y_2, \ldots, y_n; \theta)$$

in the discrete case. For either the continuous or the discrete case, the conditional p.d.f. of $Y_2, \ldots, Y_n$, given $Y_1 = y_1$, is

$$h(y_2, \ldots, y_n|y_1; \theta) = \frac{g(y_1, y_2, \ldots, y_n; \theta)}{g_1(y_1; \theta)}$$

provided $g_1(y_1; \theta) > 0$. In the interest of conservation of space, it will be tacitly assumed in the future that, if a transformation is not one-to-one, the joint p.d.f. $g(y_1, y_2, \ldots, y_n; \theta)$ would be given by a sum of k terms.

In general the conditional p.d.f. $h(y_2, \ldots, y_n|y_1; \theta)$ will depend upon θ, as is indicated by the notation. In certain cases, however, as will be seen presently, the conditional p.d.f. will not depend upon θ, and these cases will be very important to us. The following definition is made.

Definition 3. Let n denote a fixed positive integer and let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution which is one member of the family $\{f(x; \theta); \theta \in \Omega\}$. The statistic $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ is called a *sufficient statistic* for θ if and only if for all other statistics $Y_2 = u_2(X_1, X_2, \ldots, X_n), \ldots, Y_n = u_n(X_1, X_2, \ldots, X_n)$ (for which the Jacobian of the associated transformation is not identically zero) the conditional p.d.f. $h(y_2, \ldots, y_n | y_1)$ of $Y_2, \ldots, Y_n$, given $Y_1 = y_1$, does not depend upon the parameter θ, whatever be the fixed value y_1.

Perhaps it should be emphasized that when we say that a function does not depend upon θ, we mean not only that θ does not appear in the "formula" but also that the domain of the function does not involve θ. For example, the function

$$f(x) = \tfrac{1}{2}, \qquad \theta - 1 < x < \theta + 1, \ -\infty < \theta < \infty,$$
$$= 0 \text{ elsewhere},$$

does depend upon θ.

Remark. Why we use the terminology "sufficient statistic" can be explained as follows: If a statistic Y_1 satisfies the preceding definition, then the conditional p.d.f. of Y_2, say given $Y_1 = y_1$, does not depend upon the parameter θ. As a consequence, once given $Y_1 = y_1$, it is impossible to use Y_2 to make a statistical inference about θ; for example, we could not find a confidence interval for θ based on Y_2. In a sense, Y_1 exhausts all the information about θ that is contained in the sample. It is in this sense that we call Y_1 a sufficient statistic for θ. In some instances it is preferable to call Y_1 a sufficient statistic for the family $\{f(x; \theta); \theta \in \Omega\}$ of probability density functions.

When a sufficient statistic Y_1 for a parameter θ exists, we shall see later that it may serve as the basis for a best statistic for θ. At the moment, however, let us prove two theorems about a sufficient statistic for a parameter θ. The first of these theorems is called the Fisher-Neyman criterion for the existence of a sufficient statistic for a parameter. It may be stated as follows.

Theorem 1. *Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that has p.d.f. $f(x; \theta)$, $\theta \in \Omega$. Let $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ be a statistic whose p.d.f. is $g_1(y_1; \theta)$. Then $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ is a sufficient statistic for θ if, and only if,*

$$f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta) = g_1[u_1(x_1, x_2, \ldots, x_n); \theta] H(x_1, x_2, \ldots, x_n),$$

where, for every fixed value of $y_1 = u_1(x_1, x_2, \ldots, x_n)$, $H(x_1, x_2, \ldots, x_n)$ does not depend upon θ.

Proof. To prove the Fisher-Neyman criterion for random variables of the continuous type, suppose first that

$$f(x_1; \theta)f(x_2; \theta)\cdots f(x_n; \theta) = g_1[u_1(x_1, x_2, \ldots, x_n); \theta]H(x_1, x_2, \ldots, x_n).$$

In our proof we shall make the one-to-one transformation $y_1 = u_1(x_1, \ldots, x_n)$, $y_2 = u_2(x_1, \ldots, x_n), \ldots, y_n = u_n(x_1, \ldots, x_n)$ having the inverse functions $x_1 = w_1(y_1, \ldots, y_n)$, $x_2 = w_2(y_1, \ldots, y_n), \ldots, x_n = w_n(y_1, \ldots, y_n)$ and Jacobian J. Thus

$$f[w_1(y_1, \ldots, y_n); \theta]\cdots f[w_n(y_1, \ldots, y_n); \theta]|J| = g_1(y_1; \theta)H[w_1(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n)]|J|.$$

The left-hand member is the joint p.d.f. $g(y_1, y_2, \ldots, y_n; \theta)$ of $Y_1 = u_1(X_1, \ldots, X_n)$, $Y_2 = u_2(X_1, \ldots, X_n), \ldots, Y_n = u_n(X_1, \ldots, X_n)$. In the right-hand member, $g_1(y_1; \theta)$ is the p.d.f. of Y_1, so that $H(w_1, w_2, \ldots, w_n)|J|$ is the quotient of $g(y_1, y_2, \ldots, y_n; \theta)$ and $g_1(y_1; \theta)$; that is, it is the conditional p.d.f. $h(y_2, \ldots, y_n|y_1; \theta)$ of $Y_2, \ldots, Y_n$, given $Y_1 = y_1$. But $H(x_1, x_2, \ldots, x_n)$, and thus $H[w_1(y_1, \ldots, y_n), w_2(y_1, \ldots, y_n), \ldots, w_n(y_1, \ldots, y_n)]$, was given not to depend upon θ. Since θ was not introduced in the transformation and accordingly not in the Jacobian J, it follows that $h(y_2, \ldots, y_n|y_1; \theta)$ does not depend upon θ and that Y_1 is a sufficient statistic for θ.

The converse is proved by taking

$$g(y_1, y_2, \ldots, y_n; \theta) = g_1(y_1; \theta)h(y_2, \ldots, y_n|y_1),$$

where $h(y_2, \ldots, y_n|y_1)$ does not depend upon θ. Now look upon $y_1, y_2, \ldots, y_n$ as the "old" variables and $x_1, x_2, \ldots, x_n$ as the "new" variables. If J^* is the Jacobian, then the elements of J^* are the partial derivatives of the y's with respect to the x's. If we replace $y_1, y_2, \ldots, y_n$ by the functions $u_1(x_1, x_2, \ldots, x_n)$, $u_2(x_1, x_2, \ldots, x_n), \ldots, u_n(x_1, x_2, \ldots, x_n)$, and multiply by the absolute value of J^*, we have

$$g[u_1(x_1, \ldots, x_n), \ldots, u_n(x_1, \ldots, x_n); \theta]|J^*| = g_1[u_1(x_1, \ldots, x_n); \theta]h(u_2, \ldots, u_n|u_1)|J^*|.$$

The left-hand member is necessarily the joint p.d.f. $f(x_1; \theta)f(x_2; \theta)\cdots f(x_n; \theta)$ of $X_1, X_2, \ldots, X_n$. Since $h(y_2, \ldots, y_n|y_1)$, and thus $h(u_2, \ldots, u_n|u_1)$, does not depend upon θ, then

$$H(x_1, x_2, \ldots, x_n) = h(u_2, \ldots, u_n|u_1)|J^*|$$

is a function that does not depend upon θ and the proof is complete.

If the random variables are of the discrete type, of course no Jacobian is involved, but otherwise, the proof is identical with the one given.

We now give some examples that are illustrative of Theorem 1.

Example 1. Let $X_1, X_2, \ldots, X_n$ denote a random sample from the distribution that has p.d.f.

$$\begin{aligned} f(x; \theta) &= \theta^x(1 - \theta)^{1-x}, \qquad x = 0, 1; 0 < \theta < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

The statistic $Y_1 = X_1 + X_2 + \cdots + X_n$ has the p.d.f.

$$\begin{aligned} g_1(y_1; \theta) &= \frac{n!}{y_1!\,(n - y_1)!}\,\theta^{y_1}(1 - \theta)^{n-y_1}, \qquad y_1 = 0, 1, \ldots, n, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Accordingly, the joint p.d.f. of $X_1, X_2, \ldots, X_n$ may be written

$$\begin{aligned} &\theta^{x_1}(1 - \theta)^{1-x_1}\theta^{x_2}(1 - \theta)^{1-x_2}\cdots\theta^{x_n}(1 - \theta)^{1-x_n} \\ &\quad = \theta^{x_1+x_2+\cdots+x_n}(1 - \theta)^{n-(x_1+x_2+\cdots+x_n)} \\ &\quad = g_1[(x_1 + x_2 + \cdots + x_n); \theta] \\ &\qquad \times \left\{\frac{(x_1 + x_2 + \cdots + x_n)!\,[n - (x_1 + x_2 + \cdots + x_n)]!}{n!}\right\}. \end{aligned}$$

In accordance with the Fisher-Neyman criterion, $Y_1 = X_1 + X_2 + \cdots + X_n$ is a sufficient statistic for θ.

Example 2. Let $Y_1 < Y_2 < \cdots < Y_n$ denote the order statistics of a random sample $X_1, X_2, \ldots, X_n$ from the distribution that has p.d.f.

$$\begin{aligned} f(x; \theta) &= e^{-(x-\theta)}, \qquad \theta < x < \infty, -\infty < \theta < \infty, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

The p.d.f. of the statistic Y_1 is

$$\begin{aligned} g_1(y_1; \theta) &= ne^{-n(y_1-\theta)}, \qquad \theta < y_1 < \infty, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Thus the joint p.d.f. of $X_1, X_2, \ldots, X_n$ may be written

$$e^{-(x_1-\theta)}e^{-(x_2-\theta)}\cdots e^{-(x_n-\theta)} = g_1(\min x_i; \theta)\left\{\frac{\exp(-x_1 - x_2 - \cdots - x_n)}{n \exp[-n(\min x_i)]}\right\}.$$

By the Fisher-Neyman criterion, the first order statistic Y_1 is a sufficient statistic for θ.

If we are to show, by means of the Fisher-Neyman criterion, that a certain statistic Y_1 is or is not a sufficient statistic for a parameter θ,

we must first of all know the p.d.f. of Y_1, say $g_1(y_1; \theta)$. In some instances it may be quite tedious to find this p.d.f. Fortunately, this problem can be avoided if we will but prove the following *factorization theorem.*

Theorem 2. *Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that has p.d.f. $f(x; \theta)$, $\theta \in \Omega$. The statistic $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ is a sufficient statistic for θ if and only if we can find two nonnegative functions, k_1 and k_2, such that*

$$f(x_1; \theta)f(x_2; \theta)\cdots f(x_n; \theta) = k_1[u_1(x_1, x_2, \ldots, x_n); \theta]k_2(x_1, x_2, \ldots, x_n),$$

where, for every fixed value of $y_1 = u_1(x_1, x_2, \ldots, x_n)$, $k_2(x_1, x_2, \ldots, x_n)$ does not depend upon θ.

Proof. We shall prove the theorem when the random variables are of the continuous type. Assume the factorization as stated in the theorem. We make the same transformation that we made in the proof of Theorem 1. The joint p.d.f. of the statistics $Y_1, Y_2, \ldots, Y_n$ is then given by

$$g(y_1, y_2, \ldots, y_n; \theta) = k_1(y_1; \theta)k_2(w_1, w_2, \ldots, w_n)|J|,$$

where $w_i = w_i(y_1, y_2, \ldots, y_n)$, $i = 1, 2, \ldots, n$. The p.d.f. of Y_1, say $g_1(y_1; \theta)$, is given by

$$\begin{aligned} g_1(y_1; \theta) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1, y_2, \ldots, y_n; \theta)\, dy_2 \cdots dy_n \\ &= k_1(y_1; \theta) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |J| k_2(w_1, w_2, \ldots, w_n)\, dy_2 \cdots dy_n. \end{aligned}$$

Now the function k_2, for every fixed value of $y_1 = u_1(x_1, \ldots, x_n)$, does not depend upon θ. Nor is θ involved in either the Jacobian J or the limits of integration. Hence the $(n - 1)$-fold integral in the right-hand member of the preceding equation is a function of y_1 alone, say $m(y_1)$. Thus

$$g_1(y_1; \theta) = k_1(y_1; \theta)m(y_1).$$

If $m(y_1) = 0$, then $g_1(y_1; \theta) = 0$. If $m(y_1) > 0$, we can write

$$k_1[u_1(x_1, \ldots, x_n); \theta] = \frac{g_1[u_1(x_1, \ldots, x_n); \theta]}{m[u_1(x_1, \ldots, x_n)]},$$

and the assumed factorization becomes

$$f(x_1; \theta)\cdots f(x_n; \theta) = g_1[u_1(x_1, \ldots, x_n); \theta] \frac{k_2(x_1, \ldots, x_n)}{m[u_1(x_1, \ldots, x_n)]}.$$

Since neither the function k_2 nor the function m depends upon θ, then in accordance with Theorem 1, Y_1 is a sufficient statistic for the parameter θ.

Conversely, if Y_1 is a sufficient statistic for θ, the factorization can be realized by taking the function k_1 to be the p.d.f. of Y_1, namely, the function g_1. This completes the proof of the theorem.

Example 3. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution which is $n(\theta, \sigma^2)$, $-\infty < \theta < \infty$, where the variance σ^2 is known. If $\bar{x} = \sum_1^n x_i/n$, then

$$\sum_{i=1}^{n} (x_i - \theta)^2 = \sum_{i=1}^{n} [(x_i - \bar{x}) + (\bar{x} - \theta)]^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \theta)^2,$$

because

$$2 \sum_{i=1}^{n} (x_i - \bar{x})(\bar{x} - \theta) = 2(\bar{x} - \theta) \sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

Thus the joint p.d.f. of $X_1, X_2, \ldots, X_n$ may be written

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\sum_{i=1}^{n} (x_i - \theta)^2/(2\sigma^2)\right]$$

$$= \{\exp[-n(\bar{x} - \theta)^2/(2\sigma^2)]\}\left\{\frac{\exp\left[-\sum_{i=1}^{n} (x_i - \bar{x})^2/(2\sigma^2)\right]}{(\sigma\sqrt{2\pi})^n}\right\}.$$

Since the first factor of the right-hand member of this equation depends upon $x_1, x_2, \ldots, x_n$ only through $\bar{x}$, and since the second factor does not depend upon θ, the factorization theorem implies that the mean $\bar{X}$ of the sample is, for every fixed value of σ^2, a sufficient statistic for θ, the mean of the normal distribution.

We could have used the Fisher-Neyman criterion in the preceding example because we know that $\bar{X}$ is $n(\theta, \sigma^2/n)$. Let us now consider an example in which that criterion is inappropriate.

Example 4. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution with p.d.f.

$$f(x, \theta) = \theta x^{\theta - 1}, \qquad 0 < x < 1,$$
$$= 0 \text{ elsewhere},$$

where $0 < \theta$. We shall use the factorization theorem to prove that the product $u_1(X_1, X_2, \ldots, X_n) = X_1 X_2 \cdots X_n$ is a sufficient statistic for θ. The joint p.d.f. of $X_1, X_2, \ldots, X_n$ is

$$\theta^n (x_1 x_2 \cdots x_n)^{\theta - 1} = [\theta^n (x_1 x_2 \cdots x_n)^{\theta}]\left[\frac{1}{x_1 x_2 \cdots x_n}\right],$$

where $0 < x_i < 1$, $i = 1, 2, \ldots, n$. In the factorization theorem let

$$k_1[u_1(x_1, x_2, \ldots, x_n); \theta] = \theta^n(x_1x_2\cdots x_n)^\theta$$

and

$$k_2(x_1, x_2, \ldots, x_n) = \frac{1}{x_1x_2\cdots x_n}.$$

Since $k_2(x_1, x_2, \ldots, x_n)$ does not depend upon θ, the product $X_1X_2\cdots X_n$ is a sufficient statistic for θ.

There is a tendency for some students to apply incorrectly the factorization theorem in those instances in which the domain of positive probability density depends upon the parameter θ. This is due to the fact that they do not give proper consideration to the domain of the function $k_2(x_1, x_2, \ldots, x_n)$ for every fixed value of $y_1 = u_1(x_1, x_2, \ldots, x_n)$. This will be illustrated in the next example.

Example 5. In Example 2, with $f(x; \theta) = e^{-(x-\theta)}$, $\theta < x < \infty$, $-\infty < \theta < \infty$, it was found that the first order statistic Y_1 is a sufficient statistic for θ. To illustrate our point, take $n = 3$ so that the joint p.d.f. of X_1, X_2, X_3 is given by

$$e^{-(x_1-\theta)}e^{-(x_2-\theta)}e^{-(x_3-\theta)}, \qquad \theta < x_i < \infty,$$

$i = 1, 2, 3$. We can factor this in several ways. One way, with $n = 3$, is given in Example 2. Another way would be to write the joint p.d.f. as the product

$$[e^{-(\max x_i)+3\theta}][e^{-x_1-x_2-x_3+\max x_i}].$$

Certainly there is no θ in the *formula* of the second factor, and it might be assumed that $Y_3 = \max X_i$ is itself a sufficient statistic for θ. But what is the *domain* of the second factor for every fixed value of $y_3 = \max x_i$? If $\max x_i = x_1$, the domain is $\theta < x_2 < x_1$, $\theta < x_3 < x_1$; if $\max x_i = x_2$, the domain is $\theta < x_1 < x_2$, $\theta < x_3 < x_2$; and if $\max x_i = x_3$, the domain is $\theta < x_1 < x_3$, $\theta < x_2 < x_3$. That is, for each fixed $y_3 = \max x_i$, the domain of the second factor depends upon θ. Thus the factorization theorem is not satisfied.

If the student has some difficulty using the factorization theorem when the domain of positive probability density depends upon θ, we recommend that he use the Fisher-Neyman criterion even though it may be somewhat longer.

Before taking the next step in our search for a best statistic for a parameter θ, let us consider an important property possessed by a sufficient statistic $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ for θ. The conditional

p.d.f. of another statistic, say $Y_2 = u_2(X_1, X_2, \ldots, X_n)$, given $Y_1 = y_1$, does not depend upon θ. On intuitive grounds, we might surmise that the conditional p.d.f. of Y_2, given some linear function $aY_1 + b$, $a \neq 0$, of Y_1, does not depend upon θ. That is, it seems as though the random variable $aY_1 + b$ is also a sufficient statistic for θ. This conjecture is correct. In fact, every single-valued function $Z = u(Y_1)$, or $Z = u[u_1(X_1, X_2, \ldots, X_n)] = v(X_1, X_2, \ldots, X_n)$, not involving θ, with a single-valued inverse $Y_1 = w(Z)$, is also a sufficient statistic for θ. To prove this, we write, in accordance with the factorization theorem,

$$f(x_1; \theta) \cdots f(x_n; \theta) = k_1[u_1(x_1, x_2, \ldots, x_n); \theta]k_2(x_1, x_2, \ldots, x_n).$$

However, $y_1 = w(z)$ or, equivalently, $u_1(x_1, x_2, \ldots, x_n) = w[v(x_1, x_2, \ldots, x_n)]$, which is not a function of θ. Hence

$$f(x_1; \theta) \cdots f(x_n; \theta) = k_1\{w[v(x_1, \ldots, x_n)]; \theta\}k_2(x_1, x_2, \ldots, x_n).$$

Since the first factor of the right-hand member of this equation is a function of $z = v(x_1, \ldots, x_n)$ and θ, while the second factor does not depend upon θ, the factorization theorem implies that $Z = u(Y_1)$ is also a sufficient statistic for θ.

EXERCISES

7.6. Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution $n(0, \theta)$, $0 < \theta < \infty$. Show that $\sum_1^n X_i^2$ is a sufficient statistic for θ.

7.7. Prove that the sum of the items of a random sample of size n from a Poisson distribution having parameter θ, $0 < \theta < \infty$, is a sufficient statistic for θ.

7.8. Show that the nth order statistic of a random sample of size n from the uniform distribution having p.d.f. $f(x; \theta) = 1/\theta$, $0 < x < \theta$, $0 < \theta < \infty$, zero elsewhere, is a sufficient statistic for θ. Generalize this result by considering the p.d.f. $f(x; \theta) = Q(\theta)M(x)$, $0 < x < \theta$, $0 < \theta < \infty$, zero elsewhere. Here, of course,

$$\int_0^\theta M(x)\, dx = \frac{1}{Q(\theta)}.$$

7.9. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a geometric distribution that has p.d.f. $f(x; \theta) = (1 - \theta)^x\theta$, $x = 0, 1, 2, \ldots$, $0 < \theta < 1$, zero elsewhere. Show that $\sum_1^n X_i$ is a sufficient statistic for θ.

7.10. Show that the sum of the items of a random sample of size n from a gamma distribution which has p.d.f. $f(x; \theta) = (1/\theta)e^{-x/\theta}$, $0 < x < \infty$, $0 < \theta < \infty$, zero elsewhere, is a sufficient statistic for θ.

7.11. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a beta distribution with parameters $\alpha = \theta > 0$ and $\beta = 2$. Show that the product $X_1X_2\cdots X_n$ is a sufficient statistic for θ.

7.12. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution with p.d.f. $f(x; \theta) = 1/\pi[1 + (x - \theta)^2]$, $-\infty < x < \infty$, $-\infty < \theta < \infty$. Can the joint p.d.f. of $X_1, X_2, \ldots, X_n$ be written in the form given in Theorem 2? Does the parameter θ have a sufficient statistic?

7.3 The Rao-Blackwell Theorem

We shall prove the Rao-Blackwell theorem.

Theorem 3. *Let X and Y denote random variables such that Y has mean μ and positive variance σ_Y^2. Let $E(Y|x) = \varphi(x)$. Then $E[\varphi(X)] = \mu$ and $\sigma_{\varphi(X)}^2 \leq \sigma_Y^2$.*

Proof. We shall give the proof when the random variables are of the continuous type. Let $f(x, y)$, $f_1(x)$, $f_2(y)$, and $h(y|x)$ denote, respectively, the joint p.d.f. of X and Y, the two marginal probability density functions, and the conditional p.d.f. of Y, given $X = x$. Then

$$E(Y|x) = \int_{-\infty}^{\infty} yh(y|x)\,dy = \frac{\int_{-\infty}^{\infty} yf(x, y)\,dy}{f_1(x)} = \varphi(x),$$

so that

$$\int_{-\infty}^{\infty} yf(x, y)\,dy = \varphi(x)f_1(x).$$

We have

$$\begin{aligned} E[\varphi(X)] &= \int_{-\infty}^{\infty} \varphi(x)f_1(x)\,dx = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} yf(x, y)\,dy\right] dx \\ &= \int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} f(x, y)\,dx\right] dy \\ &= \int_{-\infty}^{\infty} yf_2(y)\,dy = \mu, \end{aligned}$$

and the first part of the conclusion of the theorem is established. Consider next

$$\begin{aligned} \sigma_Y^2 &= E[(Y - \mu)^2] = E\{[(Y - \varphi(X)) + (\varphi(X) - \mu)]^2\} \\ &= E\{[Y - \varphi(X)]^2\} + E\{[\varphi(X) - \mu]^2\} + 2E\{[Y - \varphi(X)][\varphi(X) - \mu]\}. \end{aligned}$$

We shall show that the last term of the right-hand member of the immediately preceding equation is zero. We have

$$E\{[Y - \varphi(X)][\varphi(X) - \mu]\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [y - \varphi(x)][\varphi(x) - \mu]f(x, y)\,dy\,dx.$$

In this integral, we shall write $f(x, y)$ in the form $h(y|x)f_1(x)$, and we shall integrate first on y to obtain

$$\int_{-\infty}^{\infty} [\varphi(x) - \mu]\left\{\int_{-\infty}^{\infty} [y - \varphi(x)]h(y|x)\, dy\right\} f_1(x)\, dx.$$

But $\varphi(x)$ is the mean of the conditional p.d.f. $h(y|x)$. Hence

$$\int_{-\infty}^{\infty} [y - \varphi(x)]h(y|x)\, dy = 0,$$

and, accordingly,

$$E\{[Y - \varphi(X)][\varphi(X) - \mu]\} = 0.$$

Moreover,

$$\sigma_{\varphi(X)}^2 = E\{[\varphi(X) - \mu]^2\}$$

and

$$E\{[Y - \varphi(X)]^2\} \geq 0.$$

Accordingly,

$$\sigma_Y^2 \geq \sigma_{\varphi(X)}^2,$$

and the theorem is proved when X and Y are random variables of the continuous type. The proof in the discrete case is identical to the proof given here with the exception that summation replaces integration.

It is interesting to note, in connection with the proof of the theorem, that unless the probability measure of the set $\{(x, y); y - \varphi(x) = 0\}$ is equal to one, then $E\{[Y - \varphi(X)]^2\} > 0$, and we have the strict inequality $\sigma_Y^2 > \sigma_{\varphi(X)}^2$.

We shall give an illustrative example.

Example 1. Let X and Y have a bivariate normal distribution with means μ_1 and μ_2, with positive variances σ_1^2 and σ_2^2, and with correlation coefficient ρ. Here $E(Y) = \mu = \mu_2$ and $\sigma_Y^2 = \sigma_2^2$. Now $E(Y|x)$ is linear in x and it is given by

$$E(Y|x) = \varphi(x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

Thus $\varphi(X) = \mu_2 + \rho(\sigma_2/\sigma_1)(X - \mu_1)$ and $E[\varphi(X)] = \mu_2$, as stated in the theorem. Moreover,

$$\begin{aligned}\sigma_{\varphi(X)}^2 &= E\{[\varphi(X) - \mu_2]^2\} \\ &= E\left\{\left[\rho\frac{\sigma_2}{\sigma_1}(X - \mu_1)\right]^2\right\} \\ &= \rho^2\sigma_2^2.\end{aligned}$$

With $-1 < \rho < 1$, we have the strict inequality $\sigma_2^2 > \rho^2\sigma_2^2$. It should be observed that $\varphi(X)$ is not a statistic if at least one of the five parameters is unknown.

We shall use the Rao-Blackwell theorem to help us in our search for a best statistic for a parameter. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that has p.d.f. $f(x; \theta)$, $\theta \in \Omega$, where it is known that $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ is a sufficient statistic for the parameter θ. Let $Y_2 = u_2(X_1, X_2, \ldots, X_n)$ be another statistic (but not a function of Y_1 alone) which is an unbiased statistic for θ; that is, $E(Y_2) = \theta$. Consider $E(Y_2|y_1)$. This expectation is a function of y_1, say $\varphi(y_1)$. Since Y_1 is a sufficient statistic for θ, the conditional p.d.f. of Y_2, given $Y_1 = y_1$, does not depend upon θ so that $E(Y_2|y_1) = \varphi(y_1)$ is a function of y_1 alone. That is to say, here $\varphi(Y_1)$ is a statistic. In accordance with the Rao-Blackwell theorem, $\varphi(Y_1)$ is an unbiased statistic for θ; and because Y_2 is not a function of Y_1 alone, the variance of $\varphi(Y_1)$ is strictly less than the variance of Y_2. We shall summarize this discussion in the following theorem.

Theorem 4. *Let $X_1, X_2, \ldots, X_n$, n a fixed positive integer, denote a random sample from a distribution (continuous or discrete) that has p.d.f. $f(x; \theta)$, $\theta \in \Omega$. Let $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ be a sufficient statistic for θ, and let $Y_2 = u_2(X_1, X_2, \ldots, X_n)$, not a function of Y_1 alone, be an unbiased statistic for θ. Then $E(Y_2|y_1) = \varphi(y_1)$ defines a statistic $\varphi(Y_1)$. This statistic $\varphi(Y_1)$ is a function of the sufficient statistic for θ; it is an unbiased statistic for θ; and its variance is less than that of Y_2.*

This theorem tells us that in our search for a best statistic for a parameter, we may, if a sufficient statistic for the parameter exists, restrict that search to functions of the sufficient statistic. For if we begin with an unbiased statistic Y_2 that is not a function of the sufficient statistic Y_1 alone, then we can always improve on this by computing $E(Y_2|y_1) = \varphi(y_1)$ so that $\varphi(Y_1)$ is an unbiased statistic with smaller variance than that of Y_2.

In Section 7.4 an additional concept, called completeness, will be introduced. Completeness, in conjunction with sufficiency, will lead us to the best statistic for a parameter.

EXERCISES

7.13. Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ be the order statistics of a random sample of size 5 from the uniform distribution having p.d.f. $f(x; \theta) = 1/\theta$, $0 < x < \theta$, $0 < \theta < \infty$, zero elsewhere. Show that $2Y_3$ is an unbiased statistic for θ. Determine the joint p.d.f. of Y_3 and the sufficient statistic Y_5 for θ. Find the conditional expectation $E(2Y_3|y_5) = \varphi(y_5)$. Compare the variances of $2Y_3$ and $\varphi(Y_5)$.

7.14. If X_1, X_2 is a random sample of size 2 from a distribution having p.d.f. $f(x; \theta) = (1/\theta)e^{-x/\theta}$, $0 < x < \infty$, $0 < \theta < \infty$, zero elsewhere, find the joint p.d.f. of the sufficient statistic $Y_1 = X_1 + X_2$ for θ and $Y_2 = X_2$. Show that Y_2 is an unbiased statistic for θ with variance θ^2. Find $E(Y_2|y_1) = \varphi(y_1)$ and the variance of $\varphi(Y_1)$.

7.15. Let the random variables X and Y have the joint p.d.f. $f(x, y) = (2/\theta^2)e^{-(x+y)/\theta}$, $0 < x < y < \infty$, zero elsewhere. (a) Show that the mean and the variance of Y are, respectively, $3\theta/2$ and $5\theta^2/4$. (b) Show that $E(Y|x) = x + \theta$. In accordance with the Rao-Blackwell theorem, the expected value of $X + \theta$ is that of Y, namely, $3\theta/2$, and the variance of $X + \theta$ is less than that of Y. Show that the variance of $X + \theta$ is in fact $\theta^2/4$.

7.16. In the notation of Theorem 4, consider $E[\varphi(Y_1)|Y_2 = y_2] = \tau(y_2)$. It follows from the Rao-Blackwell theorem that $E[\tau(Y_2)] = \theta$ and that the variance of $\tau(Y_2)$ is less than that of $\varphi(Y_1)$. Why is not $\tau(Y_2)$ a better statistic for θ than $\varphi(Y_1)$? *Hint.* Is $\tau(Y_2)$ a statistic?

7.4 Completeness

Consider the family $\{f(x; \theta); 0 < \theta < \infty\}$ of probability density functions, where $f(x; \theta) = 1/\theta$, $0 < x < \theta$, zero elsewhere. Let $u(x)$ be a continuous function of x (but not a function of θ), and suppose we are told that $E[u(X)] = 0$ for every $\theta > 0$. We shall show that this requires $u(x) = 0$ for every $x > 0$. For each x, $x \in \{x; 0 < x < \infty\}$, there exists at least one member of the family of probability density functions which is positive. So we could say we shall show that $E[u(X)] = 0$ for every $\theta > 0$ requires $u(x) = 0$ at every point x at which at least one member of the family of probability density functions is positive. Our assumption is that

$$E[u(X)] = \int_{-\infty}^{\infty} u(x)f(x; \theta)\,dx = \int_0^{\theta} u(x)\frac{1}{\theta}\,dx = 0, \qquad \text{for } \theta > 0.$$

Thus

$$\int_0^{\theta} u(x)\,dx = 0, \qquad \text{for } \theta > 0.$$

If we take the derivative of both members with respect to θ, we have, by the fundamental theorem of calculus, that

$$u(\theta) = 0, \qquad \text{for } \theta > 0.$$

That is,

$$u(x) = 0, \qquad \text{for } x > 0,$$

as we proposed to show.

Next, consider the family $\{f(x; \theta); 0 < \theta < 1\}$ of probability density functions, where $f(x; \theta) = \theta^x(1 - \theta)^{1-x}$, $x = 0, 1$, zero elsewhere. Again, let $u(x)$ be a continuous function of x (but not a function of θ), and suppose that $E[u(X)] = 0$ for every θ, $0 < \theta < 1$. We shall show that this requires $u(x)$ to be zero at every point x at which at least one member of the family of probability density functions is positive. In this example, each member of the family is positive at the same points, namely, at $x = 0$ and $x = 1$. Hence we shall actually show that $E[u(X)] = 0$, $0 < \theta < 1$, requires $u(0) = u(1) = 0$. We have

$$\begin{aligned} E[u(X)] &= \sum_x u(x)f(x; \theta) \\ &= \sum_{x=0}^{1} u(x)\theta^x(1 - \theta)^{1-x} \\ &= u(0)(1 - \theta) + u(1)\theta \\ &= \theta[u(1) - u(0)] + u(0) = 0, \qquad 0 < \theta < 1. \end{aligned}$$

If a linear function of z, say $az + b$, is equal to zero for more than one value of z, then $a = b = 0$. But we have a linear function of θ, namely, $\theta[u(1) - u(0)] + u(0)$, which vanishes for every θ, $0 < \theta < 1$. Accordingly,

$$u(1) - u(0) = 0 \qquad \text{and} \qquad u(0) = 0,$$

or

$$u(0) = u(1) = 0,$$

as we wanted to show.

We now define the concept of completeness.

Definition 4. Let $\{f(x; \theta); \theta \in \Omega\}$ denote a family of probability density functions of either the continuous type or the discrete type. Let $u(x)$ denote a continuous function of x (but not a function of θ). If $E[u(X)] = 0$ for every θ, $\theta \in \Omega$, requires $u(x)$ to be zero at each point x at which at least one member of the family of probability density functions is positive, then the family of probability density functions is called a *complete family*.

Remarks. Two remarks should be made. On p. 42, it was noted that the existence of $E[u(X)]$ implies that the integral (or sum) converge absolutely. This absolute convergence was tacitly assumed in our definition of completeness and it is needed to prove that certain families of probability density functions are complete. The second remark concerns the continuity of $u(x)$. This assumption would not be made in a more advanced textbook. The assumption does permit us to show by elementary methods that certain

families of probability density functions are complete. However, the primary purpose of the assumption of continuity is to make it unnecessary to discuss certain aspects of measure theory.

We were able to show by elementary methods that a certain family of uniform probability density functions and that a certain family of binomial probability density functions are in fact complete families. In general, however, elementary methods are not adequate. Since a proof that a family is complete is a problem in analysis rather than a problem in probability or statistics, we shall not press the point in this book. Instead, as is our practice, we shall assert theorems from other branches of mathematics whenever they are needed.

EXERCISES

7.17. If $az^2 + bz + c = 0$ for more than two values of z, then $a = b = c = 0$. Use this result to show that the family $\{b(2, \theta); 0 < \theta < 1\}$ is complete.

7.18. If the infinite series $a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots$ converges to zero for all values of z in a given interval, then $a_0 = a_1 = a_2 = a_3 = \cdots = 0$. Use this to prove that the family given by the Poisson p.d.f. $f(x; \theta) = \theta^x e^{-\theta}/x!$, $x = 0, 1, 2, \ldots, 0 < \theta < \infty$, zero elsewhere, is complete.

7.19. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from the uniform distribution having p.d.f. $f(x; \theta) = 1/\theta$, $0 < x < \theta$, $0 < \theta < \infty$, zero elsewhere. Show that the family of probability density functions of Y_n, the sufficient statistic for θ, is complete.

7.20. Let the p.d.f. of X be $f(x; \theta) = Q(\theta)M(x)$, $0 < x < \theta$, $0 < \theta < \infty$, zero elsewhere, where $M(x)$ is a continuous function of x. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from this distribution. Show that the family of probability density functions of Y_n, the sufficient statistic for θ, is complete.

7.21. Show that each of the following families $\{f(x; \theta); 0 < \theta < \infty\}$ is not complete by finding at least one nonzero continuous function $u(x)$ such that $E[u(X)] = 0$, for all $\theta > 0$.

(a) $$f(x; \theta) = \frac{1}{2\theta}, \qquad -\theta < x < \theta,$$
$$= 0 \text{ elsewhere}$$

(b) $n(0, \theta)$.

7.22. Prove that the family $\{b(n, \theta), 0 < \theta < 1\}$ is complete. *Hint.* Display $E[u(X)] = 0$, show that the constant term $u(0)$ is equal to zero, divide both members of the equation by $\theta \neq 0$, and repeat the argument.

7.5 Uniqueness

Let the parameter θ in the p.d.f. $f(x; \theta)$, $\theta \in \Omega$, have a sufficient statistic $Y_1 = u_1(X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots, X_n$ is a random sample from this distribution. Let the p.d.f. of Y_1 be $g_1(y_1; \theta)$, $\theta \in \Omega$. It has been seen that, if there is any unbiased statistic Y_2 (not a function of Y_1 alone) for θ, then there is at least one function of Y_1 that is an unbiased statistic for θ, and our search for a best statistic for θ may be restricted to functions of Y_1. Suppose it has been verified that a certain continuous function $\varphi(Y_1)$, not a function of θ, is such that $E[\varphi(Y_1)] = \theta$ for all values of θ, $\theta \in \Omega$. Let $\psi(Y_1)$ be another continuous function of the sufficient statistic Y_1 alone so that we have also $E[\psi(Y_1)] = \theta$ for all values of θ, $\theta \in \Omega$. Hence

$$E[\varphi(Y_1) - \psi(Y_1)] = 0, \qquad \theta \in \Omega.$$

If the family $\{g_1(y_1; \theta); \theta \in \Omega\}$ is complete, the continuous function $\varphi(y_1) - \psi(y_1) = 0$ at each point y_1 at which at least one member of the family is positive. That is, at all points of nonzero probability density, we have, for every continuous unbiased statistic $\psi(Y_1)$,

$$\varphi(y_1) = \psi(y_1).$$

In this sense [that is, $\varphi(y_1) = \psi(y_1)$, at all points of nonzero probability density], $\varphi(Y_1)$ is the unique continuous function of Y_1 which is an unbiased statistic for θ. In accordance with the Rao-Blackwell theorem, $\varphi(Y_1)$ has a smaller variance than every other unbiased statistic for θ. That is, the statistic $\varphi(Y_1)$ is the best statistic for θ. This fact is stated in the following theorem.

Theorem 5. *Let $X_1, X_2, \ldots, X_n$, n a fixed positive integer, denote a random sample from a distribution that has p.d.f. $f(x; \theta)$, $\theta \in \Omega$, let $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ be a sufficient statistic for θ, and let the family $\{g_1(y_1; \theta); \theta \in \Omega\}$ of probability density functions be complete. If there is a continuous function of Y_1 which is an unbiased statistic for θ, then this function of Y_1 is the unique best statistic for θ. Here "unique" is used in the sense described in the preceding paragraph.*

This theorem is a special case of a very general theorem due to Lehmann and Scheffé. Our assumption of continuity is unnecessarily restrictive, but it seems undesirable to try to relax it here. Such an attempt only creates problems, called measure theoretic problems, which we wish to avoid.

The statement that Y_1 is a sufficient statistic for a parameter θ, $\theta \in \Omega$, and that the family $\{g_1(y_1; \theta); \theta \in \Omega\}$ of probability density functions is complete is lengthy and somewhat awkward. We shall adopt the less descriptive, but more convenient, terminology that Y_1 is a *complete sufficient statistic* for θ. In the next section we shall study a fairly large class of probability density functions for which a complete sufficient statistic Y_1 for θ can be determined by inspection.

EXERCISES

7.23. Let $X_1, X_2, \ldots, X_n$ represent a random sample from each of the discrete distributions having the following probability density functions:

(a) $$f(x; \theta) = \theta^x(1 - \theta)^{1-x}, \qquad x = 0, 1, 0 < \theta < 1,$$
$$= 0 \text{ elsewhere.}$$

(b) $$f(x; \theta) = \frac{\theta^x e^{-\theta}}{x!}, \qquad x = 0, 1, 2, \ldots, 0 < \theta < \infty,$$
$$= 0 \text{ elsewhere.}$$

Show in each case that $Y_1 = \sum_1^n X_i$ is a complete sufficient statistic for θ. Find, for each distribution, the unique continuous function of Y_1 which is the best statistic for θ.

7.24. Show that the first order statistic Y_1 of a random sample of size n from the distribution having p.d.f. $f(x; \theta) = e^{-(x-\theta)}$, $\theta < x < \infty$, $-\infty < \theta < \infty$, zero elsewhere, is a complete sufficient statistic for θ. Find the unique continuous function of this statistic which is the best statistic for θ.

7.25. Let a random sample of size n be taken from a distribution of the discrete type with p.d.f. $f(x; \theta) = 1/\theta$, $x = 1, 2, \ldots, \theta$, zero elsewhere, where θ is an unknown positive integer. (a) Show that the largest item, say Y, of the sample is a complete sufficient statistic for θ. (b) Prove that

$$[Y^{n+1} - (Y - 1)^{n+1}]/[Y^n - (Y - 1)^n]$$

is the unique best statistic for θ.

7.26. In the notation of this section, let ω be a subset of the parameter space Ω. Prove the following: (a) If Y_1 is a sufficient statistic for $\theta \in \Omega$, then Y_1 is a sufficient statistic for $\theta \in \omega$. (b) If the family $\{g_1(y_1; \theta); \theta \in \omega\}$ is complete, then the family $\{g_1(y_1; \theta); \theta \in \Omega\}$ is complete.

7.27. Let Y_n denote the nth order statistic of a random sample of size n from a distribution with p.d.f. $f(x; \theta) = Q(\theta)M(x)$, $0 < x < \theta$, zero elsewhere, with $0 < \theta < \infty$. Find $\varphi(Y_n)$, the unique best statistic for θ.

7.6 The Exponential Class of Probability Density Functions

Consider a family $\{f(x; \theta); \theta \in \Omega\}$ of probability density functions, where Ω is the interval set $\Omega = \{\theta; \gamma < \theta < \delta\}$, where γ and δ are known constants, and where

$$(1) \qquad f(x; \theta) = \exp [p(\theta)K(x) + S(x) + q(\theta)], \qquad a < x < b,$$
$$= 0 \text{ elsewhere.}$$

A p.d.f. of the form (1) is said to be a member of the *exponential class* of probability density functions of the continuous type. If, in addition,

(i) neither a nor b depends upon θ, $\gamma < \theta < \delta$,
(ii) $p(\theta)$ is a nontrivial continuous function of θ, $\gamma < \theta < \delta$,
(iii) each of $K'(x) \neq 0$ and $S(x)$ is a continuous function of x, $a < x < b$,

we say that we have a *regular case* of the exponential class. A p.d.f.

$$f(x; \theta) = \exp [p(\theta)K(x) + S(x) + q(\theta)], \qquad x = a_1, a_2, a_3, \ldots,$$
$$= 0 \text{ elsewhere}$$

is said to represent a regular case of the exponential class of probability density functions of the discrete type if

(i) the set $\{x; x = a_1, a_2, \ldots\}$ does not depend upon θ,
(ii) $p(\theta)$ is a nontrivial continuous function of θ, $\gamma < \theta < \delta$,
(iii) $K(x)$ is a nontrivial function of x on the set $\{x; x = a_1, a_2, \ldots\}$.

For example, each member of the family $\{f(x; \theta); 0 < \theta < \infty\}$, where $f(x; \theta)$ is $n(0, \theta)$, represents a regular case of the exponential class of the continuous type because

$$f(x; \theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-x^2/2\theta}$$
$$= \exp\left[-\frac{1}{2\theta} x^2 - \ln \sqrt{2\pi\theta}\right], \qquad -\infty < x < \infty.$$

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that has a p.d.f. which represents a regular case of the exponential class of the continuous type. The joint p.d.f. of $X_1, X_2, \ldots, X_n$ is

$$\exp\left[p(\theta) \sum_1^n K(x_i) + \sum_1^n S(x_i) + nq(\theta)\right]$$

for $a < x_i < b$, $i = 1, 2, \ldots, n$, $\gamma < \theta < \delta$, and is zero elsewhere. At points of positive probability density, this joint p.d.f. may be written as the product of the two nonnegative functions

$$\exp\left[p(\theta)\sum_1^n K(x_i) + nq(\theta)\right]\exp\left[\sum_1^n S(x_i)\right].$$

In accordance with the factorization theorem (Theorem 2, p. 219) $Y_1 = \sum_1^n K(X_i)$ is a sufficient statistic for the parameter θ. To prove that $Y_1 = \sum_1^n K(X_i)$ is a sufficient statistic for θ in the discrete case, we take the joint p.d.f. of $X_1, X_2, \ldots, X_n$ to be positive on a discrete set of points, say, when $x_i \in \{x; x = a_1, a_2, \ldots\}$, $i = 1, 2, \ldots, n$. We then use the factorization theorem. It is left as an exercise to show that in either the continuous or the discrete case the p.d.f. of Y_1 is of the form

$$g_1(y_1; \theta) = R(y_1)\exp[p(\theta)y_1 + nq(\theta)]$$

at points of positive probability density. The points of positive probability density and the function $R(y_1)$ do not depend upon θ.

At this time we use a theorem in analysis to assert that the family $\{g_1(y_1; \theta); \gamma < \theta < \delta\}$ of probability density functions is complete. This is the same theorem we used when we asserted that a moment-generating function (when it exists) uniquely determines a distribution. In the present context it can be stated as the following theorem.

Theorem 6. *Let $f(x; \theta)$, $\gamma < \theta < \delta$, be a p.d.f. which represents a regular case of the exponential class. Then if $X_1, X_2, \ldots, X_n$ (where n is a fixed positive integer) is a random sample from a distribution with p.d.f. $f(x; \theta)$, the statistic $Y_1 = \sum_1^n K(X_i)$ is a sufficient statistic for θ and the family $\{g_1(y_1; \theta); \gamma < \theta < \delta\}$ of probability density functions of Y_1 is complete. That is, Y_1 is a complete sufficient statistic for θ.*

This theorem has useful implications. In a regular case of form (1), we can see by inspection that the sufficient statistic is $Y_1 = \sum_1^n K(X_i)$. If we can see how to form a continuous function of Y_1, say $\varphi(Y_1)$, so that $E[\varphi(Y_1)] = \theta$, then the statistic $\varphi(Y_1)$ is unique and is the best statistic for θ.

Example 1. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a normal distribution that has p.d.f.

$$f(x; \theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right], \qquad -\infty < x < \infty, -\infty < \theta < \infty,$$

or

$$f(x;\theta) = \exp\left[\frac{\theta}{\sigma^2}x - \frac{x^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2} - \frac{\theta^2}{2\sigma^2}\right].$$

Here σ^2 is any fixed positive number. This is a regular case of the exponential class with

$$p(\theta) = \frac{\theta}{\sigma^2}, \qquad K(x) = x,$$

$$S(x) = -\frac{x^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2}, \qquad q(\theta) = -\frac{\theta^2}{2\sigma^2}.$$

Accordingly, $Y_1 = X_1 + X_2 + \cdots + X_n = n\overline{X}$ is a complete sufficient statistic for the mean θ of a normal distribution for every fixed value of the variance σ^2. Since $E(Y_1) = n\theta$, then $\varphi(Y_1) = Y_1/n = \overline{X}$ is the only continuous function of Y_1 which is an unbiased statistic for θ; and being a function of the sufficient statistic Y_1, it has minimum variance. That is, $\overline{X}$ is the unique best statistic for θ. Incidentally, since Y_1 is a single-valued function of $\overline{X}$, $\overline{X}$ itself is also a complete sufficient statistic for θ.

Example 2. Consider a Poisson distribution with parameter θ, $0 < \theta < \infty$. The p.d.f. of this distribution is

$$f(x;\theta) = \frac{\theta^x e^{-\theta}}{x!} = \exp\left[(\ln\theta)x - \ln(x!) - \theta\right], \qquad x = 0, 1, 2, \ldots,$$
$$= 0 \text{ elsewhere.}$$

In accordance with Theorem 6, $Y_1 = \sum_1^n X_i$ is a complete sufficient statistic for θ. Since $E(Y_1) = n\theta$, the statistic $\varphi(Y_1) = Y_1/n = \overline{X}$, which is also a complete sufficient statistic for θ, is the unique best statistic for θ.

EXERCISES

7.28. Write the p.d.f.

$$f(x;\theta) = \frac{1}{6\theta^4}x^3e^{-x/\theta}, \qquad 0 < x < \infty, 0 < \theta < \infty,$$

zero elsewhere, in the exponential form. If $X_1, X_2, \ldots, X_n$ is a random sample from this distribution, find a complete sufficient statistic Y_1 for θ and the unique continuous function $\varphi(Y_1)$ of this statistic which is the best statistic for θ. Is $\varphi(Y_1)$ itself a complete sufficient statistic?

7.29. Let $X_1, X_2, \ldots, X_n$ denote a random sample of size $n > 2$ from a distribution with p.d.f. $f(x;\theta) = \theta e^{-\theta x}$, $0 < x < \infty$, zero elsewhere, and $\theta > 0$. Then $Y = \sum_1^n X_i$ is a sufficient statistic for θ. Prove that $(n-1)/Y$ is the best statistic for θ.

7.30. Let $\bar{X}$ denote the mean of the random sample $X_1, X_2, \ldots, X_n$ from a gamma-type distribution with parameters $\alpha > 0$ and $\beta = \theta > 0$. Compute $E[X_1|\bar{x}]$.

7.31. Let X be a random variable with a p.d.f. of a regular case of the exponential class. Show that $E[K(X)] = -q'(\theta)/p'(\theta)$, provided these derivatives exist, by differentiating both members of the equality

$$\int_a^b \exp[p(\theta)K(x) + S(x) + q(\theta)]\,dx = 1$$

with respect to θ. By a second differentiation, find the variance of $K(X)$.

7.32. Given that $f(x; \theta) = \exp[\theta K(x) + S(x) + q(\theta)]$, $a < x < b$, $\gamma < \theta < \delta$, represents a regular case of the exponential class. Show that the moment-generating function $M(t)$ of $Y = K(X)$ is $M(t) = \exp[q(\theta) - q(\theta + t)]$, $\gamma < \theta + t < \delta$.

7.33. Given, in the preceding exercise, that $E(Y) = E[K(X)] = \theta$. Prove that Y is $n(\theta, 1)$.

7.34. If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution that has a p.d.f. which is a regular case of the exponential class, show that the p.d.f. of $Y_1 = \sum_1^n K(X_i)$ is of the form $g_1(y_1; \theta) = R(y_1)\exp[p(\theta)y_1 + nq(\theta)]$. *Hint.* Let $Y_2 = X_2, \ldots, Y_n = X_n$ be $n - 1$ auxiliary random variables. Find the joint p.d.f. of $Y_1, Y_2, \ldots, Y_n$ and then the marginal p.d.f. of Y_1.

7.35. Let Y denote the median and let $\bar{X}$ denote the mean of a random sample of size $n = 2k + 1$ from a distribution which is $n(\mu, \sigma^2)$. Compute $E(Y|\bar{X} = \bar{x})$. *Hint.* See Exercise 4.61 (p. 154).

7.7 Functions of a Parameter

Up to this point we have sought an unbiased and minimum variance statistic for a parameter θ. Not always, however, are we interested in θ but rather in a function of θ. This will be illustrated in the following examples.

Example 1. Let $X_1, X_2, \ldots, X_n$ denote the items of a random sample of size $n > 1$ from a distribution which is $b(1, \theta)$, $0 < \theta < 1$. We know that if $Y = \sum_1^n X_i$, then Y/n is the unique best statistic for θ. Now the variance of Y/n is $\theta(1 - \theta)/n$. Suppose an unbiased and minimum variance statistic for this variance is sought. Because Y is a sufficient statistic for θ, it is known that we can restrict our search to functions of Y. Consider the statistic $(Y/n)(1 - Y/n)/n$. This statistic is suggested by the fact that Y/n is the best statistic for θ. The expectation of this statistic is given by

$$\frac{1}{n}E\left[\frac{Y}{n}\left(1 - \frac{Y}{n}\right)\right] = \frac{1}{n^2}E[Y] - \frac{1}{n^3}E[Y^2].$$

Now $E[Y] = n\theta$ and $E[Y^2] = n\theta(1 - \theta) + n^2\theta^2$. Hence

$$\frac{1}{n}E\left[\frac{Y}{n}\left(1 - \frac{Y}{n}\right)\right] = \frac{n-1}{n}\frac{\theta(1-\theta)}{n}.$$

If we multiply both members of this equation by $n/(n-1)$ we find that the statistic $(Y/n)(1 - Y/n)/(n-1)$ is the unique best statistic for estimating the variance of Y/n.

A somewhat different, but very important, problem in point estimation is considered in the next example. In this example, the distribution of a random variable X is described by a p.d.f. $f(x; \theta)$ which depends upon $\theta \in \Omega$. The problem is to estimate the fractional part of the probability for this distribution which is at or to the left of a fixed point c. Thus we seek an unbiased, minimum variance statistic for $F(c; \theta)$, where $F(x; \theta)$ is the distribution function of X.

Example 2. Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n > 1$ from a distribution which is $n(\theta, 1)$. Suppose we wish to find a best statistic for the function of θ defined by

$$\Pr(X \leq c) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}\, dx = N(c - \theta),$$

where c is a fixed constant. There are many unbiased statistics for $N(c - \theta)$. We first exhibit one of these, say $u(X_1)$, a function of X_1 alone. We shall then compute the conditional expectation, $E[u(X_1)|\bar{X} = \bar{x}] = \varphi(\bar{x})$, of this unbiased statistic, given the sufficient statistic $\bar{X}$, the mean of the sample. In accordance with the theorems of Rao-Blackwell and Lehmann-Scheffé, $\varphi(\bar{X})$ is the unique best statistic for $N(c - \theta)$.

Consider the function $u(x_1)$, where

$$\begin{aligned} u(x_1) &= 1, \qquad x_1 \leq c, \\ &= 0, \qquad x_1 > c. \end{aligned}$$

The expected value of the random variable $u(X_1)$ is given by

$$\begin{aligned} E[u(X_1)] &= \int_{-\infty}^{\infty} u(x_1) \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x_1 - \theta)^2}{2}\right] dx_1 \\ &= \int_{-\infty}^{c} (1) \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x_1 - \theta)^2}{2}\right] dx_1, \end{aligned}$$

because $u(x_1) = 0$, $x_1 > c$. But the latter integral has the value $N(c - \theta)$. That is, $u(X_1)$ is an unbiased statistic for $N(c - \theta)$.

We shall next discuss the joint distribution of X_1 and $\bar{X}$ and the conditional distribution of X_1, given $\bar{X} = \bar{x}$. This conditional distribution will enable us to compute $E[u(X_1)|\bar{X} = \bar{x}] = \varphi(\bar{x})$. In accordance with Exercise

4.78 (p. 162) the joint distribution of X_1 and $\bar{X}$ is bivariate normal with means θ and θ, variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 1/n$, and correlation coefficient $\rho = 1/\sqrt{n}$. Thus the conditional p.d.f. of X_1, given $\bar{X} = \bar{x}$, is normal with linear conditional mean

$$\theta + (\rho\sigma_1/\sigma_2)(\bar{x} - \theta) = \bar{x}$$

and with variance

$$\sigma_1^2(1 - \rho^2) = (n - 1)/n.$$

The conditional expectation of $u(X_1)$, given $\bar{X} = \bar{x}$, is then

$$\varphi(\bar{x}) = \int_{-\infty}^{\infty} u(x_1)\sqrt{\frac{n}{n-1}}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{n(x_1 - \bar{x})^2}{2(n-1)}\right]dx_1$$

$$= \int_{-\infty}^{c} \sqrt{\frac{n}{n-1}}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{n(x_1 - \bar{x})^2}{2(n-1)}\right]dx_1.$$

The change of variable $z = \sqrt{n}(x_1 - \bar{x})/\sqrt{n-1}$ enables us to write, with $c' = \sqrt{n}(c - \bar{x})/\sqrt{n-1}$, this conditional expectation as

$$\varphi(\bar{x}) = \int_{-\infty}^{c'} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz = N(c') = N\left[\frac{\sqrt{n}(c - \bar{x})}{\sqrt{n-1}}\right].$$

Thus the unique, unbiased, and minimum variance statistic for $N(c - \theta)$ is, for every fixed constant c, given by $\varphi(\bar{X}) = N[\sqrt{n}(c - \bar{X})/\sqrt{n-1}]$.

Remark. We should like to draw the attention of the student to a rather important fact. This has to do with the adoption of a *principle*, such as the principle of unbiasedness and minimum variance. A principle is not a theorem; and seldom does a principle yield satisfactory results in all cases. So far, this principle has provided quite satisfactory results. To see that this is not always the case, let X have a Poisson distribution with parameter θ, $0 < \theta < \infty$. We may look upon X as a random sample of size one from this distribution. Thus X is a complete sufficient statistic for θ. We seek the best statistic for $e^{-2\theta}$, best in the sense of being unbiased and having minimum variance. Consider $Y = (-1)^X$. We have

$$E[Y] = E[(-1)^X] = \sum_{x=0}^{\infty} \frac{(-\theta)^x e^{-\theta}}{x!} = e^{-2\theta}.$$

Accordingly, $(-1)^X$ is the (unique) best statistic for $e^{-2\theta}$, in the sense described. Here our principle leaves much to be desired. We are endeavoring to elicit some information about the number $e^{-2\theta}$, where $0 < e^{-2\theta} < 1$. Yet our point estimate is either -1 or $+1$, each of which is a very poor estimate of a number between zero and one. We do not wish to leave the student with the impression that an unbiased, minimum variance statistic is *bad*. That is not the case at all. We merely wish to point out that if one tries

hard enough he can find instances where such a statistic is *not good*. In the next chapter we shall consider other principles of estimation, each of which has its own merits.

EXERCISES

7.36. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution which is $b(1, \theta)$. Find the best statistic for the variance $n\theta(1 - \theta)$ of $Y = \sum X_i$.

7.37. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution which is $n(0, \theta)$. Then $Y = \sum X_i^2$ is a sufficient statistic for θ. Find the best statistic for θ^2.

7.38. In the notation of Example 2 of this section, is there a best statistic for $\Pr[-c \leq X \leq c]$? Here $c > 0$.

7.39. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with parameter $\theta > 0$. Find the best statistic for $\Pr(X \leq 1) = (1 + \theta)e^{-\theta}$. *Hint.* Let $u(x_1) = 1, x_1 \leq 1$, zero elsewhere, and find $E[u(X_1)|Y = y]$, where $Y = \sum_1^n X_i$. Make use of Example 2, p. 124.

7.8 The Case of Several Parameters

In many of the interesting problems we encounter, the p.d.f. may not depend upon a single parameter θ, but perhaps upon two (or more) parameters, say θ_1 and θ_2, where $(\theta_1, \theta_2) \in \Omega$, a two-dimensional parameter space. We now define joint sufficient statistics for the parameters. For the moment, we shall restrict ourselves to the case of two parameters. The student may wish to refer to p. 67, on which we define conditional probability density functions.

Definition 5. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that has p.d.f. $f(x; \theta_1, \theta_2)$, where $(\theta_1, \theta_2) \in \Omega$. Let $Y_1 = u_1(X_1, X_2, \ldots, X_n)$, $Y_2 = u_2(X_1, X_2, \ldots, X_n), \ldots, Y_n = u_n(X_1, X_2, \ldots, X_n)$ be n statistics such that the Jacobian of the associated transformation does not vanish identically. Let $g(y_1, y_2, \ldots, y_n; \theta_1, \theta_2)$ denote the joint p.d.f. of $Y_1, Y_2, \ldots, Y_n$, and let $g_{12}(y_1, y_2; \theta_1, \theta_2)$ be the joint p.d.f. of Y_1 and Y_2. The statistics Y_1 and Y_2 are called *joint sufficient statistics* for θ_1 and θ_2 if and only if the conditional p.d.f.

$$h(y_3, \ldots, y_n|y_1, y_2) = \frac{g(y_1, y_2, \ldots, y_n; \theta_1, \theta_2)}{g_{12}(y_1, y_2; \theta_1, \theta_2)}$$

of all other statistics $Y_3, \ldots, Y_n$, given $Y_1 = y_1$ and $Y_2 = y_2$, does not depend upon θ_1 or θ_2 or both θ_1 and θ_2, whatever be the fixed values y_1 and y_2.

As may be anticipated, there is a Fisher-Neyman criterion for the existence of joint sufficient statistics. In the notation of the preceding paragraph, this criterion can be stated as follows: The statistics $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ and $Y_2 = u_2(X_1, X_2, \ldots, X_n)$ are joint sufficient statistics for θ_1 and θ_2 if and only if

$$f(x_1; \theta_1, \theta_2)f(x_2; \theta_1, \theta_2)\cdots f(x_n; \theta_1, \theta_2)$$
$$= g_{12}[u_1(x_1, x_2, \ldots, x_n), u_2(x_1, x_2, \ldots, x_n); \theta_1, \theta_2]H(x_1, x_2, \ldots, x_n),$$

where, for fixed values of $y_1 = u_1(x_1, \ldots, x_n)$ and $y_2 = u_2(x_1, \ldots, x_n)$, the function $H(x_1, x_2, \ldots, x_n)$ does not depend upon both or either of θ_1 and θ_2.

Moreover, the factorization theorem can be extended. In our notation, it can be stated in the following manner. The statistics $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ and $Y_2 = u_2(X_1, X_2, \ldots, X_n)$ are joint sufficient statistics for the parameters θ_1 and θ_2 if and only if we can find two nonnegative functions k_1 and k_2 such that

$$f(x_1; \theta_1, \theta_2)f(x_2; \theta_1, \theta_2)\cdots f(x_n; \theta_1, \theta_2)$$
$$= k_1[u_1(x_1, x_2, \ldots, x_n), u_2(x_1, x_2, \ldots, x_n); \theta_1, \theta_2]k_2(x_1, x_2, \ldots, x_n)$$

where, for all fixed values of the functions $y_1 = u_1(x_1, x_2, \ldots, x_n)$ and $y_2 = u_2(x_1, x_2, \ldots, x_n)$, the function $k_2(x_1, x_2, \ldots, x_n)$ does not depend upon both or either of θ_1 and θ_2.

Example 1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution having p.d.f.

$$f(x; \theta_1, \theta_2) = \frac{1}{2\theta_2}, \qquad \theta_1 - \theta_2 < x < \theta_1 + \theta_2,$$
$$= 0 \text{ elsewhere},$$

where $-\infty < \theta_1 < \infty$, $0 < \theta_2 < \infty$. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics. The joint p.d.f. of Y_1 and Y_n is given by

$$g_{1n}(y_1, y_n; \theta_1, \theta_2) = \frac{n(n-1)}{(2\theta_2)^n}(y_n - y_1)^{n-2}, \qquad \theta_1 - \theta_2 < y_1 < y_n < \theta_1 + \theta_2,$$

and equals zero elsewhere. Accordingly the joint p.d.f. of $X_1, X_2, \ldots, X_n$ can be written, for points of positive probability density,

$$\left(\frac{1}{2\theta_2}\right)^n = \frac{n(n-1)[\max(x_i) - \min(x_i)]^{n-2}}{(2\theta_2)^n} \times \left(\frac{1}{n(n-1)[\max(x_i) - \min(x_i)]^{n-2}}\right).$$

Since the last factor does not depend upon the parameters, the Fisher-Neyman criterion assures us that Y_1 and Y_n are joint sufficient statistics for

θ_1 and θ_2. This result can also be established by showing that the conditional p.d.f. of $Y_2, Y_3, \ldots, Y_{n-1}$, given $Y_1 = y_1$ and $Y_n = y_n$, does not depend upon the parameters or it can be established by use of the factorization theorem.

The extension of the notion of joint sufficient statistics for more than two parameters is a natural one. Suppose that a certain p.d.f. depends upon m parameters. Let a random sample of size n be taken from the distribution that has this p.d.f. and define n statistics in such a way that the Jacobian of the associated transformation is not identically zero. Then m of these statistics are said to be joint sufficient statistics for the m parameters if and only if for fixed values of these m statistics, the conditional p.d.f. of the remaining statistics does not depend upon any or all of the m parameters. We remark that the criteria for the existence of sufficient statistics, namely, the Fisher-Neyman criterion and the factorization theorem, are readily extended.

There is an extension of the Rao-Blackwell theorem that can be adapted to joint sufficient statistics for several parameters, but that extension will not be included in this book. However, the concept of a complete family of probability density functions is generalized as follows: Let

$$\{f(v_1, v_2, \ldots, v_k; \theta_1, \theta_2, \ldots, \theta_m); (\theta_1, \theta_2, \ldots, \theta_m) \in \Omega\}$$

denote a family of probability density functions of k random variables $V_1, V_2, \ldots, V_k$ that depends upon m parameters $(\theta_1, \theta_2, \ldots, \theta_m) \in \Omega$. Let $u(v_1, v_2, \ldots, v_k)$ be a continuous function of $v_1, v_2, \ldots, v_k$ (but not a function of any or all of the parameters). If

$$E[u(V_1, V_2, \ldots, V_k)] = 0$$

for all $(\theta_1, \theta_2, \ldots, \theta_m) \in \Omega$ implies that $u(v_1, v_2, \ldots, v_k) = 0$ at all points $(v_1, v_2, \ldots, v_k)$ at which at least one member of the family of probability density functions is positive, we shall say that the family of probability density functions is a complete family.

The remainder of our treatment of the case of several parameters will be restricted to probability density functions that represent what we shall call regular cases of the exponential class. Let $X_1, X_2, \ldots, X_n$, $n > m$, denote a random sample from a distribution that depends on m parameters and has a p.d.f. of the form

$$(1) \quad f(x; \theta_1, \theta_2, \ldots, \theta_m) = \exp\left[\sum_{j=1}^{m} p_j(\theta_1, \theta_2, \ldots, \theta_m)K_j(x) + S(x) + q(\theta_1, \theta_2, \ldots, \theta_m)\right]$$

for $a < x < b$, and equals zero elsewhere.

A p.d.f. of the form (1) is said to be a member of the *exponential class* of probability density functions of the continuous type. If, in addition,

(i) neither a nor b depends upon any or all of the parameters $\theta_1, \theta_2, \ldots, \theta_m$,

(ii) the $p_j(\theta_1, \theta_2, \ldots, \theta_m)$, $j = 1, 2, \ldots, m$, are nontrivial, functionally independent, continuous functions of θ_j, $\gamma_j < \theta_j < \delta_j$, $j = 1, 2, \ldots, m$,

(iii) the $K_j'(x)$, $j = 1, 2, \ldots, m$, are continuous for $a < x < b$ and no one is a linear homogeneous function of the others,

(iv) $S(x)$ is a continuous function of x, $a < x < b$,

we say that we have a *regular case* of the exponential class.

The joint p.d.f. of $X_1, X_2, \ldots, X_n$ is given, at points of positive probability density, by

$$\exp\left[\sum_{j=1}^{m} p_j(\theta_1, \ldots, \theta_m) \sum_{i=1}^{n} K_j(x_i) + \sum_{i=1}^{n} S(x_i) + nq(\theta_1, \ldots, \theta_m)\right]$$
$$= \exp\left[\sum_{j=1}^{m} p_j(\theta_1, \ldots, \theta_m) \sum_{i=1}^{n} K_j(x_i) + nq(\theta_1, \ldots, \theta_m)\right] \exp\left[\sum_{i=1}^{n} S(x_i)\right].$$

In accordance with the factorization theorem, the statistics

$$Y_1 = \sum_{i=1}^{n} K_1(X_i), \qquad Y_2 = \sum_{i=1}^{n} K_2(X_i), \ldots, Y_m = \sum_{i=1}^{n} K_m(X_i)$$

are joint sufficient statistics for the m parameters $\theta_1, \theta_2, \ldots, \theta_m$. It is left as an exercise to prove that the joint p.d.f. of $Y_1, \ldots, Y_m$ is of the form

$$(2) \qquad R(y_1, \ldots, y_m) \exp\left[\sum_{j=1}^{m} p_j(\theta_1, \ldots, \theta_m) y_j + nq(\theta_1, \ldots, \theta_m)\right]$$

at points of positive probability density. These points of positive probability density and the function $R(y_1, \ldots, y_m)$ do not depend upon any or all of the parameters $\theta_1, \theta_2, \ldots, \theta_m$. Moreover, in accordance with a theorem in analysis, it can be asserted that, in a regular case of the exponential class, the family of probability density functions of these joint sufficient statistics $Y_1, Y_2, \ldots, Y_m$ is complete when $n > m$. In accordance with a convention previously adopted, we shall refer to $Y_1, Y_2, \ldots, Y_m$ *as joint complete sufficient statistics for the parameters* $\theta_1, \theta_2, \ldots, \theta_m$.

Example 2. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution which is $n(\theta_1, \theta_2)$, $-\infty < \theta_1 < \infty$, $0 < \theta_2 < \infty$. Thus the p.d.f. $f(x; \theta_1, \theta_2)$ of the distribution may be written as

$$f(x; \theta_1, \theta_2) = \exp\left[\frac{-1}{2\theta_2} x^2 + \frac{\theta_1}{\theta_2} x - \frac{\theta_1^2}{2\theta_2} - \ln \sqrt{2\pi\theta_2}\right].$$

Therefore, we can take $K_1(x) = x^2$ and $K_2(x) = x$. Consequently, the statistics

$$Y_1 = \sum_1^n X_i^2 \qquad \text{and} \qquad Y_2 = \sum_1^n X_i$$

are joint complete sufficient statistics for θ_1 and θ_2. Since the relations

$$Z_1 = \frac{Y_2}{n} = \bar{X}, \qquad Z_2 = \frac{Y_1 - Y_2^2/n}{n-1} = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

define a one-to-one transformation, Z_1 and Z_2 are also joint complete sufficient statistics for θ_1 and θ_2. Moreover,

$$E(Z_1) = \theta_1 \qquad \text{and} \qquad E(Z_2) = \theta_2.$$

From completeness, we have that Z_1 and Z_2 are the only continuous functions of Y_1 and Y_2 which are unbiased statistics for θ_1 and θ_2, respectively.

A p.d.f.

$$f(x; \theta_1, \theta_2, \ldots, \theta_m) = \exp\left[\sum_{j=1}^m p_j(\theta_1, \theta_2, \ldots, \theta_m)K_j(x) + S(x) + q(\theta_1, \theta_2, \ldots, \theta_m)\right], \quad x = a_1, a_2, a_3, \ldots,$$

zero elsewhere, is said to represent a regular case of the exponential class of probability density functions of the discrete type if

(i) the set $\{x; x = a_1, a_2, \ldots\}$ does not depend upon any or all of the parameters $\theta_1, \theta_2, \ldots, \theta_m$,

(ii) the $p_j(\theta_1, \theta_2, \ldots, \theta_m)$, $j = 1, 2, \ldots, m$, are nontrivial, functionally independent, continuous functions of θ_j, $\gamma_j < \theta_j < \delta_j$, $j = 1, 2, \ldots, m$,

(iii) the $K_j(x)$, $j = 1, 2, \ldots, m$, are nontrivial functions of x on the set $\{x; x = a_1, a_2, \ldots\}$ and no one is a linear function of the others.

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a discrete-type distribution which represents a regular case of the exponential class. Then the statements made above in connection with the random variable of the continuous type are also valid here.

Not always do we sample from a distribution of one random variable X. We could, for instance, sample from a distribution of two random variables V and W with joint p.d.f. $f(v, w; \theta_1, \theta_2, \ldots, \theta_m)$. Recall that by a random sample $(V_1, W_1), (V_2, W_2), \ldots, (V_n, W_n)$ from a distribution of this sort, we mean that the joint p.d.f. of these $2n$ random variables is given by

$$f(v_1, w_1; \theta_1, \ldots, \theta_m)f(v_2, w_2; \theta_1, \ldots, \theta_m)\cdots f(v_n, w_n; \theta_1, \ldots, \theta_m).$$

In particular, suppose that the random sample is taken from a distribution that has the p.d.f. of V and W of the exponential class

$$(3)\quad f(v, w; \theta_1, \ldots, \theta_m) = \exp\left[\sum_{j=1}^{m} p_j(\theta_1, \ldots, \theta_m)K_j(v, w) + S(v, w) + q(\theta_1, \ldots, \theta_m)\right]$$

for $a < v < b$, $c < w < d$, and equals zero elsewhere, where a, b, c, d do not depend on the parameters and conditions similar to (i), (ii), (iii), and (iv), p. 240, are imposed. Then the m statistics

$$Y_1 = \sum_{i=1}^{n} K_1(V_i, W_i), \qquad Y_2 = \sum_{i=1}^{n} K_2(V_i, W_i), \ldots,$$

$$Y_m = \sum_{i=1}^{n} K_m(V_i, W_i)$$

are joint complete sufficient statistics for the m parameters $\theta_1, \theta_2, \ldots, \theta_m$.

EXERCISES

7.40. Let $Y_1 < Y_2 < Y_3$ be the order statistics of a random sample of size 3 from the distribution with p.d.f.

$$f(x; \theta_1, \theta_2) = \frac{1}{\theta_2}\exp\left[-\frac{(x - \theta_1)}{\theta_2}\right],$$

$$\theta_1 < x < \infty, \quad -\infty < \theta_1 < \infty, \quad 0 < \theta_2 < \infty,$$

zero elsewhere. Find the joint p.d.f. of $Z_1 = Y_1$, $Z_2 = Y_2$, and $Z_3 = Y_1 + Y_2 + Y_3$. The corresponding transformation maps the space $\{(y_1, y_2, y_3); \theta_1 < y_1 < y_2 < y_3 < \infty\}$ onto the space $\{(z_1, z_2, z_3); \theta_1 < z_1 < z_2 < (z_3 - z_1)/2 < \infty\}$. Show that Z_1 and Z_3 are joint sufficient statistics for θ_1 and θ_2.

7.41. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that has a p.d.f. of form (1) of this section. Show that $Y_1 = \sum_{i=1}^{n} K_1(X_i), \ldots, Y_m = \sum_{i=1}^{n} K_m(X_i)$ have a joint p.d.f. of form (2) of this section.

7.42. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ denote a random sample of size n from a bivariate normal distribution with means μ_1 and μ_2, positive variances σ_1^2 and σ_2^2, and correlation coefficient ρ. Show that $\sum_1^n X_i$, $\sum_1^n Y_i$, $\sum_1^n X_i^2$, $\sum_1^n Y_i^2$, and $\sum_1^n X_iY_i$ are joint sufficient statistics for the five parameters. Are $\bar{X} = \sum_1^n X_i/n$, $\bar{Y} = \sum_1^n Y_i/n$, $S_1^2 = \sum_1^n (X_i - \bar{X})^2/n$, $S_2^2 = \sum_1^n (Y_i - \bar{Y})^2/n$,

and $\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y})/nS_1S_2$ also joint sufficient statistics for these parameters?

7.43. Let the p.d.f. $f(x; \theta_1, \theta_2)$ be of the form

$$\exp [p_1(\theta_1, \theta_2)K_1(x) + p_2(\theta_1, \theta_2)K_2(x) + S(x) + q(\theta_1, \theta_2)], \qquad a < x < b,$$

zero elsewhere. Let $K_1'(x) = cK_2'(x)$. Show that $f(x; \theta_1, \theta_2)$ can be written in the form

$$\exp [p(\theta_1, \theta_2)K(x) + S(x) + q_1(\theta_1, \theta_2)], \qquad a < x < b,$$

zero elsewhere. This is the reason why it is required that no one $K_j'(x)$ be a linear homogeneous function of the others, that is, so that the number of sufficient statistics equals the number of parameters.

7.9 Sufficiency, Completeness, and Stochastic Independence

We have seen that the properties of sufficiency and completeness can be combined to provide a most attractive theory of point estimation. We shall now investigate the impact that these two fundamental concepts have upon stochastic independence.

Theorem 7. *Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution having a p.d.f. $f(x; \theta)$, $\theta \in \Omega$, where Ω is an interval set. Let $Y_1 = u_1(X_1, X_2, \ldots, X_n)$ be a sufficient statistic for θ, and let the family $\{g_1(y_1; \theta); \theta \in \Omega\}$ of probability density functions of Y_1 be complete. Let $Z = u(X_1, X_2, \ldots, X_n)$ be any other statistic (not a function of Y_1 alone). If the distribution of Z does not depend upon θ, then Z is stochastically independent of the sufficient statistic Y_1.*

Proof. We shall prove a special case of this theorem. We assume that $f(x; \theta)$, $\theta \in \Omega$, is the p.d.f. of a random variable X of the continuous type and that the conditional p.d.f. of Z, given $Y_1 = y_1$, is a continuous function of y_1. Denote the joint p.d.f. of Y_1 and Z by $g(y_1, z; \theta)$; the conditional p.d.f. of Z, given $Y_1 = y_1$, by $h(z|y_1)$; and the p.d.f. of Z by $g_2(z)$. The integral

$$(1) \qquad \int_{-\infty}^{\infty} [g_2(z) - h(z|y_1)]g_1(y_1; \theta)\, dy_1 = 0$$

for all θ, $\theta \in \Omega$, because

$$\int_{-\infty}^{\infty} g_2(z)g_1(y_1; \theta)\, dy_1 = g_2(z)$$

and

$$\int_{-\infty}^{\infty} h(z|y_1)g_1(y_1; \theta)\, dy_1 = \int_{-\infty}^{\infty} \frac{g(y_1, z; \theta)}{g_1(y_1; \theta)} g_1(y_1; \theta)\, dy_1 = g_2(z).$$

Since Y_1 is a sufficient statistic for θ, $h(z|y_1)$ does not depend upon θ. From the hypothesis, $g_2(z)$ does not depend upon θ; thus $g_2(z) - h(z|y_1)$ does not depend upon θ. Moreover, $h(z|y_1)$, thus $g_2(z) - h(z|y_1)$, is a continuous function of y_1 by assumption. Since $\{g_1(y_1; \theta); \theta \in \Omega\}$ is complete, Equation (1) requires that

$$g_2(z) - h(z|y_1) = 0, \qquad \text{or} \qquad g_2(z) = h(z|y_1)$$

at all points y_1 where, for some θ, $g_1(y_1; \theta) > 0$. However,

$$g(y_1, z; \theta) = g_1(y_1; \theta)h(z|y_1), \qquad g_1(y_1; \theta) > 0,$$

so that

$$g(y_1, z; \theta) = g_1(y_1; \theta)g_2(z), \qquad g_1(y_1; \theta) > 0.$$

That is, Z is stochastically independent of the sufficient statistic Y_1, as was to be proved.

If it is assumed that a statistic Z is stochastically independent of a statistic Y_1, then, of course, $g_2(z) = h(z|y_1)$. It is interesting to observe that if Y_1 is a sufficient statistic for θ, then $h(z|y_1)$, and hence $g_2(z)$, does not depend upon θ whether $\{g_1(y_1; \theta); \theta \in \Omega\}$ is or is not complete. That is, the theorem may not be stated as an "if, and only if," condition for the stochastic independence of the statistic Z and the sufficient statistic Y_1. However, if we restrict $f(x; \theta)$ to represent a regular case (continuous or discrete) of the exponential class, and if $X_1, X_2, \ldots, X_n$ is a random sample from this distribution, then the family $\{g_1(y_1; \theta); \theta \in \Omega\}$ of probability density functions of the sufficient statistic $Y_1 = \sum_1^n K(X_i)$ is complete. In those cases the statistic Z is stochastically independent of the sufficient statistic Y_1 if, and only if, the distribution of Z does not depend upon θ.

It should be remarked that the theorem (including the special formulation of it for regular cases of the exponential class) extends immediately to probability density functions which involve m parameters for which there exist m joint sufficient statistics. For example, let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution having the p.d.f. $f(x; \theta_1, \theta_2)$ that represents a regular case of the exponential class such that there are two joint complete sufficient statistics for θ_1 and θ_2. Then any other statistic $Z = u(X_1, X_2, \ldots, X_n)$ is stochastically independent of the joint complete sufficient statistics if, and only if, the distribution of Z does not depend upon θ_1 or θ_2.

We now give three examples of Theorem 7. The first of these provides an alternative proof of the stochastic independence of $\bar{X}$ and S^2, the mean and the variance of a random sample of size n from a distribution which is $n(\mu, \sigma^2)$.

Example 1. Let the random variable X be $n(\mu, \sigma^2)$, and let $X_1, X_2, \ldots, X_n$ denote a random sample from this distribution. We shall show that the statistic S^2 is stochastically independent of the statistic $\bar{X}$. It has been seen (p. 232) that $\bar{X}$ is, for every fixed $\sigma^2 > 0$, a complete sufficient statistic for μ, $-\infty < \mu < \infty$. If the distribution of S^2 does not depend upon μ, then, in accordance with Theorem 7, S^2 is stochastically independent of $\bar{X}$. Now the distribution of S^2 is uniquely determined by its moment-generating function $M(t) = E(e^{tS^2})$. Our problem will be solved if we can show that $M(t)$ does not depend upon μ. Now

$$M(t) = E\left\{\exp\left[t\sum_1^n \frac{(X_i - \bar{X})^2}{n}\right]\right\}$$

$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \exp\left[\frac{t}{n}\sum_1^n (x_i - \bar{x})^2\right]\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n$$

$$\times \exp\left[-\sum_1^n \frac{(x_i - \mu)^2}{2\sigma^2}\right] dx_1\, dx_2 \cdots dx_n$$

exists provided $t < n/2\sigma^2$. This is a fairly complicated integral, and we shall make no effort to evaluate it. But by the following argument it is easy to see that $M(t)$ does not depend upon μ. Consider the one-to-one transformation $w_1 = x_1 - \mu, \ldots, w_n = x_n - \mu$, which maps the entire $x_1x_2\cdots x_n$ space onto the entire $w_1w_2\cdots w_n$ space. The Jacobian $J = 1$. Under this transformation

$$\sum_1^n (x_i - \mu)^2 = \sum_1^n w_i^2$$

and

$$\sum_1^n (x_i - \bar{x})^2 = \sum_1^n (w_i - \bar{w})^2,$$

where $\bar{w} = (w_1 + w_2 + \cdots + w_n)/n$. Thus,

$$M(t) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \exp\left[\frac{t}{n}\sum_1^n (w_i - \bar{w})^2\right]$$

$$\times \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_1^n w_i^2\right) dw_1 \cdots dw_n,$$

and $M(t)$ obviously does not depend upon μ. Accordingly, the variance S^2 of the random sample from the distribution $n(\mu, \sigma^2)$ is stochastically independent of the mean $\bar{X}$ of the sample.

Example 2. Let X_1, X_2, X_3, X_4 denote a random sample from a distribution having p.d.f.

$$f(x; \theta) = \frac{1}{\theta}, \qquad 0 < x < \theta, 0 < \theta < \infty,$$
$$= 0 \text{ elsewhere.}$$

Let $Y_1 < Y_2 < Y_3 < Y_4$ denote the order statistics of this random sample. Exercise 7.19 (p. 228), shows that $Y_4 = \max(X_i)$ is a complete sufficient statistic for θ. It will now be shown that the statistic $Z = (Y_1 + Y_2 + Y_3 + Y_4)/Y_4$ is stochastically independent of the sufficient statistic Y_4. The moment-generating function of Z is

$$M(t) = E\left[\exp\left(t\frac{Y_1 + Y_2 + Y_3 + Y_4}{Y_4}\right)\right]$$
$$= \int_0^{\theta}\int_0^{y_4}\int_0^{y_3}\int_0^{y_2} \exp\left(t\frac{y_1 + y_2 + y_3 + y_4}{y_4}\right)\left(\frac{4!}{\theta^4}\right) dy_1\, dy_2\, dy_3\, dy_4.$$

The one-to-one transformation $w_1 = y_1/\theta$, $w_2 = y_2/\theta$, $w_3 = y_3/\theta$, $w_4 = y_4/\theta$ has $|J| = \theta^4$, and it maps $0 < y_1 < y_2 < y_3 < y_4 < \theta$ onto $0 < w_1 < w_2 < w_3 < w_4 < 1$. Consequently,

$$M(t) = \int_0^1\int_0^{w_4}\int_0^{w_3}\int_0^{w_2} \exp\left(t\frac{w_1 + w_2 + w_3 + w_4}{w_4}\right)(4!)\, dw_1\, dw_2\, dw_3\, dw_4,$$

which clearly does not depend upon θ. Thus the distribution of Z does not depend upon θ, and so Z is stochastically independent of Y_4, the complete sufficient statistic for θ.

Example 3. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution which is $n(\theta_1, \theta_2)$, $-\infty < \theta_1 < \infty$, $0 < \theta_2 < \infty$. In Example 2, p. 240, it was proved that the mean $\bar{X}$ and the variance S^2 of the sample are joint complete sufficient statistics for θ_1 and θ_2. Consider the statistic

$$Z = \frac{\sum_1^{n-1}(X_{i+1} - X_i)^2}{\sum_1^n (X_i - \bar{X})^2}.$$

The one-to-one transformation $y_i = (x_i - \theta_1)/\sqrt{\theta_2}$, $i = 1, 2, \ldots, n$, may be used to show that $E(e^{tZ})$ does not depend upon θ_1, θ_2. That is, Z is stochastically independent of both $\bar{X}$ and S^2.

EXERCISES

7.44. Let X_1, X_2 be a random sample of size 2 from a distribution having the p.d.f. $f(x; \theta) = (1/\theta)e^{-x/\theta}$, $0 < x < \infty$, $0 < \theta < \infty$, zero elsewhere. Show that $Z = X_1/(X_1 + X_2)$ has a moment-generating function that does not depend upon θ. Thus $Y = X_1 + X_2$, a complete sufficient statistic for θ, is stochastically independent of Z.

7.45. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample from the normal distribution $n(\theta, \sigma^2)$, $-\infty < \theta < \infty$. Show that the moment-generating function of $Z = Y_n - \bar{Y}$ does not depend upon θ. Thus $\bar{Y} = \sum_1^n Y_i/n$, a complete sufficient statistic for θ, is stochastically independent of Z.

7.46. Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution $n(\theta, \sigma^2)$, $-\infty < \theta < \infty$. Prove that a necessary and sufficient condition that the statistics $Z = \sum_1^n a_i X_i$ and $Y = \sum_1^n X_i$, a complete sufficient statistic for θ, be stochastically independent is that $\sum_1^n a_i = 0$.

7.47. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from the distribution having p.d.f. $f(x; \theta) = e^{-(x-\theta)}$, $\theta < x < \infty$, $-\infty < \theta < \infty$, zero elsewhere. (a) Show that $Z = \sum_1^n (Y_i - Y_1)$ and the complete sufficient statistic Y_1 for θ are stochastically independent. (b) Write $\sum_1^n (Y_i - \theta) = n(Y_1 - \theta) + \sum_1^n (Y_i - Y_1)$. Find the moment-generating functions of $\sum_1^n (Y_i - \theta)$ and $n(Y_1 - \theta)$. Use the fact that Y_1 and $\sum_1^n (Y_i - Y_1)$ are stochastically independent to find the moment-generating function of $\sum_1^n (Y_i - Y_1)$.

7.48. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from a distribution that has p.d.f. $f(x; \theta) = 1/\theta$, $0 < x < \theta$, $0 < \theta < \infty$, zero elsewhere. Show that Y_1/Y_n and the complete sufficient statistic Y_n for θ are stochastically independent.

7.49. Let X and Y be random variables such that $E(X^k)$ and $E(Y^k) \neq 0$ exist for $k = 1, 2, 3, \ldots$. If the ratio X/Y and its denominator Y are stochastically independent, prove that $E[(X/Y)^k] = E(X^k)/E(Y^k)$, $k = 1, 2, 3, \ldots$. *Hint.* Write $E(X^k) = E[Y^k(X/Y)^k]$.

7.50. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from a distribution that has p.d.f. $f(x; \theta) = (1/\theta)e^{-x/\theta}$, $0 < x < \infty$, $0 < \theta < \infty$, zero elsewhere. Show that the ratio $R = nY_1/\sum_1^n Y_i$ and its denominator (a complete sufficient statistic for θ) are stochastically independent. Use the result of the preceding exercise to determine $E(R^k)$, $k = 1, 2, 3, \ldots$.

7.51. Let $X_1, X_2, \ldots, X_5$ be a random sample of size 5 from the distribution that has p.d.f. $f(x) = e^{-x}$, $0 < x < \infty$, zero elsewhere. Show that $(X_1 + X_2)/(X_1 + X_2 + \cdots + X_5)$ and its denominator are stochastically independent. *Hint.* The p.d.f. $f(x)$ is a member of $\{f(x; \theta); 0 < \theta < \infty\}$, where $f(x; \theta) = (1/\theta)e^{-x/\theta}$, $0 < x < \infty$, zero elsewhere.

Chapter 8

Further Topics in Point Estimation

8.1 The Rao-Cramér Inequality

In chapter 7 we used the problem of point estimation of a parameter θ to motivate the study of sufficient statistics. In that chapter we arbitrarily adopted the principles of unbiasedness and minimum variance as being the most desirable properties that a statistic should have if that statistic is to be called a "best" statistic for the parameter θ. Not always, however, will there exist such a best statistic. In this chapter we shall investigate other principles of point estimation. But first we wish to establish an important inequality.

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution with p.d.f. $f(x; \theta)$, $\theta \in \Omega = \{\theta; \gamma < \theta < \delta\}$, where γ and δ are known. Let $Y = u(X_1, X_2, \ldots, X_n)$ be an unbiased statistic for θ. We shall show that the variance of Y, say σ_Y^2, satisfies the inequality

$$\sigma_Y^2 \geq \frac{1}{nE\{[\partial \ln f(X; \theta)/\partial\theta]^2\}}. \tag{1}$$

Throughout this chapter, unless otherwise specified, it will be assumed that we may differentiate, with respect to a parameter, under an integral or a summation symbol. This means, among other things, that the domain of positive probability density does not depend upon θ. In the terminology of Chapter 7, we are dealing with regular cases, although not necessarily regular cases of the exponential class.

We shall now give a proof of Inequality (1) when X is a random variable of the continuous type. The student can easily handle the

discrete case by changing integrals to sums. Let $g(y; \theta)$ denote the p.d.f. of the unbiased statistic Y. We are given that

$$1 = \int_{-\infty}^{\infty} f(x_i; \theta)\, dx_i, \qquad i = 1, 2, \ldots, n,$$

and

$$\begin{aligned}\theta &= \int_{-\infty}^{\infty} y g(y; \theta)\, dy \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_1, x_2, \ldots, x_n) f(x_1; \theta) \cdots f(x_n; \theta)\, dx_1 \cdots dx_n.\end{aligned}$$

The final form of the right-hand member of the second equation is justified by the discussion on p. 155. If we differentiate both members of each of these equations with respect to θ, we have

$$(2)\qquad \begin{aligned} 0 &= \int_{-\infty}^{\infty} \frac{\partial f(x_i; \theta)}{\partial \theta}\, dx_i = \int_{-\infty}^{\infty} \frac{\partial \ln f(x_i; \theta)}{\partial \theta} f(x_i; \theta)\, dx_i, \\ 1 &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_1, x_2, \cdots, x_n) \left[\sum_1^n \frac{1}{f(x_i; \theta)} \frac{\partial f(x_i; \theta)}{\partial \theta}\right] \\ &\qquad \times f(x_1; \theta) \cdots f(x_n; \theta)\, dx_1 \cdots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_1, x_2, \ldots, x_n) \left[\sum_1^n \frac{\partial \ln f(x_i; \theta)}{\partial \theta}\right] \\ &\qquad \times f(x_1; \theta) \cdots f(x_n; \theta)\, dx_1 \cdots dx_n. \end{aligned}$$

Define the random variable Z by $Z = \sum_1^n [\partial \ln f(X_i; \theta)/\partial \theta]$. In accordance with the first of Equations (2) we have $E[Z] = \sum_1^n E[\partial \ln f(X_i; \theta)/\partial \theta] = 0$. Moreover, Z is the sum of n mutually stochastically independent random variables each with mean zero and consequently with variance $E\{[\partial \ln f(X; \theta)/\partial \theta]^2\}$. Hence the variance of Z is the sum of the n variances, namely,

$$\sigma_Z^2 = nE\left[\left(\frac{\partial \ln f(X; \theta)}{\partial \theta}\right)^2\right].$$

Because $Y = u(X_1, \ldots, X_n)$ and $Z = \sum_1^n [\partial \ln f(X_i; \theta)/\partial \theta]$, the second of Equations (2) shows that $E[YZ] = 1$. Recall (p. 69) that

$$E[YZ] = E[Y]E[Z] + \rho \sigma_Y \sigma_Z,$$

where ρ is the correlation coefficient of Y and Z. Since $E[Y] = \theta$ and $E[Z] = 0$ we have

$$1 = \theta \cdot 0 + \rho \sigma_Y \sigma_Z \qquad \text{or} \qquad \rho = \frac{1}{\sigma_Y \sigma_Z}.$$

Now $\rho^2 \leq 1$. Hence

$$\frac{1}{\sigma_Y^2 \sigma_Z^2} \leq 1, \qquad \text{or} \qquad \frac{1}{\sigma_Z^2} \leq \sigma_Y^2.$$

If we replace σ_Z^2 by its value, we have Inequality (1), namely,

$$\sigma_Y^2 \geq \frac{1}{nE\left[\left(\frac{\partial \ln f(X; \theta)}{\partial \theta}\right)^2\right]}.$$

Inequality (1) is known as the Rao-Cramér inequality. It provides, in regular cases, a lower bound on the variance of an unbiased statistic for a parameter. We call this lower bound the Rao-Cramér lower bound.

We now make the following definitions.

Definition 1. Let Y be an unbiased statistic for a parameter θ in a regular case of point estimation. The statistic Y is called an efficient statistic for θ if and only if the variance of Y attains the Rao-Cramér lower bound.

It is left as an exercise to show, in regular cases of point estimation, that $E\{[\partial \ln f(X; \theta)/\partial\theta]^2\} = -E[\partial^2 \ln f(X; \theta)/\partial\theta^2]$. In some instances the latter is much easier to compute.

Remark. If, in a regular case of point estimation, one assumes that a certain unbiased statistic for a parameter has the Rao-Cramér lower bound for its variance, one can *prove* that this statistic must first of all be a sufficient statistic for the parameter. That is, the class of efficient statistics can be proved to be a subclass of the class of sufficient statistics.

Definition 2. In a regular case of point estimation, the ratio of the Rao-Cramér lower bound, to the actual variance of any unbiased statistic for a parameter is called the *efficiency* of that statistic.

Example 1. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a Poisson distribution that has the mean $\theta > 0$. It is known that $Y = \sum_1^n X_i$ is a sufficient statistic for θ. We shall show that $Y/n = \bar{X}$ is an efficient statistic for θ. We have

$$\frac{\partial \ln f(x; \theta)}{\partial \theta} = \frac{\partial}{\partial \theta}[x \ln \theta - \theta - \ln x!]$$
$$= \frac{x}{\theta} - 1 = \frac{(x - \theta)}{\theta}.$$

Accordingly,

$$E\left[\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2\right] = \frac{E(X-\theta)^2}{\theta^2} = \frac{\sigma^2}{\theta^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta}.$$

The Rao-Cramér lower bound in this case is $1/[n(1/\theta)] = \theta/n$. But θ/n is the variance $\sigma_{\bar{X}}^2$ of $\bar{X}$. Hence $\bar{X}$ is an efficient statistic for θ.

Example 2. Let S^2 denote the variance of a random sample of size $n > 1$ from a distribution which is $n(\mu, \theta)$, $0 < \theta < \infty$. We know that $E[nS^2/(n-1)] = \theta$. What is the efficiency of the statistic $nS^2/(n-1)$? We have

$$\ln f(x;\theta) = -\frac{(x-\mu)^2}{2\theta} - \frac{\ln(2\pi\theta)}{2},$$

$$\frac{\partial \ln f(x;\theta)}{\partial\theta} = \frac{(x-\mu)^2}{2\theta^2} - \frac{1}{2\theta},$$

and

$$\frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2} = -\frac{(x-\mu)^2}{\theta^3} + \frac{1}{2\theta^2}.$$

Accordingly,

$$-E\left[\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right] = \frac{\theta}{\theta^3} - \frac{1}{2\theta^2} = \frac{1}{2\theta^2}.$$

Thus the Rao-Cramér lower bound is $2\theta^2/n$. Now nS^2/θ is $\chi^2(n-1)$, so that the variance of nS^2/θ is $2(n-1)$. Accordingly, the variance of $nS^2/(n-1)$ is $2(n-1)[\theta^2/(n-1)^2] = 2\theta^2/(n-1)$. Thus, the efficiency of the statistic $nS^2/(n-1)$ is $(n-1)/n$.

Example 3. Let $X_1, X_2, \ldots, X_n$ denote a random sample of size $n > 2$ from a distribution with p.d.f.

$$f(x;\theta) = \theta x^{\theta-1} = \exp[\theta \ln x - \ln x + \ln\theta], \qquad 0 < x < 1,$$
$$= 0 \text{ elsewhere.}$$

It is easy to verify that the Rao-Cramér lower bound is θ^2/n. Let $Y_i = -\ln X_i$. We shall indicate that each Y_i has a gamma distribution. The associated transformation $y_i = -\ln x_i$, with inverse $x_i = e^{-y_i}$, is one-to-one and the transformation maps the space $\{x_i; 0 < x_i < 1\}$ onto the space $\{y_i; 0 < y_i < \infty\}$. We have $|J| = e^{-y_i}$. Thus Y_i has a gamma distribution with $\alpha = 1$ and $\beta = 1/\theta$. Let $Z = -\sum_1^n \ln X_i$. Then Z is a sufficient statistic for θ and Z has a gamma distribution with $\alpha = n$ and $\beta = 1/\theta$. Accordingly, we have $E[Z] = \alpha\beta = n/\theta$. This suggests that we compute the expectation of $1/Z$ to see if we can find an unbiased statistic for θ. A simple integration shows that $E[1/Z] = \theta/(n-1)$. Hence $(n-1)/Z$ is an unbiased statistic for θ. With $n > 2$, the variance of $(n-1)/Z$ exists and is found to be $\theta^2/(n-2)$, so that the efficiency of $(n-1)/Z$ is $(n-2)/n$. This efficiency tends to one as

n increases. In such an instance, the statistic is said to be *asymptotically efficient.*

The concept of joint efficient statistics for several parameters has been developed along with the associated concept of joint efficiency of several statistics. But limitations of space prevent their inclusion in this book.

We have investigated the notion of a sufficient statistic and that of an efficient statistic. We shall now define a consistent statistic.

Definition 3. Any statistic which converges stochastically to a parameter θ is called a *consistent* statistic for that parameter θ.

For example, let S_n^2 denote the variance of a random sample of size n from a distribution which is $n(\mu, \sigma^2)$. In Exercise 5.8 (p. 178) it was proved that the unbiased statistic for σ^2, namely, $nS_n^2/(n-1)$, converges stochastically to σ^2. Thus $nS_n^2/(n-1)$ is an unbiased consistent statistic for σ^2. In Exercise 5.17 (p. 181) it was proved that S_n^2 also converges stochastically to σ^2. Thus S_n^2 is a biased consistent statistic for σ^2.

A reexamination of Theorem 1, p. 176, shows that an equivalent definition in terms of probabilities is that a statistic Y_n is a consistent statistic for a parameter θ if and only if

$$\lim_{n\to\infty} \Pr(|Y_n - \theta| \geq \epsilon) = 0 \qquad \text{for every } \epsilon > 0.$$

If Y_n is an unbiased statistic for θ, we have the following theorem.

Theorem 1. *An unbiased statistic Y_n for the parameter θ is a consistent statistic for θ if the variance $\sigma_{Y_n}^2$ of Y_n approaches zero as the sample size becomes infinite.*

Proof. In accordance with Chebyshev's inequality we have, for every $k > 0$,

$$\Pr[|Y_n - \theta| \geq k\sigma_{Y_n}] \leq \frac{1}{k^2}.$$

Let $\epsilon > 0$ denote an arbitrary positive number and take $k = \epsilon/\sigma_{Y_n}$. Accordingly, $\Pr[|Y_n - \theta| \geq \epsilon] \leq \sigma_{Y_n}^2/\epsilon^2$. Since $\lim_{n\to\infty} \sigma_{Y_n}^2 = 0$, then, for every $\epsilon > 0$, we have the following limit:

$$\lim_{n\to\infty} \Pr[|Y_n - \theta| \geq \epsilon] = 0.$$

This proves the theorem.

The student will show in Exercise 8.7 that if $\lim_{n\to\infty} E(Y_n) = \theta$ and if $\lim_{n\to\infty} \sigma_{Y_n}^2 = 0$, then Y_n is a consistent statistic for θ.

EXERCISES

8.1. Prove that $\bar{X}$, the mean of a random sample of size n from a distribution which is $n(\theta, \sigma^2)$, $-\infty < \theta < \infty$, is, for every fixed $\sigma^2 > 0$, an efficient statistic for θ.

8.2. Show that the mean $\bar{X}$ of a random sample of size n from a distribution which is $b(1, \theta)$, $0 < \theta < 1$, is an efficient statistic for θ.

8.3. Given $f(x; \theta) = 1/\theta$, $0 < x < \theta$, zero elsewhere, with $\theta > 0$, formally compute the reciprocal of

$$nE\left\{\left[\frac{\partial \ln f(X; \theta)}{\partial \theta}\right]^2\right\}.$$

Compare this with the variance of $(n + 1)Y_n/n$, where Y_n is the largest item of a random sample of size n from this distribution. Comment.

8.4. Given the p.d.f.

$$f(x; \theta) = \frac{1}{\pi[1 + (x - \theta)^2]}, \qquad -\infty < x < \infty, -\infty < \theta < \infty.$$

Show that the Rao-Cramér lower bound is $2/n$, where n is the size of a random sample from this Cauchy distribution.

8.5. Show, in a regular case of point estimation, that

$$E\left\{\left[\frac{\partial \ln f(X; \theta)}{\partial \theta}\right]^2\right\} = -E\left[\frac{\partial^2 \ln f(X; \theta)}{\partial \theta^2}\right].$$

8.6. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution with p.d.f. $f(x; \theta_1, \theta_2, \ldots, \theta_r)$, $(\theta_1, \ldots, \theta_r) \in \Omega$. The expectation $E(X^k)$ may be called the kth moment of the distribution, say M_k, $k = 1, 2, \ldots$. The sum $\sum_1^n X_i^k/n$ is called the kth moment of the sample, say m_k, $k = 1, 2, \ldots$. A method of point estimation, called "the method of moments," can be described as follows: Equate M_k to the experimental value of m_k, beginning with $k = 1$ and continuing until there are enough equations to provide unique solutions for $\theta_1, \theta_2, \ldots, \theta_r$, say $\theta_i = h_i(m_1, m_2, \ldots)$, $i = 1, 2, \ldots, r$. Subject to certain conditions on the function h_i, the random variable $h_i(m_1, m_2, \ldots)$, where $m_k = \sum_1^n X_i^k/n$, is a consistent statistic for θ_i. In each of the following problems, find the functions h_i such that $\theta_i = h_i(m_1, m_2, \ldots)$ and show that $h_i(m_1, m_2, \ldots)$ is a consistent statistic for θ_i: (a) X is Poisson with parameter $\theta > 0$; (b) X is $n(\theta_1, \theta_2)$; (c) X has the p.d.f. $f(x; \theta) = \theta x^{\theta - 1}$, $0 < x < 1$, zero elsewhere, where $0 < \theta$.

8.7. Let Y_n be a statistic such that $\lim_{n\to\infty} E(Y_n) = \theta$ and $\lim_{n\to\infty} \sigma_{Y_n}^2 = 0$. Prove that Y_n is a consistent statistic for θ.

8.2 Maximum Likelihood Estimation of Parameters

In this section we present a very general method of point estimation which is applicable whether the regularity conditions are or are not satisfied. This method of finding point estimates of parameters is called the *method of maximum likelihood*. Consider a random sample $X_1, X_2, \ldots, X_n$ from a distribution having p.d.f. $f(x; \theta)$, $\theta \in \Omega$. The joint p.d.f. of $X_1, X_2, \ldots, X_n$ is $f(x_1; \theta)f(x_2; \theta)\cdots f(x_n; \theta)$. This joint p.d.f. may be regarded as a function of θ. When so regarded, it is called the likelihood function L of the random sample, and we write

$$L(\theta; x_1, x_2, \ldots, x_n) = f(x_1; \theta)f(x_2; \theta)\cdots f(x_n; \theta), \qquad \theta \in \Omega.$$

Suppose we can find a nontrivial function of $x_1, x_2, \ldots, x_n$, say $u(x_1, x_2, \ldots, x_n)$, such that, when θ is replaced by $u(x_1, x_2, \ldots, x_n)$, the likelihood function L is a maximum. That is, $L[u(x_1, x_2, \ldots, x_n); x_1, x_2, \ldots, x_n]$ is at least as great as $L(\theta; x_1, x_2, \ldots, x_n)$ for every $\theta \in \Omega$. Then the statistic $u(X_1, X_2, \ldots, X_n)$ will be called a *maximum likelihood statistic* for θ. A maximum likelihood statistic for θ will be denoted by the symbol $\hat{\theta} = u(X_1, X_2, \ldots, X_n)$. We remark that in many instances there will be a unique maximum likelihood statistic $\hat{\theta}$ for a parameter θ, and often it may be obtained by the process of differentiation. A few examples should help clarify these definitions and statements.

Example 1. Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution $n(\theta, 1)$, $-\infty < \theta < \infty$. Here

$$L(\theta; x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\sum_1^n (x_i - \theta)^2/2\right].$$

This function L can be maximized by setting the first derivative of L, with respect to θ, equal to zero and solving the resulting equation for θ. We note, however, that each of the functions L and $\ln L$ is a maximum for the same value θ. So, it may be easier to solve

$$\frac{d \ln L(\theta; x_1, x_2, \ldots, x_n)}{d\theta} = 0.$$

For this example,

$$\frac{d \ln L(\theta; x_1, x_2, \ldots, x_n)}{d\theta} = \sum_1^n (x_i - \theta).$$

If this derivative is equated to zero, the solution for θ is $u(x_1, x_2, \ldots, x_n) = \sum_1^n x_i/n$. That $\sum_1^n x_i/n$ actually maximizes L is easily seen. Thus the statistic

$$\hat{\theta} = u(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_1^n X_i = \overline{X}$$

is the unique maximum likelihood statistic for the mean θ. In this instance the maximum likelihood statistic is unbiased, sufficient, efficient, and consistent.

Example 2. Let

$$f(x; \theta) = \frac{1}{\theta}, \qquad 0 < x \leq \theta, 0 < \theta < \infty,$$
$$= 0 \text{ elsewhere},$$

and let $X_1, X_2, \ldots, X_n$ denote a random sample from this distribution. Note that we have taken $0 < x \leq \theta$ instead of $0 < x < \theta$ so as to avoid a discussion of supremum versus maximum. Here

$$L(\theta; x_1, x_2, \ldots, x_n) = \frac{1}{\theta^n}, \qquad 0 < x_i \leq \theta,$$

which is an ever-decreasing function of θ. The maximum of such functions cannot be found by differentiation but by selecting θ as small as possible. Now $\theta \geq$ each x_i; in particular, then, $\theta \geq \max(x_i)$. Thus L can be made no larger than

$$\frac{1}{[\max(x_i)]^n}$$

and the unique maximum likelihood statistic $\hat{\theta}$ for θ in this example is the nth order statistic $\max(X_i)$. We know that $E[\max(X_i)] = n\theta/(n+1)$. Thus, in this instance, the maximum likelihood statistic for the parameter θ is biased. That is, the property of unbiasedness is not in general a property of a maximum likelihood statistic.

Example 3. Let

$$f(x; \theta) = 1, \qquad \theta - \tfrac{1}{2} \leq x \leq \theta + \tfrac{1}{2}, -\infty < \theta < \infty,$$
$$= 0 \text{ elsewhere},$$

and let $X_1, X_2, \ldots, X_n$ denote a random sample from this distribution. Then

$$L(\theta; x_1, x_2, \ldots, x_n) = 1, \qquad \theta - \tfrac{1}{2} \leq x_i \leq \theta + \tfrac{1}{2},$$
$$= 0 \text{ elsewhere}.$$

Thus L attains its maximum provided

$$\theta - \tfrac{1}{2} \leq \min(x_i) \quad \text{and} \quad \max(x_i) \leq \theta + \tfrac{1}{2},$$

or when

$$\theta \leq \min(x_i) + \tfrac{1}{2} \quad \text{and} \quad \max(x_i) - \tfrac{1}{2} \leq \theta.$$

So, every statistic $u(X_1, X_2, \ldots, X_n)$ such that

$$\max(X_i) - \tfrac{1}{2} \leq u(X_1, X_2, \ldots, X_n) \leq \min(X_i) + \tfrac{1}{2}$$

is a maximum likelihood statistic for θ. The length of the random interval $[\max(X_i) - \frac{1}{2}, \min(X_i) + \frac{1}{2}]$ is $1 + \min(X_i) - \max(X_i)$. If b is a positive proper fraction, the statistic

$$\max(X_i) - \tfrac{1}{2} + b[1 + \min(X_i) - \max(X_i)]$$
$$= b \min(X_i) + (1 - b) \max(X_i) + b - \tfrac{1}{2}$$

is in this interval. Accordingly, for each b, $0 < b < 1$, this statistic is a maximum likelihood statistic for the parameter θ. One such statistic is $[\min(X_i) + \max(X_i)]/2$. Thus uniqueness is not in general a property of a maximum likelihood statistic for a parameter.

One of the important properties of the method of maximum likelihood is contained in the following theorem.

Theorem 2. *Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that has p.d.f. $f(x; \theta)$, $\theta \in \Omega$. If a sufficient statistic $Y = u(X_1, X_2, \ldots, X_n)$ for θ exists, and if a maximum likelihood statistic $\hat{\theta}$ for θ also exists uniquely, then $\hat{\theta}$ is a function of $Y = u(X_1, X_2, \ldots, X_n)$.*

Proof. Let $g(y; \theta)$ be the p.d.f. of the sufficient statistic Y. Then, by the Fisher-Neyman criterion,

$$L(\theta; x_1, x_2, \ldots, x_n) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta)$$
$$= g[u(x_1, x_2, \ldots, x_n); \theta] H(x_1, x_2, \ldots, x_n),$$

where $H(x_1, x_2, \ldots, x_n)$ does not depend on θ. That is, in maximizing L with respect to θ, $H(x_1, x_2, \ldots, x_n)$ plays the role of a constant, so that each of the functions L and $g[u(x_1, x_2, \ldots, x_n); \theta]$ is a maximum simultaneously. Apart from trivial solutions (that is, θ a constant) any θ that maximizes $g[u(x_1, x_2, \ldots, x_n); \theta]$ will be a function of $u(x_1, x_2, \ldots, x_n)$. Thus the maximum likelihood statistic $\hat{\theta}$ is a function of the sufficient statistic $Y = u(X_1, X_2, \ldots, X_n)$. This important fact was illustrated in Examples 1 and 2 of this section.

The preceding definitions and properties are easily generalized. Let $X, Y, \ldots, Z$ denote random variables which may or may not be stochastically independent and which may or may not be identically distributed. Let the joint p.d.f. $g(x, y, \ldots, z; \theta_1, \theta_2, \ldots, \theta_m)$, $(\theta_1, \theta_2, \ldots, \theta_m) \in \Omega$, depend on m parameters. This joint p.d.f., when regarded as a function of $(\theta_1, \theta_2, \ldots, \theta_m) \in \Omega$, is called the likelihood function of the random variables. Those functions $u_1(x, y, \ldots, z)$, $u_2(x, y, \ldots, z), \ldots, u_m(x, y, \ldots, z)$ which maximize this likelihood function with respect to $\theta_1, \theta_2, \ldots, \theta_m$, respectively, define maximum likelihood statistics

$$\hat{\theta}_1 = u_1(X, Y, \ldots, Z), \quad \hat{\theta}_2 = u_2(X, Y, \ldots, Z), \ldots,$$
$$\hat{\theta}_m = u_m(X, Y, \ldots, Z)$$

for the m parameters. If joint sufficient statistics for the parameters exist and if the maximum likelihood statistics exist, then these maximum likelihood statistics are functions of the joint sufficient statistics. An illustrative example follows.

Example 4. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $n(\theta_1, \theta_2)$, $-\infty < \theta_1 < \infty$, $0 < \theta_2 < \infty$. We shall find $\hat{\theta}_1$ and $\hat{\theta}_2$, the maximum likelihood statistics for θ_1 and θ_2. The logarithm of the likelihood function may be written in the form

$$\ln L(\theta_1, \theta_2; x_1, \ldots, x_n) = -\frac{\sum_1^n (x_i - \theta_1)^2}{2\theta_2} - \frac{n \ln (2\pi\theta_2)}{2}.$$

We observe that we may maximize by differentiation. We have

$$\frac{\partial \ln L}{\partial \theta_1} = \frac{\sum_1^n (x_i - \theta_1)}{\theta_2}, \qquad \frac{\partial \ln L}{\partial \theta_2} = \frac{\sum_1^n (x_i - \theta_1)^2}{2\theta_2^2} - \frac{n}{2\theta_2}.$$

If we equate these partial derivatives to zero and solve simultaneously the two equations thus obtained, the solutions for θ_1 and θ_2 are found to be $\sum_1^n x_i/n = \bar{x}$ and $\sum_1^n (x_i - \bar{x})^2/n = s^2$, respectively. It can be verified that these solutions maximize L. Thus the maximum likelihood statistics for θ_1 and θ_2 are, respectively, the mean and the variance of the sample, namely, $\hat{\theta}_1 = \bar{X}$ and $\hat{\theta}_2 = S^2$. These are two joint sufficient statistics for θ_1 and θ_2. Whereas $\hat{\theta}_1$ is an unbiased statistic for θ_1, the statistic $\hat{\theta}_2$ is biased. It has previously been shown that $\hat{\theta}_1 = \bar{X}$ and $\hat{\theta}_2 = S^2$ converge stochastically to θ_1 and θ_2, respectively. In this example, we can say that the statistics $\bar{X}$ and S^2 are joint sufficient and consistent statistics for θ_1 and θ_2. In all cases of practical interest, consistency is a property of a maximum likelihood statistic.

EXERCISES

8.8. Let $X_1, X_2, \ldots, X_n$ represent a random sample from each of the distributions having the following probability density functions:

(a) $f(x; \theta) = \theta^x e^{-\theta}/x!, x = 0, 1, 2, \ldots, 0 \leq \theta < \infty$, zero elsewhere, where $f(0; 0) = 1$.

(b) $f(x; \theta) = \theta x^{\theta - 1}, 0 < x < 1, 0 < \theta < \infty$, zero elsewhere.

(c) $f(x; \theta) = (1/\theta)e^{-x/\theta}, 0 < x < \infty, 0 < \theta < \infty$, zero elsewhere.

(d) $f(x; \theta) = \frac{1}{2}e^{-|x-\theta|}, -\infty < x < \infty, -\infty < \theta < \infty$.

(e) $f(x; \theta) = e^{-(x-\theta)}, \theta \leq x < \infty, -\infty < \theta < \infty$, zero elsewhere.

(f) $f(x; \theta) = \theta^x(1 - \theta)^{1-x}, x = 0, 1, 0 \leq \theta \leq 1$, zero elsewhere, where $f(0; 0) = f(1, 1) = 1$.

In each case find the maximum likelihood statistic $\hat{\theta}$ for θ. If a sufficient statistic Y also exists, express $\hat{\theta}$ as a function of Y.

8.9. Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution having p.d.f. $f(x; \theta_1, \theta_2) = (1/\theta_2)e^{-(x-\theta_1)/\theta_2}$, $\theta_1 \leq x < \infty$, $-\infty < \theta_1 < \infty$, $0 < \theta_2 < \infty$, zero elsewhere. Find the maximum likelihood statistics for θ_1 and θ_2.

8.10. (Madow) Prove the following: Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution having p.d.f. $f(x; \theta)$, $\theta \in \Omega$. Let the likelihood function $L(\theta) = f(x_1; \theta)f(x_2; \theta)\cdots f(x_n; \theta)$. Let $Y = u(X_1, X_2, \ldots, X_n)$ be a sufficient statistic for the parameter θ. If $g(y; \theta_0)$ is the p.d.f. of the sufficient statistic Y in the special case $\theta = \theta_0$, then the p.d.f. $g(y; \theta)$ of the sufficient statistic Y for every $\theta \in \Omega$ is given by

$$g(y; \theta) = g(y; \theta_0)\frac{L(\theta)}{L(\theta_0)}.$$

This result proves useful in certain difficult distribution problems. The result shows that if we can find the p.d.f. of a sufficient statistic when the parameter is assigned a special value, we can find the p.d.f. of the sufficient statistic for all values of the parameter. This will be illustrated with the distribution of the sample correlation coefficient in Chapter 10. The symbol θ may represent any number of parameters.

8.3 Decision Functions

In this section we shall discuss the problem of point estimation of a parameter from a different standpoint. Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution that has the p.d.f. $f(x; \theta)$, $\theta \in \Omega$. The distribution may be either of the continuous or the discrete type. Let $Y = u(X_1, X_2, \ldots, X_n)$ be a statistic (sufficient, if one exists) on which we wish to base a point estimate of the parameter θ. Let $w(y)$ be that function of the observed value of the statistic Y which is the point estimate of θ. Thus the function w *decides* the value of our point estimate of θ and w is called a *decision function* or a *decision rule*. One value of the decision function, say $w(y)$, is called a *decision*. Thus a numerically determined point estimate of a parameter θ is a decision. Now a decision may be correct or it may be wrong. It would be useful to have a measure of the seriousness of the difference, if any, between the true value of θ and the point estimate $w(y)$. Accordingly, with each pair, $[\theta, w(y)]$, $\theta \in \Omega$, we associate a nonnegative number $\mathscr{L}[\theta, w(y)]$ which reflects this seriousness. We call the function $\mathscr{L}$ the *loss function*. The expected (mean) value of the loss function is called the *risk function*. If $g(y; \theta)$, $\theta \in \Omega$, is the p.d.f. of Y, the risk function $R(\theta, w)$ is given by

$$R(\theta, w) = E\{\mathscr{L}[\theta, w(Y)]\} = \int_{-\infty}^{\infty} \mathscr{L}[\theta, w(y)]g(y; \theta)\, dy,$$

if Y is a random variable of the continuous type. It would be desirable to select a decision function that minimizes the risk $R(\theta, w)$ for all values of θ, $\theta \in \Omega$. But this is usually impossible because the decision function w which minimizes $R(\theta, w)$ for one value of θ may not minimize $R(\theta, w)$ for another value of θ. Accordingly, we need either to restrict our decision function to a certain class or to consider methods of ordering the risk functions. The following example, while very simple, dramatizes these difficulties.

Example 1. Let $X_1, X_2, \ldots, X_{25}$ be a random sample from a distribution that is $n(\theta, 1)$, $-\infty < \theta < \infty$. Let $Y = \bar{X}$, the mean of the random sample, and let $\mathscr{L}[\theta, w(y)] = [\theta - w(y)]^2$. We will compare the two decision functions given by $w_1(y) = y$ and $w_2(y) = 0$ for $-\infty < y < \infty$. The corresponding risk functions are

$$R(\theta, w_1) = E[(\theta - Y)^2] = \tfrac{1}{25}$$

and

$$R(\theta, w_2) = E[(\theta - 0)^2] = \theta^2.$$

Obviously, if, in fact, $\theta = 0$, then $w_2(y) = 0$ is an excellent decision and we have $R(0, w_2) = 0$. However, if θ differs from zero by very much, it is equally clear that $w_2(y) = 0$ is a poor decision. For example, if, in fact, $\theta = 2$, $R(2, w_2) = 4 > R(2, w_1) = \frac{1}{25}$. In general, we see that $R(\theta, w_2) < R(\theta, w_1)$, provided $-\frac{1}{5} < \theta < \frac{1}{5}$, and that otherwise $R(\theta, w_2) \geq R(\theta, w_1)$. That is, one of these decision functions is better than the other for some values of θ and the other decision function is better for other values of θ. If, however, we had restricted our consideration to decision functions w such that $E[w(Y)] = \theta$ for all values of θ, $\theta \in \Omega$, then the decision $w_2(y) = 0$ is not allowed. Under this restriction and with the given $\mathscr{L}[\theta, w(y)]$, the risk function is the variance of the unbiased statistic. Thus we know from the results of Chapter 7, that the decision function given by $w_1(y) = y = \bar{x}$ is the best possible. Suppose, however, we do not want to restrict ourselves to decision functions w such that $E[w(Y)] = \theta$ for all values of θ, $\theta \in \Omega$. Instead, let us say that the decision function that minimizes the maximum of the risk function is the best decision function. Because, in this example, $R(\theta, w_2) = \theta^2$ is unbounded, $w_2(y) = 0$ is not, in accordance, with this criterion, a good decision function. On the other hand, with $-\infty < \theta < \infty$, we have

$$\max_{\theta} R(\theta, w_1) = \max_{\theta} \left(\tfrac{1}{25}\right) = \tfrac{1}{25}.$$

Accordingly, $w_1(y) = y = \bar{x}$ seems to be a very good decision in accordance with this criterion because $\frac{1}{25}$ is small. As a matter of fact, it can be proved that w_1 is the best decision function, as measured by this ***minimax criterion***, when the loss function is $\mathscr{L}[\theta, w(y)] = [\theta - w(y)]^2$.

In this example we illustrated the following:

(a) Without some restriction on the decision function, it is difficult to find a decision function that has a risk function which is uniformly less than the risk function of another decision function.

(b) A principle of selecting a best decision function which is called the *minimax principle.* This principle may be stated as follows: If the decision function given by $w_0(y)$ is such that, for all $\theta \in \Omega$,

$$\max_{\theta} R[\theta, w_0(y)] \leq \max_{\theta} R[\theta, w(y)]$$

for every other decision function $w(y)$, then $w_0(y)$ is called a *minimax decision function.*

With the restriction $E[w(Y)] = \theta$ and the loss function $\mathscr{L}[\theta, w(y)] = [\theta - w(y)]^2$, the decision function which minimizes the risk function yields an unbiased statistic with minimum variance. If, however, the restriction $E[w(Y)] = \theta$ is replaced by some other condition, the decision function $w(Y)$, if it exists, which minimizes $E\{[\theta - w(Y)]^2\}$ uniformly in θ is sometimes called the *minimum mean-square-error* statistic. Exercises 8.12, 8.13, and 8.14 provide examples of this type of statistic.

Another principle for selecting the decision function which may be called a best decision function will be stated in Section 8.4.

EXERCISES

8.11. In Example 1 of this section, take $\mathscr{L}[\theta, w(y)] = |\theta - w(y)|$. Show that $R(\theta, w_1) = \frac{1}{5}\sqrt{2/\pi}$ and $R(\theta, w_2) = |\theta|$. Of these two decision functions w_1 and w_2, which yields the smaller maximum risk?

8.12. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a Poisson distribution with parameter θ, $0 < \theta < \infty$. Let $Y = \sum_1^n X_i$ and let $\mathscr{L}[\theta, w(y)] = [\theta - w(y)]^2$. If we restrict our considerations to decision functions of the form $w(y) = b + y/n$, where b does not depend upon y, show that $R(\theta, w) = b^2 + \theta/n$. What decision function of this form yields a uniformly smaller risk than every other decision function of this form? With this solution, say w, and $0 < \theta < \infty$, determine $\max_{\theta} R(\theta, w)$ if it exists.

8.13. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $n(\mu, \theta)$, $0 < \theta < \infty$, where μ is unknown. Let $Y = \sum_1^n (X_i - \bar{X})^2/n = S^2$ and let $\mathscr{L}[\theta, w(y)] = [\theta - w(y)]^2$. If we consider decision functions of the form $w(y) = by$, where b does not depend upon y, show that $R(\theta, w) = (\theta^2/n^2)[(n^2 - 1)b^2 - 2n(n - 1)b + n^2]$. Show that $b = n/(n + 1)$ yields a

minimum risk for decision functions of this form. Note that $nY/(n+1)$ is not an unbiased statistic for θ. With $w(y) = ny/(n+1)$ and $0 < \theta < \infty$, determine $\max_{\theta} R(\theta, w)$ if it exists.

8.14. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $b(1, \theta)$, $0 \leq \theta \leq 1$. Let $Y = \sum_1^n X_i$ and let $\mathscr{L}[\theta, w(y)] = [\theta - w(y)]^2$. Consider decision functions of the form $w(y) = by$, where b does not depend upon y. Prove that $R(\theta, w) = b^2 n\theta(1-\theta) + (bn-1)^2\theta^2$. Show that

$$\max_{\theta} R(\theta, w) = \frac{b^4 n^2}{4[b^2 n - (bn-1)^2]},$$

provided the value b is such that $b^2 n \geq 2(bn-1)^2$. Prove that $b = 1/n$ does not minimize $\max_{\theta} R(\theta, w)$.

8.4 Bayesian Procedures

In Section 6.5 we introduced a principle of statistical inference that we called Bayesian statistics. Recall, in the notation of that section, that the conditional p.d.f. of the random variable Θ, given the statistic $Y = y$, namely,

$$k(\theta|y) = \frac{h(\theta)g(y|\theta)}{k_1(y)}, \qquad \theta \in \Omega,$$

was the basis for our inference. We determined two functions, $u(y)$ and $v(y)$, such that the probability $\Pr[u(y) < \Theta < v(y)|Y = y]$ was close to one. The interval $u(y)$ to $v(y)$ was then used as an interval estimate of θ.

Suppose, however, that we want a point estimate of θ instead of an interval estimate. From the Bayesian viewpoint, this really amounts to selecting a decision function w so that $w(y)$ is a predicted value of θ (an experimental value of the random variable Θ) when both the computed value y and the conditional p.d.f. $k(\theta|y)$ are known. Now, in general, how would we predict an experimental value of any random variable, say W, if we want our prediction to be "reasonably close" to the value to be observed? Many statisticians would predict the mean, $E(W)$, of the distribution of W; others would predict a median (perhaps unique) of the distribution of W; some would predict a mode (perhaps unique) of the distribution of W; and some would have other predictions. However, it seems desirable that the choice of the decision function should depend upon the loss function $\mathscr{L}[\theta, w(y)]$. One way in which this dependence upon the loss function can be reflected is to

select the decision function w in such a way that the conditional expectation of the loss is a minimum. A *Bayes' solution* is a decision function w that minimizes

$$E\{\mathscr{L}[\Theta, w(y)]|Y = y\} = \int_{-\infty}^{\infty} \mathscr{L}[\theta, w(y)]k(\theta|y)\, d\theta,$$

if Θ is a random variable of the continuous type. The usual modification of the right-hand member of this equation is made for random variables of the discrete type. If, for example, the loss function is given by $\mathscr{L}[\theta, w(y)] = [\theta - w(y)]^2$, the Bayes' solution is given by $w(y) = E(\Theta|y)$, the mean of the conditional distribution of Θ, given $Y = y$. This follows from the fact (Exercise 1.81) that $E[(W - b)^2]$, if it exists, is a minimum when $b = E(W)$. If the loss function is given by $\mathscr{L}[\theta, w(y)] = |\theta - w(y)|$, then a median of the conditional distribution of Θ, given $Y = y$, is the Bayes' solution. This follows from the fact (Exercise 1.71) that $E(|W - b|)$, if it exists, is a minimum when b is equal to any median of the distribution of W.

The conditional expectation of the loss, given $Y = y$, defines a random variable which is a function of the statistic Y. The expected value of that function of Y, in the notation of this section, is given by

$$\int_{-\infty}^{\infty} \left\{\int_{-\infty}^{\infty} \mathscr{L}[\theta, w(y)]k(\theta|y)\, d\theta\right\} k_1(y)\, dy$$
$$= \int_{-\infty}^{\infty} \left\{\int_{-\infty}^{\infty} \mathscr{L}[\theta, w(y)]g(y|\theta)\, dy\right\} h(\theta)\, d\theta,$$

in the continuous case. The integral within the braces in the latter expression is, for every given $\theta \in \Omega$, the risk function $R(\theta, w)$; accordingly, the latter expression is the mean value of the risk, or the expected risk. Because a Bayes' solution minimizes

$$\int_{-\infty}^{\infty} \mathscr{L}[\theta, w(y)]k(\theta|y)\, d\theta$$

for every y for which $k_1(y) > 0$, it is evident that a Bayes' solution $w(y)$ minimizes this mean value of the risk. We now give an illustrative example.

Example 1. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution which is $b(1, \theta)$, $0 < \theta < 1$. We seek a decision function w which is a Bayes' solution. If $Y = \sum_1^n X_i$, then Y is $b(n, \theta)$. That is, the conditional p.d.f. of Y, given $\Theta = \theta$, is

$$g(y|\theta) = \binom{n}{y}\theta^y(1 - \theta)^{n-y}, \qquad y = 0, 1, \ldots, n,$$
$$= 0 \text{ elsewhere.}$$

We take the prior p.d.f. of the random variable Θ to be

$$h(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}, \qquad 0 < \theta < 1,$$
$$= 0 \text{ elsewhere.}$$

where α and β are assigned positive constants. Thus the joint p.d.f. of Y and Θ is given by $g(y|\theta)h(\theta)$ and the marginal p.d.f. of Y is

$$\begin{aligned} k_1(y) &= \int_0^1 h(\theta)g(y|\theta)\,d\theta \\ &= \binom{n}{y} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta^{y+\alpha-1}(1 - \theta)^{n-y+\beta-1}\,d\theta \\ &= \binom{n}{y} \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + y)\Gamma(n + \beta - y)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(n + \alpha + \beta)}, \qquad y = 0, 1, 2, \ldots, n, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Finally, the conditional p.d.f. of Θ, given $Y = y$, is, at points of positive probability density,

$$\begin{aligned} k(\theta|y) &= \frac{g(y|\theta)h(\theta)}{k_1(y)} \\ &= \frac{\Gamma(n + \alpha + \beta)}{\Gamma(\alpha + y)\Gamma(n + \beta - y)} \theta^{\alpha+y-1}(1 - \theta)^{\beta+n-y-1}, \qquad 0 < \theta < 1, \end{aligned}$$

and $y = 0, 1, \ldots, n$. We take the loss function to be $\mathscr{L}[\theta, w(y)] = [\theta - w(y)]^2$. Because Y is a random variable of the discrete type, whereas Θ is of the continuous type, we have for the expected risk

$$\int_0^1 \left\{ \sum_{y=0}^{n} [\theta - w(y)]^2 \binom{n}{y} \theta^y (1 - \theta)^{n-y} \right\} h(\theta)\,d\theta$$
$$= \sum_{y=0}^{n} \left\{ \int_0^1 [\theta - w(y)]^2 k(\theta|y)\,d\theta \right\} k_1(y).$$

The Bayes' solution $w(y)$ is the mean of the conditional distribution of Θ, given $Y = y$. Thus

$$\begin{aligned} w(y) &= \int_0^1 \theta k(\theta|y)\,d\theta \\ &= \frac{\Gamma(n + \alpha + \beta)}{\Gamma(\alpha + y)\Gamma(n + \beta - y)} \int_0^1 \theta^{\alpha+y}(1 - \theta)^{\beta+n-y-1}\,d\theta \\ &= \frac{\alpha + y}{\alpha + \beta + n}. \end{aligned}$$

This decision function $w(y)$ minimizes

$$\int_0^1 [\theta - w(y)]^2 k(\theta|y)\,d\theta$$

for $y = 0, 1, \ldots, n$ and, accordingly, it minimizes the expected risk.

EXERCISES

8.15. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution which is $n(\theta, \sigma^2)$, $-\infty < \theta < \infty$, where σ^2 is a given positive number. Let $Y = \bar{X}$, the mean of the random sample. Take the loss function to be $\mathscr{L}[\theta, w(y)] = |\theta - w(y)|$. If θ is an observed value of the random variable Θ which is $n(\mu, \tau^2)$, where $\tau^2 > 0$ and μ are known numbers, find the Bayes' solution $w(y)$ for a point estimate of θ.

8.16. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a Poisson distribution with mean θ, $0 < \theta < \infty$. Let $Y = \sum_1^n X_i$ and take the loss function to be $\mathscr{L}[\theta, w(y)] = [\theta - w(y)]^2$. Let θ be an observed value of the random variable Θ. If Θ has the p.d.f. $h(\theta) = \theta^{\alpha-1}e^{-\theta/\beta}/\Gamma(\alpha)\beta^\alpha$, $0 < \theta < \infty$, zero elsewhere, where $\alpha > 0$, $\beta > 0$ are known numbers, find the Bayes' solution $w(y)$ for a point estimate of θ.

8.17. Let Y_n be the nth order statistic of a random sample of size n from a distribution with p.d.f. $f(x|\theta) = 1/\theta$, $0 < x < \theta$, zero elsewhere. Take the loss function to be $\mathscr{L}[\theta, w(y_n)] = [\theta - w(y_n)]^2$. Let θ be an observed value of the random variable Θ, which has p.d.f. $h(\theta) = \beta\alpha^\beta/\theta^{\beta+1}$, $\alpha < \theta < \infty$, zero elsewhere, with $\alpha > 0$, $\beta > 0$. Find the Bayes' solution $w(y_n)$ for a point estimate of θ.

Chapter 9

Statistical Hypotheses

9.1 Some Examples and Definitions

The two principal areas of statistical inference are the areas of estimation of parameters and of tests of statistical hypotheses. The problem of estimation of parameters, both point and interval estimation, has been treated. In this chapter, some aspects of statistical hypotheses and tests of statistical hypotheses will be considered. The subject will be introduced by way of example.

Example 1. Let it be known that the outcome X of a random experiment is $n(\theta, 100)$. For instance, X may denote a score on a test, which score we assume to be normally distributed with mean θ and variance 100. Let us say that past experience with this random experiment indicates that $\theta = 75$. Suppose, owing possibly to some research in the area pertaining to this experiment, some changes are made in the method of performing this random experiment. It is then suspected that no longer does $\theta = 75$ but that now $\theta > 75$. There is as yet no formal experimental evidence that $\theta > 75$; hence the statement $\theta > 75$ is a conjecture or a *statistical hypothesis*. In admitting that the statistical hypothesis $\theta > 75$ may be false, we allow, in effect, the possibility that $\theta \leq 75$. Thus there are actually two statistical hypotheses. First, that the unknown parameter $\theta \leq 75$; that is, there has been no increase in θ. Second, that the unknown parameter $\theta > 75$. Accordingly, the parameter space is $\Omega = \{\theta; -\infty < \theta < \infty\}$. We denote the first of these hypotheses by the symbols H_0: $\theta \leq 75$ and the second by the symbols H_1: $\theta > 75$. Since the values $\theta > 75$ are alternatives to those where $\theta \leq 75$, the hypothesis H_1: $\theta > 75$ is called the *alternative hypothesis*. Needless to say, H_0 could be called the alternative to H_1; however, the conjecture, here $\theta > 75$, which is

made by the research worker is usually taken to be the alternative hypothesis. In any case, the problem is to decide which of these hypotheses is to be accepted. To reach a decision, the random experiment is to be repeated a number of independent times, say n, and the results observed. That is, we consider a random sample $X_1, X_2, \ldots, X_n$ from a distribution which is $n(\theta, 100)$, and we devise a rule which will tell us what decision to make once the experimental values, say $x_1, x_2, \ldots, x_n$, have been determined. Such a rule is called a *test* of the hypothesis H_0: $\theta \leq 75$ against the alternative hypothesis H_1: $\theta > 75$. There is no bound on the number of rules or tests which can be constructed. We shall consider three such tests. Our tests will be constructed around the following notion. We shall partition the sample space $\mathscr{A}$ into a subset C and its complement C^*. If the experimental values of $X_1, X_2, \ldots, X_n$, say $x_1, x_2, \ldots, x_n$, are such that the point $(x_1, x_2, \ldots, x_n) \in C$, we shall reject the hypothesis H_0 (accept the hypothesis H_1). If we have $(x_1, x_2, \ldots, x_n) \in C^*$, we shall accept the hypothesis H_0 (reject the hypothesis H_1).

Test 1. Let $n = 25$. The sample space $\mathscr{A}$ is the set

$$\{(x_1, x_2, \ldots, x_{25}); -\infty < x_i < \infty, i = 1, 2, \ldots, 25\}.$$

Let the subset C of the sample space be

$$C = \{(x_1, x_2, \ldots, x_{25}); x_1 + x_2 + \cdots + x_{25} > (25)(75)\}.$$

We shall reject the hypothesis H_0 if and only if our 25 experimental values are such that $(x_1, x_2, \ldots, x_{25}) \in C$. If $(x_1, x_2, \ldots, x_{25})$ is not an element of C, we shall accept the hypothesis H_0. This subset C of the sample space that leads to the rejection of the hypothesis H_0: $\theta \leq 75$ is called the *critical region* of Test 1. Now $\sum_1^{25} x_i > (25)(75)$ if and only if $\bar{x} > 75$, where $\bar{x} = \sum_1^{25} x_i/25$. Thus we can much more conveniently say that we shall reject the hypothesis H_0: $\theta \leq 75$ and accept the hypothesis H_1: $\theta > 75$ if and only if the experimentally determined value of the sample mean $\bar{x}$ is greater than 75. If $\bar{x} \leq 75$, we accept the hypothesis H_0: $\theta \leq 75$. Our test then amounts to this: We shall reject the hypothesis H_0: $\theta \leq 75$ if the mean of the sample exceeds the maximum value of the mean of the distribution when the hypothesis H_0 is true.

It would help us to evaluate a test of a statistical hypothesis if we knew the probability of rejecting that hypothesis (and hence of accepting the alternative hypothesis). In our Test 1, this means that we want to compute the probability

$$\Pr[(X_1, \ldots, X_{25}) \in C] = \Pr(\bar{X} > 75).$$

Obviously, this probability is a function of the parameter θ and we shall denote it by $K_1(\theta)$. The function $K_1(\theta) = \Pr(\bar{X} > 75)$ is called the *power*

function of Test 1, and the value of the power function at a parameter point is called the *power* of Test 1 at that point. Because $\bar{X}$ is $n(\theta, 4)$, we have

$$K_1(\theta) = \Pr\left(\frac{\bar{X} - \theta}{2} > \frac{75 - \theta}{2}\right) = 1 - N\left(\frac{75 - \theta}{2}\right).$$

So, for illustration, we have, by Table III of the Appendix, the power at $\theta = 75$ to be $K_1(75) = 0.500$. Other powers are $K_1(73) = 0.159$, $K_1(77) = 0.841$, and $K_1(79) = 0.977$. The graph of $K_1(\theta)$ of Test 1 is depicted in Figure 9.1. Among other things, this means that, if $\theta = 75$, the probability of

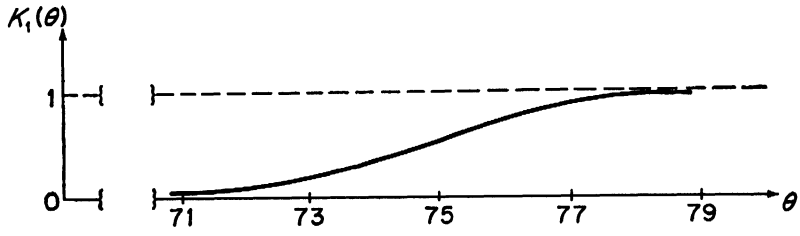


FIGURE 9.1

rejecting the hypothesis H_0: $\theta \leq 75$ is $\frac{1}{2}$. That is, if $\theta = 75$ so that H_0 is true, the probability of rejecting this true hypothesis H_0 is $\frac{1}{2}$. Many statisticians and research workers find it very undesirable to have such a high probability as $\frac{1}{2}$ assigned to this kind of mistake; namely, the rejection of H_0 when H_0 is a true hypothesis. Thus, Test 1 does not appear to be a very satisfactory test. Let us try to devise another test which does not have this objectionable feature. We shall do this by making it more difficult to reject the hypothesis H_0, with the hope that this will give a smaller probability of rejecting H_0 when that hypothesis is true.

Test 2. Let $n = 25$. We shall reject the hypothesis H_0: $\theta \leq 75$ and accept the hypothesis H_1: $\theta > 75$ if and only if $\bar{x} > 78$. Here the critical region is $C = \{(x_1, \ldots, x_{25}); x_1 + \cdots + x_{25} > (25)(78)\}$. The power function of Test 2 is, because $\bar{X}$ is $n(\theta, 4)$,

$$K_2(\theta) = \Pr(\bar{X} > 78) = 1 - N\left(\frac{78 - \theta}{2}\right).$$

Some values of the power function of Test 2 are $K_2(73) = 0.006$, $K_2(75) = 0.067$, $K_2(77) = 0.309$, and $K_2(79) = 0.691$. That is, if $\theta = 75$, the probability of rejecting H_0: $\theta \leq 75$ is 0.067; this is much more desirable than the corresponding probability $\frac{1}{2}$ which resulted from Test 1. However, if H_0 is false and, in fact, $\theta = 77$, the probability of rejecting H_0: $\theta \leq 75$ (and hence of accepting H_1: $\theta > 75$) is only 0.309. In certain instances, this low probability 0.309 of a correct decision (the acceptance of H_1 when H_1 is true) is objectionable. That is, Test 2 is not wholly satisfactory. Perhaps

we can overcome the undesirable features of Test 1 and Test 2 if we proceed as in Test 3.

Test 3. Let us first select a power function $K_3(\theta)$ which has the features of a small value at $\theta = 75$ and a large value at $\theta = 77$. For instance, take $K_3(75) = 0.159$ and $K_3(77) = 0.841$. To determine a test with such a power function, let us reject H_0: $\theta \leq 75$ if and only if the experimental value $\bar{x}$ of the mean of a random sample of size n is greater than some constant c. Thus the critical region is $C = \{(x_1, x_2, \ldots, x_n); x_1 + x_2 + \cdots + x_n > nc\}$. It should be noted that the sample size n and the constant c have not been determined as yet. However, since $\bar{X}$ is $n(\theta, 100/n)$, the power function is

$$K_3(\theta) = \Pr(\bar{X} > c) = 1 - N\left(\frac{c - \theta}{10/\sqrt{n}}\right).$$

The conditions $K_3(75) = 0.159$ and $K_3(77) = 0.841$ require

$$1 - N\left(\frac{c - 75}{10/\sqrt{n}}\right) = 0.159, \qquad 1 - N\left(\frac{c - 77}{10/\sqrt{n}}\right) = 0.841.$$

Equivalently, from Table III of the Appendix, we have

$$\frac{c - 75}{10/\sqrt{n}} = 1, \qquad \frac{c - 77}{10/\sqrt{n}} = -1.$$

The solution to these two equations in n and c is $n = 100$, $c = 76$. With these values of n and c, other powers of Test 3 are $K_3(73) = 0.001$ and $K_3(79) = 0.999$. It is important to observe that although Test 3 has a more desirable power function than those of Test 1 and Test 2, a certain "price" has been paid; namely, a sample size of $n = 100$ is required in Test 3, whereas we had $n = 25$ in the earlier tests.

We have now given illustrations of the following concepts:

(i) A statistical hypothesis.

(ii) A test of an hypothesis against an alternative hypothesis and the associated concept of the critical region of the test.

(iii) The power of a test.

These concepts will now be formally defined.

Definition 1. A *statistical hypothesis* is an assertion about the distribution of one or more random variables. If the statistical hypothesis completely specifies the distribution, it is called a *simple statistical hypothesis*; if it does not, it is called a *composite statistical hypothesis.*

If we refer to Example 1, we see that both H_0: $\theta \leq 75$ and H_1: $\theta > 75$ are composite statistical hypotheses since neither of them completely

specifies the distribution. If there, instead of H_0: $\theta \leq 75$, we had H_0: $\theta = 75$, then H_0 would have been a simple statistical hypothesis.

Definition 2. A *test* of a statistical hypothesis is a rule which, when the experimental sample values have been obtained, leads to a decision to accept or to reject the hypothesis under consideration.

Definition 3. Let C be that subset of the sample space which, in accordance with a prescribed test, leads to the rejection of the hypothesis under consideration. Then C is called the *critical region* of the test.

Definition 4. The *power function* of a test of a statistical hypothesis H_0 against an alternative hypothesis H_1 is that function, defined for all distributions under consideration, which yields the probability that the sample point falls in the critical region C of the test, that is, a function which yields the probability of rejecting the hypothesis under consideration. The value of the power function at a parameter point is called the *power* of the test at that point.

Definition 5. Let H_0 denote an hypothesis which is to be tested against an alternative hypothesis H_1 in accordance with a prescribed test. The *significance level* of the test (or the *size* of the critical region C) is the maximum value (actually supremum) of the power function of the test when H_0 is true.

If again we refer to Example 1, we see that the significance levels of Tests 1, 2, and 3 of that example are 0.500, 0.067, and 0.159, respectively. An additional example may help clarify these definitions.

Example 2. It is known that the random variable X has a p.d.f. of the form

$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}, \qquad 0 < x < \infty,$$
$$= 0 \text{ elsewhere.}$$

It is desired to test the simple hypothesis H_0: $\theta = 2$ against the alternative simple hypothesis H_1: $\theta = 4$. Thus $\Omega = \{\theta; \theta = 2, 4\}$. A random sample X_1, X_2 of size $n = 2$ will be used. The test to be used is defined by taking the critical region to be $C = \{(x_1, x_2); 9.5 \leq x_1 + x_2 < \infty\}$. The power function of the test and the significance level of the test will be determined.

There are but two probability density functions under consideration, namely, $f(x; 2)$ specified by H_0 and $f(x; 4)$ specified by H_1. Thus the power function is defined at but two points $\theta = 2$ and $\theta = 4$. The power function

of the test is given by $\Pr[(X_1, X_2) \in C]$. If H_0 is true, that is, $\theta = 2$, the joint p.d.f. of X_1 and X_2 is

$$f(x_1; 2)f(x_2; 2) = \tfrac{1}{4}e^{-(x_1+x_2)/2}, \qquad 0 < x_1 < \infty, 0 < x_2 < \infty,$$
$$= 0 \text{ elsewhere,}$$

and

$$\Pr[(X_1, X_2) \in C] = 1 - \Pr[(X_1, X_2) \in C^*]$$
$$= 1 - \int_0^{9.5} \int_0^{9.5-x_2} \tfrac{1}{4}e^{-(x_1+x_2)/2}\, dx_1\, dx_2$$
$$= 0.05, \text{ approximately.}$$

If H_1 is true, that is, $\theta = 4$, the joint p.d.f. of X_1 and X_2 is

$$f(x_1; 4)f(x_2; 4) = \tfrac{1}{16}e^{-(x_1+x_2)/4}, \qquad 0 < x_1 < \infty, 0 < x_2 < \infty,$$
$$= 0 \text{ elsewhere,}$$

and

$$\Pr[(X_1, X_2) \in C] = 1 - \int_0^{9.5} \int_0^{9.5-x_2} \tfrac{1}{16}e^{-(x_1+x_2)/4}\, dx_1\, dx_2$$
$$= 0.31, \text{ approximately.}$$

Thus the power of the test is given by 0.05 for $\theta = 2$ and by 0.31 for $\theta = 4$. That is, the probability of rejecting H_0 when H_0 is true is 0.05, and the probability of rejecting H_0 when H_0 is false is 0.31. Since the significance level of this test (or the size of the critical region) is the power of the test when H_0 is true, the significance level of this test is 0.05.

The fact that the power of this test, when $\theta = 4$, is only 0.31 immediately suggests that a search be made for another test which, with the same power when $\theta = 2$, would have a power greater than 0.31 when $\theta = 4$. However, Section 9.2 will make clear that such a search would be fruitless. That is, there is no test with a significance level of 0.05 and based on a random sample of size $n = 2$ that has a greater power at $\theta = 4$. The only manner in which the situation may be improved is to have recourse to a random sample of size n greater than 2.

Our computations of the powers of this test at the two points $\theta = 2$ and $\theta = 4$ were purposely done the hard way to focus attention on fundamental concepts. A procedure that is computationally simpler is the following. When the hypothesis H_0 is true, the random variable X is $\chi^2(2)$. Thus, the random variable $X_1 + X_2 = Y$, say, is $\chi^2(4)$. Accordingly, the power of the test when H_0 is true is given by

$$\Pr(Y \geq 9.5) = 1 - \Pr(Y < 9.5) = 1 - 0.95 = 0.05,$$

from Table II of the Appendix. When the hypothesis H_1 is true, the random variable $X/2$ is $\chi^2(2)$; so the random variable $(X_1 + X_2)/2 = Z$, say, is $\chi^2(4)$. Accordingly, the power of the test when H_1 is true is given by

$$\Pr(X_1 + X_2 \geq 9.5) = \Pr(Z \geq 4.75)$$
$$= \int_{4.75}^{\infty} \tfrac{1}{4}ze^{-z/2}\, dz,$$

which is equal to 0.31, approximately.

Remark. The rejection of the hypothesis H_0 when that hypothesis is true is, of course, an incorrect decision or an error. This incorrect decision is often called a Type-I error; accordingly, the significance level of the test is the probability of committing an error of Type I. The acceptance of H_0 when H_0 is false (H_1 is true) is called an error of Type II. Thus the probability of a Type-II error is one minus the power of the test when H_1 is true.

EXERCISES

9.1. Let X have a p.d.f. of the form $f(x; \theta) = \theta x^{\theta-1}$, $0 < x < 1$, zero elsewhere, where $\theta \in \{\theta; \theta = 1, 2\}$. To test the simple hypothesis $H_0: \theta = 1$ against the alternative simple hypothesis $H_1: \theta = 2$ use a random sample X_1, X_2 of size $n = 2$ and define the critical region to be $C = \{(x_1, x_2); 3/(4x_1) \leq x_2\}$. Find the power function of the test.

9.2. Let X have a binomial distribution with parameters $n = 10$ and $p \in \{p; p = \frac{1}{4}, \frac{1}{2}\}$. The simple hypothesis $H_0: p = \frac{1}{2}$ is rejected, and the alternative simple hypothesis $H_1: p = \frac{1}{4}$ is accepted, if the observed value of X_1, a random sample of size 1, is less than or equal to 3. Find the power function of the test.

9.3. Let X_1, X_2 be a random sample of size $n = 2$ from the distribution having p.d.f. $f(x; \theta) = (1/\theta)e^{-x/\theta}$, $0 < x < \infty$, zero elsewhere. We reject $H_0: \theta = 2$ and accept $H_1: \theta = 1$ if the observed values of X_1, X_2, say x_1, x_2, are such that

$$\frac{f(x_1; 2)f(x_2; 2)}{f(x_1; 1)f(x_2; 1)} \leq \frac{1}{2}.$$

Here $\Omega = \{\theta; \theta = 1, 2\}$. Find the significance level of the test and the power of the test when H_0 is false.

9.4. Sketch, as in Figure 9.1, the graphs of the power functions of Tests 1, 2, and 3 of Example 1 of this section.

9.5. Let us assume that the life of a tire in miles, say X, is normally distributed with mean θ and standard deviation 5000. Past experience indicates that $\theta = 30{,}000$. The manufacturer claims that the tires made by a new process have mean $\theta > 30{,}000$, and it is very possible that $\theta = 35{,}000$. Let us check his claim by testing $H_0: \theta \leq 30{,}000$ against $H_1: \theta > 30{,}000$. We shall observe n independent values of X, say $x_1, \ldots, x_n$, and we shall reject H_0 (thus accept H_1) if and only if $\bar{x} \geq c$. Determine n and c so that the power function $K(\theta)$ of the test has the values $K(30{,}000) = 0.01$ and $K(35{,}000) = 0.98$.

9.6. Let X have a Poisson distribution with mean θ. Consider the simple hypothesis $H_0: \theta = \frac{1}{2}$ and the alternative composite hypothesis $H_1: \theta < \frac{1}{2}$. Thus $\Omega = \{\theta; 0 < \theta \leq \frac{1}{2}\}$. Let $X_1, \ldots, X_{12}$ denote a random sample of size 12 from this distribution. We reject H_0 if and only if the

observed value of $Y = X_1 + \cdots + X_{12} \leq 2$. If $K(\theta)$ is the power function of the test, find the powers $K(\frac{1}{2})$, $K(\frac{1}{3})$, $K(\frac{1}{4})$, $K(\frac{1}{6})$, and $K(\frac{1}{12})$. Sketch the graph of $K(\theta)$. What is the significance level of the test?

9.2 Certain Best Tests

In this section we require that both the hypothesis H_0, which is to be tested, and the alternative hypothesis H_1 be simple hypotheses. Thus, in all instances, the parameter space is a set that consists of exactly two points. Under this restriction, we shall do three things:

(a) Define a best test for testing H_0 against H_1.

(b) Prove a theorem which provides a method of determining a best test.

(c) Give two examples.

Before we define a best test, one important observation should be made. Certainly, a test specifies a critical region; but it can also be said that a choice of a critical region defines a test. For instance, if one is given the critical region $C = \{(x_1, x_2, x_3); x_1^2 + x_2^2 + x_3^2 \geq 1\}$, the test is determined: Three random variables X_1, X_2, X_3 are to be considered; if the observed values are x_1, x_2, x_3, accept H_0 if $x_1^2 + x_2^2 + x_3^2 < 1$; otherwise, reject H_0. That is, the terms "test" and "critical region" can, in this sense, be used interchangeably. Thus, if we define a best critical region, we have defined a best test.

Let $f(x; \theta)$ denote the p.d.f. of a random variable X. Let $X_1, X_2, \ldots, X_n$ denote a random sample from this distribution, and consider the two simple hypotheses $H_0: \theta = \theta'$ and $H_1: \theta = \theta''$. Thus $\Omega = \{\theta; \theta = \theta', \theta''\}$. We now define a best critical region (and hence a best test) for testing the simple hypothesis H_0 against the alternative simple hypothesis H_1. In this definition the symbols $\Pr[(X_1, X_2, \ldots, X_n) \in C; H_0]$ and $\Pr[(X_1, X_2, \ldots, X_n) \in C; H_1]$ mean $\Pr[(X_1, X_2, \ldots, X_n) \in C]$ when, respectively, H_0 and H_1 are true.

Definition 6. Let C denote a subset of the sample space. Then C is called a *best critical region* of size α for testing the simple hypothesis $H_0: \theta = \theta'$ against the alternative simple hypothesis $H_1: \theta = \theta''$ if, for every subset A of the sample space for which $\Pr[(X_1, \ldots, X_n) \in A; H_0] = \alpha$,

(i) $\Pr[(X_1, X_2, \ldots, X_n) \in C; H_0] = \alpha$;

(ii) $\Pr[(X_1, X_2, \ldots, X_n) \in C; H_1] \geq \Pr[(X_1, X_2, \ldots, X_n) \in A; H_1]$.

This definition states, in effect, the following: First assume H_0 to be true. There will in general be a multiplicity of subsets A of the sample space such that $\Pr[(X_1, X_2, \ldots, X_n) \in A] = \alpha$. Suppose there is one of these subsets, say C, such that when H_1 is true, the power of the test associated with C is at least as great as the power of the test associated with each other A. Then C is defined as a best critical region of size α for testing H_0 against H_1.

In the following example, we shall examine this definition in some detail and in a very simple case.

Example 1. Consider the one random variable X which has a binomial distribution with $n = 5$ and $p = \theta$. Let $f(x; \theta)$ denote the p.d.f. of X and let H_0: $\theta = \frac{1}{2}$ and H_1: $\theta = \frac{3}{4}$. The following tabulation gives, at points of positive probability density, the values of $f(x; \frac{1}{2})$, $f(x; \frac{3}{4})$, and the ratio $f(x; \frac{1}{2})/f(x; \frac{3}{4})$.

x	0	1	2	3	4	5
$f(x; \frac{1}{2})$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$
$f(x; \frac{3}{4})$	$\frac{1}{1024}$	$\frac{15}{1024}$	$\frac{90}{1024}$	$\frac{270}{1024}$	$\frac{405}{1024}$	$\frac{243}{1024}$
$\frac{f(x; \frac{1}{2})}{f(x; \frac{3}{4})}$	32	$\frac{32}{3}$	$\frac{32}{9}$	$\frac{32}{27}$	$\frac{32}{81}$	$\frac{32}{243}$

We shall use one random value of X to test the simple hypothesis H_0: $\theta = \frac{1}{2}$ against the alternative simple hypothesis H_1: $\theta = \frac{3}{4}$ and we shall first assign the significance level of the test to be $\alpha = \frac{1}{32}$. We seek a best critical region of size $\alpha = \frac{1}{32}$. If $A_1 = \{x; x = 0\}$ and $A_2 = \{x; x = 5\}$, then $\Pr(X \in A_1; H_0) = \Pr(X \in A_2; H_0) = \frac{1}{32}$ and there is no other subset A_3 of the space $\{x; x = 0, 1, 2, 3, 4, 5\}$ such that $\Pr(X \in A_3; H_0) = \frac{1}{32}$. Then either A_1 or A_2 is the best critical region C of size $\alpha = \frac{1}{32}$ for testing H_0 against H_1. We note that $\Pr(X \in A_1; H_0) = \frac{1}{32}$ and that $\Pr(X \in A_1; H_1) = \frac{1}{1024}$. Thus if the set A_1 is used as a critical region of size $\alpha = \frac{1}{32}$, we have the intolerable situation that the probability of rejecting H_0 when H_1 is true (H_0 is false) is much less than the probability of rejecting H_0 when H_0 is true.

On the other hand, if the set A_2 is used as a critical region, then $\Pr(X \in A_2; H_0) = \frac{1}{32}$ and $\Pr(X \in A_2; H_1) = \frac{243}{1024}$. That is, the probability of rejecting H_0 when H_1 is true is much greater than the probability of rejecting H_0 when H_0 is true. Certainly, this is a more desirable state of affairs and actually A_2 is the best critical region of size $\alpha = \frac{1}{32}$. The latter statement follows from the fact that, when H_0 is true, there are but two subsets A_1 and A_2 of the sample space each of whose probability measure is $\frac{1}{32}$ and the fact that

$$\frac{243}{1024} = \Pr(X \in A_2; H_1) > \Pr(X \in A_1; H_1) = \frac{1}{1024}.$$

It should be noted, in this problem, that the best critical region $C = A_2$ of size $\alpha = \frac{1}{32}$ is found by including in C the point (or points) at which $f(x; \frac{1}{2})$

is *small* in comparison with $f(x; \frac{3}{4})$. This is seen to be true once it is observed that the ratio $f(x; \frac{1}{2})/f(x; \frac{3}{4})$ is a minimum at $x = 5$. Accordingly, the ratio $f(x; \frac{1}{2})/f(x; \frac{3}{4})$, which is given in the last line of the above tabulation, provides us with a precise tool by which to find a best critical region C for certain given values of α. To illustrate this, take $\alpha = \frac{6}{32}$. When H_0 is true, each of the subsets $\{x; x = 0, 1\}, \{x; x = 0, 4\}, \{x; x = 1, 5\}, \{x; x = 4, 5\}$ has probability measure $\frac{6}{32}$. By direct computation it is found that the best critical region of this size is $\{x; x = 4, 5\}$. This reflects the fact that the ratio $f(x; \frac{1}{2})/f(x; \frac{3}{4})$ has its two smallest values for $x = 4$ and $x = 5$. The power of this test, which has $\alpha = \frac{6}{32}$, is

$$\Pr(X = 4, 5; H_1) = \frac{405}{1024} + \frac{243}{1024} = \frac{648}{1024}.$$

The preceding example should make the following theorem, due to Neyman and Pearson, easier to understand. It is an important theorem because it provides a systematic method of determining a best critical region.

Neyman-Pearson Theorem. *Let $X_1, X_2, \ldots, X_n$, where n is a fixed positive integer, denote a random sample from a distribution that has p.d.f. $f(x; \theta)$. Then the joint p.d.f. of $X_1, X_2, \ldots, X_n$ is*

$$L(\theta; x_1, x_2, \ldots, x_n) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta).$$

Let θ' and θ'' be distinct fixed values of θ so that $\Omega = \{\theta; \theta = \theta', \theta''\}$, and let k be a positive number. Let C be a subset of the sample space such that:

(i) $\dfrac{L(\theta'; x_1, x_2, \ldots, x_n)}{L(\theta''; x_1, x_2, \ldots, x_n)} \leq k$ *for each point* $(x_1, x_2, \ldots, x_n) \in C$;

(ii) $\dfrac{L(\theta'; x_1, x_2, \ldots, x_n)}{L(\theta''; x_1, x_2, \ldots, x_n)} \geq k$ *for each point* $(x_1, x_2, \ldots, x_n) \in C^*$;

(iii) $\alpha = \Pr[(X_1, X_2, \ldots, X_n) \in C; H_0]$.

Then C is a best critical region of size α for testing the simple hypothesis H_0: $\theta = \theta'$ against the alternative simple hypothesis H_1: $\theta = \theta''$.

Proof. We shall give the proof when the random variables are of the continuous type. If C is the only critical region of size α, the theorem is proved. If there is another critical region of size α, denote it by A. For convenience, we shall let $\int_R \cdots \int L(\theta; x_1, \ldots, x_n)\, dx_1 \cdots dx_n$ be denoted by $\int_R L(\theta)$. In this notation, we wish to show that

$$\int_C L(\theta'') - \int_A L(\theta'') \geq 0.$$

Since C is the union of the disjoint sets $C \cap A$ and $C \cap A^*$ and A is the union of the disjoint sets $A \cap C$ and $A \cap C^*$, we have

$$\begin{aligned}(1)\quad &\int_C L(\theta'') - \int_A L(\theta'') \\ &= \int_{C\cap A} L(\theta'') + \int_{C\cap A^*} L(\theta'') - \int_{A\cap C} L(\theta'') - \int_{A\cap C^*} L(\theta'') \\ &= \int_{C\cap A^*} L(\theta'') - \int_{A\cap C^*} L(\theta'').\end{aligned}$$

However, by the hypothesis of the theorem, $L(\theta'') \geq (1/k)L(\theta')$ at each point of C, and hence at each point of $C \cap A^*$; thus,

$$\int_{C\cap A^*} L(\theta'') \geq \frac{1}{k}\int_{C\cap A^*} L(\theta').$$

But $L(\theta'') \leq (1/k)L(\theta')$ at each point of C^*, and hence at each point of $A \cap C^*$; accordingly,

$$\int_{A\cap C^*} L(\theta'') \leq \frac{1}{k}\int_{A\cap C^*} L(\theta').$$

These inequalities imply that

$$\int_{C\cap A^*} L(\theta'') - \int_{A\cap C^*} L(\theta'') \geq \frac{1}{k}\int_{C\cap A^*} L(\theta') - \frac{1}{k}\int_{A\cap C^*} L(\theta');$$

and, from Equation (1), we obtain

$$(2)\qquad \int_C L(\theta'') - \int_A L(\theta'') \geq \frac{1}{k}\left[\int_{C\cap A^*} L(\theta') - \int_{A\cap C^*} L(\theta')\right].$$

However,

$$\begin{aligned}&\int_{C\cap A^*} L(\theta') - \int_{A\cap C^*} L(\theta') \\ &\quad= \int_{C\cap A^*} L(\theta') + \int_{C\cap A} L(\theta') - \int_{A\cap C} L(\theta') - \int_{A\cap C^*} L(\theta') \\ &\quad= \int_C L(\theta') - \int_A L(\theta') \\ &\quad= \alpha - \alpha = 0.\end{aligned}$$

If this result is substituted in Inequality (2), we obtain the desired result

$$\int_C L(\theta'') - \int_A L(\theta'') \geq 0.$$

If the random variables are of the discrete type, the proof is the same, with integration replaced by summation.

One aspect of the theorem to be emphasized is that if we take C to be the set of all points $(x_1, x_2, \ldots, x_n)$ which satisfy

$$\frac{L(\theta'; x_1, x_2, \ldots, x_n)}{L(\theta''; x_1, x_2, \ldots, x_n)} \leq k, \qquad k > 0,$$

then, in accordance with the theorem, C will be a best critical region. This inequality can frequently be expressed in one of the forms (where c_1 and c_2 are constants)

$$u_1(x_1, x_2, \ldots, x_n; \theta', \theta'') \leq c_1,$$

or

$$u_2(x_1, x_2, \ldots, x_n; \theta', \theta'') \geq c_2.$$

Suppose that it is the first form, $u_1 \leq c_1$. Since θ' and θ'' are given constants, $u_1(X_1, X_2, \ldots, X_n; \theta', \theta'')$ is a statistic; and if the p.d.f. of this statistic can be found when H_0 is true, then the significance level of the test of H_0 against H_1 can be determined from this distribution. That is,

$$\alpha = \Pr [u_1(X_1, X_2, \ldots, X_n; \theta', \theta'') \leq c_1; H_0].$$

Moreover, the test may be based on this statistic; for, if the observed values of $X_1, X_2, \ldots, X_n$ are $x_1, x_2, \ldots, x_n$, we reject H_0 (accept H_1) if $u_1(x_1, x_2, \ldots, x_n) \leq c_1$.

A positive number k determines a best critical region C whose size is $\alpha = \Pr [(X_1, X_2, \ldots, X_n) \in C; H_0]$ for that particular k. It may be that this value of α is unsuitable for the purpose at hand; that is, it is too large or too small. However, if there is a statistic $u_1(X_1, X_2, \ldots, X_n)$, as in the preceding paragraph, whose p.d.f. can be determined when H_0 is true, we need not experiment with various values of k to obtain a desirable significance level. For if the distribution of the statistic is known, or can be found, we may determine c_1 such that $\Pr [u_1(X_1, X_2, \ldots, X_n) \leq c_1; H_0]$ is a desirable significance level.

An illustrative example follows.

Example 2. Let $X_1, X_2, \ldots, X_n$ denote a random sample from the distribution that has the p.d.f.

$$f(x; \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \theta)^2}{2}\right), \qquad -\infty < x < \infty.$$

It is desired to test the simple hypothesis H_0: $\theta = \theta' = 0$ against the alternative simple hypothesis H_1: $\theta = \theta'' = 1$. Now

$$\frac{L(\theta'; x_1, \ldots, x_n)}{L(\theta''; x_1, \ldots, x_n)} = \frac{(1/\sqrt{2\pi})^n \exp\left[-\left(\sum_1^n x_i^2\right)/2\right]}{(1/\sqrt{2\pi})^n \exp\left[-\left(\sum_1^n (x_i - 1)^2\right)/2\right]}$$

$$= \exp\left[-\sum_1^n x_i + \frac{n}{2}\right].$$

If $k > 0$, the set of all points $(x_1, x_2, \ldots, x_n)$ such that

$$\exp\left[-\sum_1^n x_i + \frac{n}{2}\right] \leq k$$

is a best critical region. This inequality holds if and only if

$$-\sum_1^n x_i + \frac{n}{2} \leq \ln k$$

or, equivalently,

$$\sum_1^n x_i \geq \frac{n}{2} - \ln k = c.$$

In this case, a best critical region is the set $C = \{(x_1, x_2, \ldots, x_n); \sum_1^n x_i \geq c\}$, where c is a constant that can be determined so that the size of the critical region is a desired number α. The event $\sum_1^n X_i \geq c$ is equivalent to the event $\overline{X} \geq c/n = c_1$, say, so the test may be based upon the statistic $\overline{X}$. If H_0 is true, that is, $\theta = \theta' = 0$, then $\overline{X}$ has a distribution which is $n(0, 1/n)$. For a given positive integer n, the size of the sample, and a given significance level α, the number c_1 can be found from Table III in the Appendix, so that $\Pr(\overline{X} \geq c_1; H_0) = \alpha$. Hence, if the experimental values of $X_1, X_2, \ldots, X_n$ were, respectively, $x_1, x_2, \ldots, x_n$, we would compute $\bar{x} = \sum_1^n x_i/n$. If $\bar{x} \geq c_1$, the simple hypothesis H_0: $\theta = \theta' = 0$ would be rejected at the significance level α; if $\bar{x} < c_1$, the hypothesis H_0 would be accepted. The probability of rejecting H_0, when H_0 is true, is α; the probability of rejecting H_0, when H_0 is false, is the value of the power of the test at $\theta = \theta'' = 1$. That is,

$$\Pr(\overline{X} \geq c_1; H_1) = \int_{c_1}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1/n}} \exp\left[-\frac{(\bar{x} - 1)^2}{2(1/n)}\right] d\bar{x}.$$

For example, if $n = 25$ and if α is selected to be 0.05, then from Table III we find $c_1 = 1.645/\sqrt{25} = 0.329$. Thus the power of this best test of H_0 against H_1 is 0.05, when H_0 is true, and is

$$\int_{0.329}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{\frac{1}{25}}} \exp\left[-\frac{(\bar{x} - 1)^2}{2(\frac{1}{25})}\right] d\bar{x} = \int_{-3.355}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw = 0.999+,$$

when H_1 is true.

There is another aspect of this theorem that warrants special mention. It has to do with the number of parameters that appear in the p.d.f. Our notation suggests that there is but one parameter. However, a careful review of the proof will reveal that nowhere was this needed nor assumed. The p.d.f. may depend upon any finite number of parameters. What is essential is that the hypothesis H_0 and the alternative hypothesis H_1 be simple, namely, that they completely specify the distributions. With this in mind, we see that the simple hypotheses

H_0 and H_1 do not need to be hypotheses about the parameters of a distribution, nor, as a matter of fact, do the random variables $X_1, X_2, \ldots, X_n$ need to be mutually stochastically independent. That is, if H_0 is the simple hypothesis that the joint p.d.f. is $g(x_1, x_2, \ldots, x_n)$, and if H_1 is the alternative simple hypothesis that the joint p.d.f. is $h(x_1, x_2, \ldots, x_n)$, then C is a best critical region of size α for testing H_0 against H_1 if, for $k > 0$;

$$\text{(i)}' \quad \frac{g(x_1, x_2, \ldots, x_n)}{h(x_1, x_2, \ldots, x_n)} \leq k \qquad \text{for } (x_1, x_2, \ldots, x_n) \in C;$$

$$\text{(ii)}' \quad \frac{g(x_1, x_2, \ldots, x_n)}{h(x_1, x_2, \ldots, x_n)} \geq k \qquad \text{for } (x_1, x_2, \ldots, x_n) \in C^*;$$

$$\text{(iii)}' \quad \alpha = \Pr[(X_1, X_2, \ldots, X_n) \in C; H_0].$$

An illustrative example follows.

Example 3. Let $X_1, \ldots, X_n$ denote a random sample from a distribution which has a p.d.f. $f(x)$ which is positive on and only on the nonnegative integers. It is desired to test the simple hypothesis

$$H_0: f(x) = \frac{e^{-1}}{x!}, \qquad x = 0, 1, 2, \ldots,$$
$$= 0 \text{ elsewhere,}$$

against the alternative simple hypothesis

$$H_1: f(x) = (\tfrac{1}{2})^{x+1}, \qquad x = 0, 1, 2, \ldots,$$
$$= 0 \text{ elsewhere.}$$

Here

$$\frac{g(x_1, \ldots, x_n)}{h(x_1, \ldots, x_n)} = \frac{e^{-n}/(x_1!\, x_2! \cdots x_n!)}{(\frac{1}{2})^n (\frac{1}{2})^{x_1 + x_2 + \cdots + x_n}}$$
$$= \frac{(2e^{-1})^n 2^{\Sigma x_i}}{\prod_1^n (x_i!)}.$$

If $k > 0$, the set of points $(x_1, x_2, \ldots, x_n)$ such that

$$\left(\sum_1^n x_i\right) \ln 2 - \ln\left[\prod_1^n (x_i!)\right] \leq \ln k - n \ln (2e^{-1}) = c$$

is a best critical region C. Consider the case of $k = 1$ and $n = 1$. The preceding inequality may be written $2^{x_1}/x_1! \leq e/2$. This inequality is satisfied by all points in the set $C = \{x_1; x_1 = 0, 3, 4, 5, \ldots\}$. Thus the power of the test when H_0 is true is

$$\Pr(X_1 \in C; H_0) = 1 - \Pr(X_1 = 1, 2; H_0) = 0.448,$$

approximately, in accordance with Table I of the Appendix. The power of the test when H_1 is true is given by

$$\begin{aligned}\Pr(X_1 \in C; H_1) &= 1 - \Pr(X_1 = 1, 2; H_1) \\ &= 1 - (\tfrac{1}{4} + \tfrac{1}{8}) = 0.625.\end{aligned}$$

EXERCISES

9.7. In Example 2 of this section, let the simple hypotheses read H_0: $\theta = \theta' = 0$ and H_1: $\theta = \theta'' = -1$. Show that the best test of H_0 against H_1 may be carried out by use of the statistic $\bar{X}$, and that if $n = 25$ and $\alpha = 0.05$, the power of the test is 0.999+ when H_1 is true.

9.8. Let the random variable X have the p.d.f. $f(x; \theta) = (1/\theta)e^{-x/\theta}$, $0 < x < \infty$, zero elsewhere. Consider the simple hypothesis H_0: $\theta = \theta' = 2$ and the alternative hypothesis H_1: $\theta = \theta'' = 4$. Let X_1, X_2 denote a random sample of size 2 from this distribution. Show that the best test of H_0 against H_1 may be carried out by use of the statistic $X_1 + X_2$ and that the assertion in Example 2 of Section 9.1 is correct.

9.9. Repeat Exercise 9.8 when H_1: $\theta = \theta'' = 6$. Generalize this for every $\theta'' > 2$.

9.10. Let $X_1, X_2, \ldots, X_{10}$ be a random sample of size 10 from a normal distribution $n(0, \sigma^2)$. Find a best critical region of size $\alpha = 0.05$ for testing H_0: $\sigma^2 = 1$ against H_1: $\sigma^2 = 2$. Is this a best critical region of size 0.05 for testing H_0: $\sigma^2 = 1$ against H_1: $\sigma^2 = 4$? Against H_1: $\sigma^2 = \sigma_1^2 > 1$?

9.11. If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution having p.d.f. of the form $f(x; \theta) = \theta x^{\theta - 1}$, $0 < x < 1$, zero elsewhere, show that a best critical region for testing H_0: $\theta = 1$ against H_1: $\theta = 2$ is $C = \left\{(x_1, x_2, \ldots, x_n); c \leq \prod_{i=1}^{n} x_i\right\}$.

9.12. Let $X_1, X_2, \ldots, X_{10}$ be a random sample from a distribution which is $n(\theta_1, \theta_2)$. Find a best test of the simple hypothesis H_0: $\theta_1 = \theta_1' = 0$, $\theta_2 = \theta_2' = 1$ against the alternative simple hypothesis H_1: $\theta_1 = \theta_1'' = 1$, $\theta_2 = \theta_2'' = 4$.

9.13. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a normal distribution $n(\theta, 100)$. Show that $C = \{(x_1, x_2, \ldots, x_n); c \leq \bar{x} = \sum_1^n x_i/n\}$ is a best critical region for testing H_0: $\theta = 75$ against H_1: $\theta = 78$. Find n and c so that

$$\Pr[(X_1, X_2, \ldots, X_n) \in C; H_0] = \Pr[\bar{X} \geq c; H_0] = 0.05$$

and

$$\Pr[(X_1, X_2, \ldots, X_n) \in C; H_1] = \Pr[\bar{X} \geq c; H_1] = 0.90, \text{ approximately.}$$

9.14. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution having the p.d.f. $f(x; p) = p^x(1-p)^{1-x}$, $x = 0, 1$, zero elsewhere. Show that $C = \{(x_1, \ldots, x_n); \sum_1^n x_i \leq c\}$ is a best critical region for testing $H_0: p = \frac{1}{2}$ against $H_1: p = \frac{1}{3}$. Use the central limit theorem to find n and c so that approximately $\Pr\left(\sum_1^n X_i \leq c; H_0\right) = 0.10$ and $\Pr\left(\sum_1^n X_i \leq c; H_1\right) = 0.80$.

9.15. Let $X_1, X_2, \ldots, X_{10}$ denote a random sample of size 10 from a Poisson distribution with mean θ. Show that the critical region C defined by $\sum_1^{10} x_i \geq 3$ is a best critical region for testing $H_0: \theta = 0.1$ against $H_1: \theta = 0.5$. Determine, for this test, the significance level α and the power at $\theta = 0.5$.

9.3 Uniformly Most Powerful Tests

This section will take up the problem of a test of a simple hypothesis H_0 against an alternative composite hypothesis H_1. We begin with an example.

Example 1. Consider the p.d.f.

$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}, \qquad 0 < x < \infty,$$
$$= 0 \text{ elsewhere},$$

of Example 2, p. 269. It is desired to test the simple hypothesis $H_0: \theta = 2$ against the alternative composite hypothesis $H_1: \theta > 2$. Thus $\Omega = \{\theta; \theta \geq 2\}$. A random sample, X_1, X_2, of size $n = 2$ will be used, and the critical region is $C = \{(x_1, x_2); 9.5 \leq x_1 + x_2 < \infty\}$. It was shown in the example cited that the significance level of the test is approximately 0.05 and that the power of the test when $\theta = 4$ is approximately 0.31. The power function $K(\theta)$ of the test for all $\theta \geq 2$ will now be obtained. We have

$$K(\theta) = 1 - \int_0^{9.5} \int_0^{9.5 - x_2} \frac{1}{\theta^2} \exp\left(-\frac{x_1 + x_2}{\theta}\right) dx_1\, dx_2$$
$$= \left(\frac{\theta + 9.5}{\theta}\right) e^{-9.5/\theta}, \qquad 2 \leq \theta.$$

Thus, for instance, $K(2) = 0.05$, $K(4) = 0.31$, and $K(9.5) = 2/e$. It is known, Exercise 9.8, p. 279, that $C = \{(x_1, x_2); 9.5 \leq x_1 + x_2 < \infty\}$ is a best critical region of size 0.05 for testing the simple hypothesis $H_0: \theta = 2$ against each simple hypothesis in the composite hypothesis $H_1: \theta > 2$.

The preceding example affords an illustration of a test of a simple hypothesis H_0 which is a best test of H_0 against every simple hypothesis in the alternative composite hypothesis H_1. We now define a critical

region, when it exists, which is a best critical region for testing a simple hypothesis H_0 against an alternative composite hypothesis H_1. It seems desirable that this critical region should be a best critical region for testing H_0 against each simple hypothesis in H_1. That is, the power function of the test which corresponds to this critical region should be at least as great as the power function of any other test with the same significance level for every simple hypothesis in H_1.

Definition 7. The critical region C is a *uniformly most powerful critical region* of size α for testing the simple hypothesis H_0 against an alternative composite hypothesis H_1 if the set C is a best critical region of size α for testing H_0 against each simple hypothesis in H_1. A test, defined by this critical region C, is called a *uniformly most powerful test*, with significance level α, for testing the simple hypothesis H_0 against the alternative composite hypothesis H_1.

As will be seen presently, uniformly most powerful tests do not always exist. However, when they do exist, the Neyman-Pearson theorem provides a technique for finding them. Some illustrative examples are given here.

Example 2. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $n(0, \theta)$, where the variance θ is an unknown positive number. It will be shown that there exists a uniformly most powerful test with significance level α for testing the simple hypothesis H_0: $\theta = \theta'$, where θ' is a fixed positive number, against the alternative composite hypothesis H_1: $\theta > \theta'$. Thus $\Omega = \{\theta; \theta \geq \theta'\}$. The joint p.d.f. of $X_1, X_2, \ldots, X_n$ is

$$L(\theta; x_1, x_2, \ldots, x_n) = \left(\frac{1}{2\pi\theta}\right)^{n/2} \exp\left(-\frac{\sum_1^n x_i^2}{2\theta}\right).$$

Let θ'' represent a number greater than θ', and let k denote a positive number. Let C be the set of points where

$$\frac{L(\theta'; x_1, x_2, \ldots, x_n)}{L(\theta''; x_1, x_2, \ldots, x_n)} \leq k,$$

that is, the set of points where

$$\left(\frac{\theta''}{\theta'}\right)^{n/2} \exp\left[-\left(\frac{\theta'' - \theta'}{2\theta'\theta''}\right)\sum_1^n x_i^2\right] \leq k$$

or, equivalently,

$$\sum_1^n x_i^2 \geq \frac{2\theta'\theta''}{\theta'' - \theta'}\left[\frac{n}{2}\ln\left(\frac{\theta''}{\theta'}\right) - \ln k\right] = c.$$

The set $C = \{(x_1, x_2, \ldots, x_n); \sum_1^n x_i^2 \geq c\}$ is then a best critical region for testing the simple hypothesis H_0: $\theta = \theta'$ against the simple hypothesis $\theta = \theta''$. It remains to determine c so that this critical region has the desired size α. If H_0 is true, the random variable $\sum_1^n X_i^2/\theta'$ has a chi-square distribution with n degrees of freedom. Since $\alpha = \Pr\left(\sum_1^n X_i^2/\theta' \geq c/\theta'; H_0\right)$, c/θ' may be read from Table II in the Appendix and c determined. Then $C = \{(x_1, x_2, \ldots, x_n); \sum_1^n x_i^2 \geq c\}$ is a best critical region of size α for testing H_0: $\theta = \theta'$ against the hypothesis $\theta = \theta''$. Moreover, for each number θ'' greater than θ', the foregoing argument holds. That is, if θ''' is another number greater than θ', then $C = \{(x_1, \ldots, x_n); \sum_1^n x_i^2 \geq c\}$ is a best critical region of size α for testing H_0: $\theta = \theta'$ against the hypothesis $\theta = \theta'''$. Accordingly, $C = \{(x_1, \ldots, x_n); \sum_1^n x_i^2 \geq c\}$ is a uniformly most powerful critical region of size α for testing H_0: $\theta = \theta'$ against H_1: $\theta > \theta'$. If $x_1, x_2, \ldots, x_n$ denote the experimental values of $X_1, X_2, \ldots, X_n$, then H_0: $\theta = \theta'$ is rejected at the significance level α, and H_1: $\theta > \theta'$ is accepted, if $\sum_1^n x_i^2 \geq c$; otherwise H_0: $\theta = \theta'$ is accepted.

If in the preceding discussion we take $n = 15$, $\alpha = 0.05$, and $\theta' = 3$, then here the two hypotheses will be H_0: $\theta = 3$ and H_1: $\theta > 3$. From Table II, $c/3 = 25$ and hence $c = 75$.

***Example* 3.** Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $n(\theta, 1)$, where the mean θ is unknown. It will be shown that there is no uniformly most powerful test of the simple hypothesis H_0: $\theta = \theta'$, where θ' is a fixed number, against the alternative composite hypothesis H_1: $\theta \neq \theta'$. Thus $\Omega = \{\theta; -\infty < \theta < \infty\}$. Let θ'' be a number not equal to θ'. Let k be a positive number and consider

$$\frac{(1/2\pi)^{n/2} \exp\left[-\sum_1^n (x_i - \theta')^2/2\right]}{(1/2\pi)^{n/2} \exp\left[-\sum_1^n (x_i - \theta'')^2/2\right]} \leq k.$$

The preceding inequality may be written as

$$\exp\left\{-(\theta'' - \theta') \sum_1^n x_i + \frac{n}{2}[(\theta'')^2 - (\theta')^2]\right\} \leq k,$$

or

$$(\theta'' - \theta') \sum_1^n x_i \geq \frac{n}{2}[(\theta'')^2 - (\theta')^2] - \ln k.$$

This last inequality is equivalent to

$$\sum_1^n x_i \geq \frac{n}{2}(\theta'' + \theta') - \frac{\ln k}{\theta'' - \theta'},$$

provided $\theta'' > \theta'$, and it is equivalent to

$$\sum_1^n x_i \leq \frac{n}{2}(\theta'' + \theta') - \frac{\ln k}{\theta'' - \theta'}$$

if $\theta'' < \theta'$. The first of these two expressions defines a best critical region for testing H_0: $\theta = \theta'$ against the hypothesis $\theta = \theta''$, provided $\theta'' > \theta'$, while the second expression defines a best critical region for testing H_0: $\theta = \theta'$ against the hypothesis $\theta = \theta''$, provided $\theta'' < \theta'$. That is, a best critical region for testing the simple hypothesis against an alternative simple hypothesis, say $\theta = \theta' + 1$, will not serve as a best critical region for testing H_0: $\theta = \theta'$ against the alternative simple hypothesis $\theta = \theta' - 1$, say. By definition, there is then no uniformly most powerful test in the case under consideration.

It should be noted that had the alternative composite hypothesis been either H_1: $\theta > \theta'$ or H_1: $\theta < \theta'$, then in each instance, a uniformly most powerful test would exist.

This section will conclude with an observation which, although obvious when pointed out, is important. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that has p.d.f. $f(x; \theta)$, where θ is an element of a prescribed set. Let $Y = u(X_1, \ldots, X_n)$ be a sufficient statistic for θ. In accordance with the Fisher-Neyman criterion (p. 216), the joint p.d.f. of $X_1, X_2, \ldots, X_n$ may be written

$$\begin{aligned} L(\theta; x_1, \ldots, x_n) &= f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta) \\ &= g[u(x_1, x_2, \ldots, x_n); \theta] H(x_1, x_2, \ldots, x_n), \end{aligned}$$

where $g(y; \theta)$ is the p.d.f. of the sufficient statistic Y and $H(x_1, x_2, \ldots, x_n)$ does not depend upon θ. Consequently, the ratio

$$\frac{L(\theta'; x_1, \ldots, x_n)}{L(\theta''; x_1, \ldots, x_n)} = \frac{g[u(x_1, x_2, \ldots, x_n); \theta']}{g[u(x_1, x_2, \ldots, x_n); \theta'']}$$

depends upon $x_1, x_2, \ldots, x_n$ only through $y = u(x_1, x_2, \ldots, x_n)$. Accordingly, if there is a sufficient statistic Y for the parameter θ, and if a best test or a uniformly most powerful test is desired, there is no need to consider tests which are based upon any statistic other than the sufficient statistic $Y = u(X_1, X_2, \ldots, X_n)$.

EXERCISES

9.16. Let X have the p.d.f. $f(x; \theta) = \theta^x(1 - \theta)^{1-x}$, $x = 0, 1$, zero elsewhere. We test the simple hypothesis H_0: $\theta = \frac{1}{4}$ against the alternative composite hypothesis H_1: $\theta < \frac{1}{4}$ by taking a random sample of size 10 and rejecting H_0: $\theta = \frac{1}{4}$ if and only if the observed values $x_1, x_2, \ldots, x_{10}$ of the sample items are such that $\sum_1^{10} x_i \leq 1$. Find the power function $K(\theta)$, $0 < \theta \leq \frac{1}{4}$, of this test.

9.17. Let X have a p.d.f. of the form $f(x; \theta) = 1/\theta$, $0 < x < \theta$, zero elsewhere. Let $Y_1 < Y_2 < Y_3 < Y_4$ denote the order statistics of a random sample of size 4 from this distribution. Let the observed value of Y_4 be y_4. We reject H_0: $\theta = 1$ and accept H_1: $\theta \neq 1$ if either $y_4 \leq \frac{1}{2}$ or $y_4 \geq 1$. Find the power function $K(\theta)$, $0 < \theta$, of the test.

9.18. Consider a normal distribution of the form $n(\theta, 4)$. The simple hypothesis H_0: $\theta = 0$ is rejected, and the alternative composite hypothesis H_1: $\theta > 0$ is accepted if and only if the observed mean $\bar{x}$ of a random sample of size 25 is greater than or equal to $\frac{3}{5}$. Find the power function $K(\theta)$, $0 \leq \theta$, of this test.

9.19. Consider the two independent normal distributions $n(\mu_1, 400)$ and $n(\mu_2, 225)$. Let $\theta = \mu_1 - \mu_2$. Let $\bar{x}$ and $\bar{y}$ denote the observed means of two independent random samples, each of size n, from these two distributions. We reject H_0: $\theta = 0$ and accept H_1: $\theta > 0$ if and only if $\bar{x} - \bar{y} \geq c$. If $K(\theta)$ is the power function of this test, find n and c so that $K(0) = 0.05$ and $K(10) = 0.90$, approximately.

9.20. If, in Example 2 of this section, H_0: $\theta = \theta'$, where θ' is a fixed positive number, and H_1: $\theta < \theta'$, show that the set $\{(x_1, x_2, \ldots, x_n); \sum_1^n x_i^2 \leq c\}$ is a uniformly most powerful critical region for testing H_0 against H_1.

9.21. If, in Example 2 of this section, H_0: $\theta = \theta'$, where θ' is a fixed positive number, and H_1: $\theta \neq \theta'$, show that there is no uniformly most powerful test for testing H_0 against H_1.

9.22. Let $X_1, X_2, \ldots, X_{25}$ denote a random sample of size 25 from a normal distribution $n(\theta, 100)$. Find a uniformly most powerful critical region of size $\alpha = 0.10$ for testing H_0: $\theta = 75$ against H_1: $\theta > 75$.

9.23. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a normal distribution $n(\theta, 16)$. Find the sample size n and a uniformly most powerful test of H_0: $\theta = 25$ against H_1: $\theta < 25$ with power function $K(\theta)$ so that approximately $K(25) = 0.10$ and $K(23) = 0.90$.

9.24. Consider a distribution having a p.d.f. of the form $f(x; \theta) = \theta^x(1 - \theta)^{1-x}$, $x = 0, 1$, zero elsewhere. Let H_0: $\theta = \frac{1}{20}$ and H_1: $\theta > \frac{1}{20}$. Use the central limit theorem to determine the sample size n of a random sample

so that a uniformly most powerful test of H_0 against H_1 has a power function $K(\theta)$, with approximately $K(\frac{1}{20}) = 0.05$ and $K(\frac{1}{10}) = 0.90$.

9.25. Let $X_1, X_2, \ldots, X_{20}$ be a random sample of size 20 from a Poisson distribution with mean θ. Show that the critical region defined by $\sum_1^{20} x_i \geq 5$ is a uniformly most powerful critical region for testing H_0: $\theta = \frac{1}{10}$ against H_1: $\theta > \frac{1}{10}$. What is α, the significance level of the test?

9.26. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution which has p.d.f. $f(x; \theta_1, \theta_2) = \theta_2 e^{-\theta_2(x-\theta_1)}$, $\theta_1 \leq x < \infty$, zero elsewhere. Find a best critical region for testing the simple hypothesis H_0: $\theta_1 = \theta_1'$, $\theta_2 = \theta_2' > 0$, against the alternative simple hypothesis H_1: $\theta_1 = \theta_1'' < \theta_1'$, $\theta_2 = \theta_2'' > \theta_2'$. Is this a uniformly most powerful critical region for testing H_0: $\theta_1 = \theta_1'$, $\theta_2 = \theta_2' > 0$, against H_1: $\theta_1 < \theta_1'$, $\theta_2 > \theta_2'$?

9.27. Illustrative Example 1 of this section dealt with a random sample of size $n = 2$ from a gamma distribution with $\alpha = 1$, $\beta = \theta$. Thus the moment-generating function of the distribution is $(1 - \theta t)^{-1}$, $t < 1/\theta$, $\theta \geq 2$. Let $Z = X_1 + X_2$. Show that Z has a gamma distribution with $\alpha = 2$, $\beta = \theta$. Express the power function $K(\theta)$ of Example 1 in terms of a single integral. Generalize this for a random sample of size n.

9.28. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with p.d.f. $f(x; \theta) = \exp[p(\theta)K(x) + S(x) + q(\theta)]$, $a < x < b$, zero elsewhere, which is a regular case of the exponential class. Let $\Omega = \{\theta; \theta' \leq \theta\}$. (a) If $p(\theta)$ is an increasing function of θ, show that $C = \{(x_1, x_2, \ldots, x_n); \sum_1^n K(x_i) \geq c\}$ is a uniformly most powerful critical region for testing H_0: $\theta = \theta'$ against H_1: $\theta > \theta'$. (b) Consider the case in which $p(\theta)$ is a decreasing function of θ.

9.4 The Sequential Probability Ratio Test

In Section 9.2 we proved a theorem which provided us with a method for determining a best critical region for testing a simple hypothesis against an alternative simple hypothesis. The theorem was: Let $X_1, X_2, \ldots, X_n$ be a random sample with fixed sample size n from a distribution which has p.d.f. $f(x; \theta)$, where $\theta \in \{\theta; \theta = \theta', \theta''\}$ and θ' and θ'' are known numbers. Let the joint p.d.f. of $X_1, X_2, \ldots, X_n$ be denoted by

$$L(\theta, n) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta),$$

a notation that reveals both the parameter θ and the sample size n. If we reject H_0: $\theta = \theta'$ and accept H_1: $\theta = \theta''$ when and only when

$$\frac{L(\theta', n)}{L(\theta'', n)} \leq k,$$

where $k > 0$, then this is a best test of H_0 against H_1.

Let us now suppose that the sample size n is *not* fixed in advance. In fact, let the sample size be a random variable N with sample space $\{n; n = 1, 2, 3, \ldots\}$. An interesting procedure for testing the simple hypothesis H_0: $\theta = \theta'$ against the simple hypothesis H_1: $\theta = \theta''$ is the following: Let k_0 and k_1 be two positive constants with $k_0 < k_1$. Observe the mutually stochastically independent outcomes $X_1, X_2, X_3, \ldots$ in sequence, say $x_1, x_2, x_3, \ldots$, and compute

$$\frac{L(\theta', 1)}{L(\theta'', 1)}, \frac{L(\theta', 2)}{L(\theta'', 2)}, \frac{L(\theta', 3)}{L(\theta'', 3)}, \ldots.$$

The hypothesis H_0: $\theta = \theta'$ is rejected (and H_1: $\theta = \theta''$ is accepted) if and only if there exists a positive integer n so that $(x_1, x_2, \ldots, x_n)$ belongs to the set

$$C_n = \left\{(x_1, \ldots, x_n); k_0 < \frac{L(\theta', j)}{L(\theta'', j)} < k_1, j = 1, \ldots, n-1, \text{ and } \frac{L(\theta', n)}{L(\theta'', n)} \leq k_0\right\}.$$

On the other hand, the hypothesis H_0: $\theta = \theta'$ is accepted (and H_1: $\theta = \theta''$ is rejected) if and only if there exists a positive integer n so that $(x_1, x_2, \ldots, x_n)$ belongs to the set

$$B_n = \left\{(x_1, \ldots, x_n); k_0 < \frac{L(\theta', j)}{L(\theta'', j)} < k_1, j = 1, \ldots, n-1, \text{ and } \frac{L(\theta', n)}{L(\theta'', n)} \geq k_1\right\}.$$

That is, we continue to observe sample items as long as

$$k_0 < \frac{L(\theta', n)}{L(\theta'', n)} < k_1. \tag{1}$$

We stop these observations in one of two ways:

(a) With rejection of H_0: $\theta = \theta'$ as soon as

$$\frac{L(\theta', n)}{L(\theta'', n)} \leq k_0,$$

or

(b) with acceptance of H_0: $\theta = \theta'$ as soon as

$$\frac{L(\theta', n)}{L(\theta'', n)} \geq k_1.$$

A test of this kind is called a *sequential probability ratio test.* Now, frequently Inequality (1) can be conveniently expressed in an equivalent form

$$c_0(n) < u(x_1, x_2, \ldots, x_n) < c_1(n),$$

where $u(X_1, X_2, \ldots, X_n)$ is a statistic and $c_0(n)$ and $c_1(n)$ depend on the constants k_0, k_1, θ', θ'', and on n. Then the observations are stopped and a decision is reached as soon as

$$u(x_1, x_2, \ldots, x_n) \leq c_0(n), \qquad \text{or} \qquad u(x_1, x_2, \ldots, x_n) \geq c_1(n).$$

We now give an illustrative example.

Example 1. Let X have the p.d.f.

$$f(x; \theta) = \theta^x(1 - \theta)^{1-x}, \qquad x = 0, 1,$$
$$= 0 \text{ elsewhere.}$$

In the preceding discussion of a sequential probability ratio test, let H_0: $\theta = \frac{1}{3}$ and H_1: $\theta = \frac{2}{3}$; then, with $\sum x_i = \sum_1^n x_i$,

$$\frac{L(\frac{1}{3}, n)}{L(\frac{2}{3}, n)} = \frac{(\frac{1}{3})^{\sum x_i}(\frac{2}{3})^{n-\sum x_i}}{(\frac{2}{3})^{\sum x_i}(\frac{1}{3})^{n-\sum x_i}} = 2^{n-2\sum x_i}.$$

If we take logarithms to the base 2, the inequality

$$k_0 < \frac{L(\frac{1}{3}, n)}{L(\frac{2}{3}, n)} < k_1,$$

with $0 < k_0 < k_1$, becomes

$$\log_2 k_0 < n - 2\sum_1^n x_i < \log_2 k_1,$$

or, equivalently,

$$c_0(n) = \frac{n}{2} - \frac{1}{2}\log_2 k_1 < \sum_1^n x_i < \frac{n}{2} - \frac{1}{2}\log_2 k_0 = c_1(n).$$

Note that $L(\frac{1}{3}, n)/L(\frac{2}{3}, n) \leq k_0$ if and only if $c_1(n) \leq \sum_1^n x_i$; and $L(\frac{1}{3}, n)/L(\frac{2}{3}, n) \geq k_1$ if and only if $c_0(n) \geq \sum_1^n x_i$. Thus we continue to observe outcomes as long as $c_0(n) < \sum_1^n x_i < c_1(n)$. The observation of outcomes is discontinued with the first value n of N for which either $c_1(n) \leq \sum_1^n x_i$ or $c_0(n) \geq \sum_1^n x_i$. The inequality $c_1(n) \leq \sum_1^n x_i$ leads to the rejection of H_0: $\theta = \frac{1}{3}$ (the acceptance of H_1), and the inequality $c_0(n) \geq \sum_1^n x_i$ leads to the acceptance of H_0: $\theta = \frac{1}{3}$ (the rejection of H_1).

Remarks. At this point, the student undoubtedly sees that there are many questions which should be raised in connection with the sequential probability ratio test. Some of these questions are possibly among the following:

(a) What is the probability of the procedure continuing indefinitely?

(b) What is the value of the power function of this test at each of the points $\theta = \theta'$ and $\theta = \theta''$?

(c) If θ'' is one of several values of θ specified by an alternative composite hypothesis, say H_1: $\theta > \theta'$, what is the power function at each point $\theta \geq \theta'$?

(d) Since the sample size N is a random variable, what are some of the properties of the distribution of N? In particular, what is the expected value $E(N)$ of N?

(e) How does this test compare with tests that have a fixed sample size n?

A course in sequential analysis would investigate these and many other problems. However, in this book our objective is largely that of acquainting the student with this kind of test procedure. Accordingly, we assert that the answer to question (a) is zero. Moreover, it can be proved that if $\theta = \theta'$ or if $\theta = \theta''$, $E(N)$ is smaller, for this sequential procedure, than the sample size of a fixed sample size test which has the same values of the power function at those points. We now consider question (b) in some detail.

In this section we shall denote the power of the test when H_0 is true by the symbol α and the power of the test when H_1 is true by the symbol $1 - \beta$. Thus α is the probability of committing a Type-I error (the rejection of H_0 when H_0 is true), and β is the probability of committing a Type-II error (the acceptance of H_0 when H_0 is false). With the sets C_n and B_n as previously defined, and with random variables of the continuous type, we then have

$$\alpha = \sum_{n=1}^{\infty} \int_{C_n} L(\theta', n), \qquad 1 - \beta = \sum_{n=1}^{\infty} \int_{C_n} L(\theta'', n).$$

Since the probability is one that the procedure will terminate, we also have

$$1 - \alpha = \sum_{n=1}^{\infty} \int_{B_n} L(\theta', n), \qquad \beta = \sum_{n=1}^{\infty} \int_{B_n} L(\theta'', n).$$

If $(x_1, x_2, \ldots, x_n) \in C_n$, we have $L(\theta', n) \leq k_0 L(\theta'', n)$; hence it is clear that

$$\alpha = \sum_{n=1}^{\infty} \int_{C_n} L(\theta', n) \leq \sum_{n=1}^{\infty} \int_{C_n} k_0 L(\theta'', n) = k_0(1 - \beta).$$

Because $L(\theta', n) \geq k_1 L(\theta'', n)$ at each point of the set B_n, we have

$$1 - \alpha = \sum_{n=1}^{\infty} \int_{B_n} L(\theta', n) \geq \sum_{n=1}^{\infty} \int_{B_n} k_1 L(\theta'', n) = k_1 \beta.$$

Accordingly, it follows that

$$\frac{\alpha}{1-\beta} \leq k_0, \qquad k_1 \leq \frac{1-\alpha}{\beta}, \tag{2}$$

provided β is not equal to zero or one.

Now let α_a and β_a be preassigned proper fractions; some typical values in the applications are 0.01, 0.05, and 0.10. If we take

$$k_0 = \frac{\alpha_a}{1-\beta_a}, \qquad k_1 = \frac{1-\alpha_a}{\beta_a},$$

then Inequalities (2) become

$$\frac{\alpha}{1-\beta} \leq \frac{\alpha_a}{1-\beta_a}, \qquad \frac{1-\alpha_a}{\beta_a} \leq \frac{1-\alpha}{\beta}; \tag{3}$$

or, equivalently,

$$\alpha(1-\beta_a) \leq (1-\beta)\alpha_a, \qquad \beta(1-\alpha_a) \leq (1-\alpha)\beta_a.$$

If we add corresponding members of the immediately preceding inequalities, we find that

$$\alpha + \beta - \alpha\beta_a - \beta\alpha_a \leq \alpha_a + \beta_a - \beta\alpha_a - \alpha\beta_a$$

and hence

$$\alpha + \beta \leq \alpha_a + \beta_a.$$

That is, the sum $\alpha + \beta$ of the probabilities of the two kinds of errors is bounded above by the sum $\alpha_a + \beta_a$ of the preassigned numbers. Moreover, since α and β are positive proper fractions, Inequalities (3) imply that

$$\alpha \leq \frac{\alpha_a}{1-\beta_a}, \qquad \beta \leq \frac{\beta_a}{1-\alpha_a};$$

consequently, we have an upper bound on each of α and β. Various investigations of the sequential probability ratio test seem to indicate that in most regular cases, the values of α and β are quite close to α_a and β_a. This prompts us to approximate the power function at the points $\theta = \theta'$ and $\theta = \theta''$ by α_a and $1 - \beta_a$, respectively.

Example 2. Let X be $n(\theta, 100)$. To find the sequential probability ratio test for testing H_0: $\theta = 75$ against H_1: $\theta = 78$ such that each of α and β is approximately equal to 0.10, take

$$k_0 = \frac{0.10}{1-0.10} = \frac{1}{9}, \qquad k_1 = \frac{1-0.10}{0.10} = 9.$$

Since

$$\frac{L(75, n)}{L(78, n)} = \frac{\exp\{-\sum (x_i - 75)^2/2(100)\}}{\exp\{-\sum (x_i - 78)^2/2(100)\}} = \exp\left[-\frac{6\sum x_i - 459n}{200}\right],$$

the inequality

$$k_0 = \frac{1}{9} < \frac{L(75, n)}{L(78, n)} < 9 = k_1$$

can be rewritten, by taking logarithms, as

$$-\ln 9 < -\frac{6\sum x_i - 459n}{200} < \ln 9.$$

This inequality is equivalent to the inequality

$$c_0(n) = \tfrac{153}{2}n - \tfrac{100}{3}\ln 9 < \sum_1^n x_i < \tfrac{153}{2}n + \tfrac{100}{3}\ln 9 = c_1(n).$$

Moreover, $L(75, n)/L(78, n) \leq k_0$ and $L(75, n)/L(78, n) \geq k_1$ are equivalent to the inequalities $\sum_1^n x_i \geq c_1(n)$ and $\sum_1^n x_i \leq c_0(n)$, respectively. Thus, the observation of outcomes is discontinued with the first value n of N for which either $\sum_1^n x_i \geq c_1(n)$ or $\sum_1^n x_i \leq c_0(n)$. The inequality $\sum_1^n x_i \geq c_1(n)$ leads to the rejection of H_0: $\theta = 75$, and the inequality $\sum_1^n x_i \leq c_0(n)$ leads to the acceptance of H_0: $\theta = 75$. The power of the test is approximately 0.10 when H_0 is true, and approximately 0.90 when H_1 is true.

EXERCISES

9.29. Let X be $n(0, \theta)$ and, in the notation of this section, let $\theta' = 4$, $\theta'' = 9$, $\alpha_a = 0.05$, and $\beta_a = 0.10$. Show that the sequential probability ratio test can be based upon the statistic $\sum_1^n X_i^2$. Determine $c_0(n)$ and $c_1(n)$.

9.30. Let X have a Poisson distribution with mean θ. Find the sequential probability ratio test for testing H_0: $\theta = 0.02$ against H_1: $\theta = 0.07$. Show that this test can be based upon the statistic $\sum_1^n X_i$. If $\alpha_a = 0.20$ and $\beta_a = 0.10$, find $c_0(n)$ and $c_1(n)$.

9.31. Let the stochastically independent random variables Y and Z be $n(\mu_1, 1)$ and $n(\mu_2, 1)$, respectively. Let $\theta = \mu_1 - \mu_2$. Let us observe mutually stochastically independent items from each distribution, say $Y_1, Y_2, \ldots$ and $Z_1, Z_2, \ldots$. To test sequentially the hypothesis H_0: $\theta = 0$ against H_1: $\theta = \frac{1}{2}$, use the sequence $X_i = Y_i - Z_i$, $i = 1, 2, \ldots$. If $\alpha_a = \beta_a = 0.05$, show that the test can be based upon $\overline{X} = \overline{Y} - \overline{Z}$. Find $c_0(n)$ and $c_1(n)$.

9.5 Minimax and Bayesian Tests

In Chapter 8 we considered several procedures which may be used in problems of point estimation. Among these were decision function

procedures (in particular, minimax decisions) and Bayesian procedures. In this section, we apply these same principles to the problem of testing a simple hypothesis H_0 against an alternative simple hypothesis H_1. It is important to observe that each of these procedures yields, in accordance with the Neyman-Pearson theorem, a best test of H_0 against H_1.

We first investigate the decision function approach to the problem of testing a simple hypothesis against a simple alternative hypothesis. Let the joint p.d.f. of n random variables $X_1, X_2, \ldots, X_n$ depend upon the parameter θ. Here n is a fixed positive integer. This p.d.f. is denoted by $L(\theta; x_1, x_2, \ldots, x_n)$ or, for brevity, by $L(\theta)$. Let θ' and θ'' be distinct and fixed values of θ. We wish to test the simple hypothesis H_0: $\theta = \theta'$ against the simple hypothesis H_1: $\theta = \theta''$. Thus the parameter space is $\Omega = \{\theta; \theta = \theta', \theta''\}$. In accordance with the decision function procedure, we need a function w of the observed values of $X_1, \ldots, X_n$ (or, of the observed value of a statistic Y) that decides which of the two values of θ, θ' or θ'', to accept. That is, the function w selects either H_0: $\theta = \theta'$ or H_1: $\theta = \theta''$. We denote these decisions by $w = \theta'$ and $w = \theta''$, respectively. Let $\mathscr{L}(\theta, w)$ represent the loss function associated with this decision problem. Because the pairs $(\theta = \theta', w = \theta')$ and $(\theta = \theta'', w = \theta'')$ represent correct decisions, we shall always take $\mathscr{L}(\theta', \theta') = \mathscr{L}(\theta'', \theta'') = 0$. On the other hand, if either $w = \theta''$ when $\theta = \theta'$ or $w = \theta'$ when $\theta = \theta''$, then a positive value should be assigned to the loss function; that is, $\mathscr{L}(\theta', \theta'') > 0$ and $\mathscr{L}(\theta'', \theta') > 0$.

It has previously been emphasized that a test of H_0: $\theta = \theta'$ against H_1: $\theta = \theta''$ can be described in terms of a critical region in the sample space. We can do the same kind of thing with the decision function. That is, we can choose a subset C of the sample space and if $(x_1, x_2, \ldots, x_n) \in C$, we make the decision $w = \theta''$; whereas, if $(x_1, x_2, \ldots, x_n) \in C^*$, the complement of C, we make the decision $w = \theta'$. Thus a given critical region C determines the decision function. In this sense, we may denote the risk function by $R(\theta, C)$ instead of $R(\theta, w)$. That is, in a notation used on p. 274,

$$R(\theta, C) = R(\theta, w) = \int_{C \cup C^*} \mathscr{L}(\theta, w) L(\theta).$$

Since $w = \theta''$ if $(x_1, \ldots, x_n) \in C$ and $w = \theta'$ if $(x_1, \ldots, x_n) \in C^*$, we have

$$R(\theta, C) = \int_C \mathscr{L}(\theta, \theta'') L(\theta) + \int_{C^*} \mathscr{L}(\theta, \theta') L(\theta). \tag{1}$$

If, in Equation (1), we take $\theta = \theta'$, then $\mathscr{L}(\theta', \theta') = 0$ and hence

$$R(\theta', C) = \int_C \mathscr{L}(\theta', \theta'') L(\theta') = \mathscr{L}(\theta', \theta'') \int_C L(\theta').$$

On the other hand, if in Equation (1) we let $\theta = \theta''$, then $\mathscr{L}(\theta'', \theta'') = 0$ and, accordingly,

$$R(\theta'', C) = \int_{C^*} \mathscr{L}(\theta'', \theta')L(\theta'') = \mathscr{L}(\theta'', \theta') \int_{C^*} L(\theta'').$$

It is enlightening to note that, if $K(\theta)$ is the power function of the test associated with the critical region C, then

$$R(\theta', C) = \mathscr{L}(\theta', \theta'')K(\theta') = \mathscr{L}(\theta', \theta'')\alpha,$$

where $\alpha = K(\theta')$ is the significance level; and

$$R(\theta'', C) = \mathscr{L}(\theta'', \theta')[1 - K(\theta'')] = \mathscr{L}(\theta'', \theta')\beta,$$

where $\beta = 1 - K(\theta'')$ is the probability of the Type-II error.

Let us now see if we can find a minimax solution to our problem. That is, we want to find a critical region C so that

$$\max [R(\theta', C), R(\theta'', C)]$$

is minimized. We shall show that the solution is the region

$$C = \left\{(x_1, \ldots, x_n); \frac{L(\theta'; x_1, \ldots, x_n)}{L(\theta''; x_1, \ldots, x_n)} \leq k\right\},$$

provided the positive constant k is selected so that $R(\theta', C) = R(\theta'', C)$. That is, if k is chosen so that

$$\mathscr{L}(\theta', \theta'') \int_C L(\theta') = \mathscr{L}(\theta'', \theta') \int_{C^*} L(\theta''),$$

then the critical region C provides a minimax solution. In the case of random variables of the continuous type, k can always be selected so that $R(\theta', C) = R(\theta'', C)$. However, with random variables of the discrete type, we may need to consider an auxiliary random experiment when $L(\theta')/L(\theta'') = k$ in order to achieve the exact equality $R(\theta', C) = R(\theta'', C)$.

To see that this region C is the minimax solution, consider every other region A for which $R(\theta', C) \geq R(\theta', A)$. Obviously, a region A for which $R(\theta', C) < R(\theta', A)$ is not a candidate for a minimax solution, for then $R(\theta', C) = R(\theta'', C) < \max [R(\theta', A), R(\theta'', A)]$. Since $R(\theta', C) \geq R(\theta', A)$ means that

$$\mathscr{L}(\theta', \theta'') \int_C L(\theta') \geq \mathscr{L}(\theta', \theta'') \int_A L(\theta'),$$

we have

$$\alpha = \int_C L(\theta') \geq \int_A L(\theta').$$

That is, the significance level of the test associated with the critical region A is less than or equal to α. But C, in accordance with the Neyman-Pearson theorem, is a best critical region of size α. Thus,

$$\int_C L(\theta'') \geq \int_A L(\theta'')$$

and

$$\int_{C^*} L(\theta'') \leq \int_{A^*} L(\theta'').$$

Accordingly,

$$\mathscr{L}(\theta'', \theta') \int_{C^*} L(\theta'') \leq \mathscr{L}(\theta'', \theta') \int_{A^*} L(\theta''),$$

or, equivalently,

$$R(\theta'', C) \leq R(\theta'', A).$$

That is,

$$R(\theta', C) = R(\theta'', C) \leq R(\theta'', A).$$

This means that

$$\max [R(\theta', C), R(\theta'', C)] \leq R(\theta'', A).$$

Then certainly

$$\max [R(\theta', C), R(\theta'', C)] \leq \max [R(\theta', A), R(\theta'', A)],$$

and the critical region C provides a minimax solution, as we wanted to show.

Example 1. Let $X_1, X_2, \ldots, X_{100}$ denote a random sample of size 100 from a distribution that is $n(\theta, 100)$. We again consider the problem of testing H_0: $\theta = 75$ against H_1: $\theta = 78$. We seek a minimax solution with $\mathscr{L}(75, 78) = 3$ and $\mathscr{L}(78, 75) = 1$. Since $L(75)/L(78) \leq k$ is equivalent to $\bar{x} \geq c$, we want to determine c, and thus k, so that

$$3 \Pr (\bar{X} \geq c; \theta = 75) = \Pr (\bar{X} < c; \theta = 78).$$

Because $\bar{X}$ is $n(\theta, 1)$, the preceding equation can be rewritten as

$$3[1 - N(c - 75)] = N(c - 78).$$

If we use Table III of the Appendix, we see, by trial and error, that the solution is $c = 76.8$, approximately. The significance level of the test is $1 - N(1.8) = 0.036$, approximately, and the power of the test when H_1 is true is $1 - N(-1.2) = 0.885$, approximately.

Next, let us consider the Bayesian approach to the problem of testing the simple hypothesis H_0: $\theta = \theta'$ against the simple hypothesis H_1: $\theta = \theta''$. We continue to use the notation already presented in

this section. In addition, we recall that we need the p.d.f. $h(\theta)$ of the random variable Θ. Since the parameter space consists of but two points, θ' and θ'', Θ is a random variable of the discrete type; and we have $h(\theta') + h(\theta'') = 1$. Since $L(\theta; x_1, x_2, \ldots, x_n) = L(\theta)$ is the conditional p.d.f. of $X_1, X_2, \ldots, X_n$, given $\Theta = \theta$, the joint p.d.f. of $X_1, X_2, \ldots, X_n$ and Θ is

$$h(\theta)L(\theta; x_1, x_2, \ldots, x_n) = h(\theta)L(\theta).$$

Because

$$\sum_{\Omega} h(\theta)L(\theta) = h(\theta')L(\theta') + h(\theta'')L(\theta'')$$

is the marginal p.d.f. of $X_1, X_2, \ldots, X_n$, the conditional p.d.f. of Θ, given $X_1 = x_1, \ldots, X_n = x_n$, is

$$k(\theta|x_1, \ldots, x_n) = \frac{h(\theta)L(\theta)}{h(\theta')L(\theta') + h(\theta'')L(\theta'')}.$$

Now a Bayes' solution to a decision problem is defined in Section 8.4 as a $w(y)$ such that $E[\mathscr{L}(\theta, w(y))|Y = y]$ is a minimum. In this problem if $w = \theta'$, the conditional expectation of $\mathscr{L}(\theta, w)$, given $X_1 = x_1, \ldots, X_n = x_n$, is

$$\sum_{\Omega} \mathscr{L}(\theta, \theta')k(\theta|x_1, \ldots, x_n) = \frac{\mathscr{L}(\theta'', \theta')h(\theta'')L(\theta'')}{h(\theta')L(\theta') + h(\theta'')L(\theta'')},$$

because $\mathscr{L}(\theta', \theta') = 0$; and if $w = \theta''$, this expectation is

$$\sum_{\Omega} \mathscr{L}(\theta, \theta'')k(\theta|x_1, \ldots, x_n) = \frac{\mathscr{L}(\theta', \theta'')h(\theta')L(\theta')}{h(\theta')L(\theta') + h(\theta'')L(\theta'')},$$

because $\mathscr{L}(\theta'', \theta'') = 0$. Accordingly, the Bayes' solution requires that the decision $w = \theta''$ be made if

$$\frac{\mathscr{L}(\theta', \theta'')h(\theta')L(\theta')}{h(\theta')L(\theta') + h(\theta'')L(\theta'')} < \frac{\mathscr{L}(\theta'', \theta')h(\theta'')L(\theta'')}{h(\theta')L(\theta') + h(\theta'')L(\theta'')},$$

or, equivalently, if

$$\frac{L(\theta')}{L(\theta'')} < \frac{\mathscr{L}(\theta'', \theta')h(\theta'')}{\mathscr{L}(\theta', \theta'')h(\theta')}. \tag{2}$$

If the sign of inequality in expression (2) is reversed, we make the decision $w = \theta'$; and if the two members of expression (2) are equal, we can use some auxiliary random experiment to make the decision. It is important to note that expression (2) describes, in accordance with the Neyman-Pearson theorem, a best test.

Example 2. In addition to the information given in Example 1, suppose we know the prior probabilities for $\theta = \theta' = 75$ and for $\theta = \theta'' = 78$ to be given, respectively, by $h(75) = \frac{1}{7}$ and $h(78) = \frac{6}{7}$. Then the Bayes' solution is, in this case,

$$\frac{L(75)}{L(78)} < \frac{(1)(\frac{6}{7})}{(3)(\frac{1}{7})} = 2,$$

which is equivalent to $\bar{x} > 76.3$, approximately. The power of the test when H_0 is true is $1 - N(1.3) = 0.097$, approximately, and the power of the test when H_1 is true is $1 - N(-1.7) = N(1.7) = 0.955$, approximately.

In summary, we make the following comments. In testing the simple hypothesis H_0: $\theta = \theta'$ against the simple hypothesis H_1: $\theta = \theta''$, it is emphasized that each principle leads to critical regions of the form

$$\left\{(x_1, x_2, \ldots, x_n); \frac{L(\theta'; x_1, \ldots, x_n)}{L(\theta''; x_1, \ldots, x_n)} \leq k\right\},$$

where k is a positive constant. In the classical approach, we determine k by requiring that the power function of the test have a certain value at the point $\theta = \theta'$ or at the point $\theta = \theta''$ (usually, the value α at the point $\theta = \theta'$). The minimax decision requires k to be selected so that

$$\mathscr{L}(\theta', \theta'') \int_C L(\theta') = \mathscr{L}(\theta'', \theta') \int_{C^*} L(\theta'').$$

Finally, the Bayes' procedure requires that

$$k = \frac{\mathscr{L}(\theta'', \theta')h(\theta'')}{\mathscr{L}(\theta', \theta'')h(\theta')}.$$

Each of these tests is a best test for testing a simple hypothesis H_0: $\theta = \theta'$ against a simple alternative hypothesis H_1: $\theta = \theta''$.

EXERCISES

9.32. Let $X_1, X_2, \ldots, X_{20}$ be a random sample of size 20 from a distribution which is $n(\theta, 5)$. Let $L(\theta)$ represent the joint p.d.f. of $X_1, X_2, \ldots, X_{20}$. The problem is to test H_0: $\theta = 1$ against H_1: $\theta = 0$. Thus $\Omega = \{\theta; \theta = 0, 1\}$.

(a) Show that $L(1)/L(0) \leq k$ is equivalent to $\bar{x} \leq c$.

(b) Find c so that the significance level is $\alpha = 0.05$. Compute the power of this test if H_1 is true.

(c) If the loss function is such that $\mathscr{L}(1, 1) = \mathscr{L}(0, 0) = 0$ and $\mathscr{L}(1, 0) = \mathscr{L}(0, 1) > 0$, find the minimax test. Evaluate the power function of this test at the points $\theta = 1$ and $\theta = 0$.

(d) If, in addition, the prior probabilities of $\theta = 1$ and $\theta = 0$ are, respectively, $h(1) = \frac{3}{4}$ and $h(0) = \frac{1}{4}$, find the Bayes' test. Evaluate the power function of this test at the points $\theta = 1$ and $\theta = 0$.

9.33. Let $X_1, X_2, \ldots, X_{10}$ be a random sample of size 10 from a Poisson distribution with parameter θ. Let $L(\theta)$ be the joint p.d.f. of $X_1, X_2, \ldots, X_{10}$. The problem is to test H_0: $\theta = \frac{1}{2}$ against H_1: $\theta = 1$.

(a) Show that $L(\frac{1}{2})/L(1) \leq k$ is equivalent to $y = \sum_1^{10} x_i \geq c$.

(b) In order to make $\alpha = 0.05$, show that H_0 is rejected if $y > 9$ and, if $y = 9$, reject H_0 with probability $\frac{1}{2}$ (using some auxiliary random experiment).

(c) If the loss function is such that $\mathscr{L}(\frac{1}{2}, \frac{1}{2}) = \mathscr{L}(1, 1) = 0$ and $\mathscr{L}(\frac{1}{2}, 1) = 1$ and $\mathscr{L}(1, \frac{1}{2}) = 2$, show that the minimax procedure is to reject H_0 if $y > 6$ and, if $y = 6$, reject H_0 with probability 0.08 (using some auxiliary random experiment).

(d) If, in addition, we are given that the prior probabilities of $\theta = \frac{1}{2}$ and $\theta = 1$ are $h(\frac{1}{2}) = \frac{1}{3}$ and $h(1) = \frac{2}{3}$, respectively, show that the Bayes' solution is to reject H_0 if $y > 5.2$, that is, reject H_0 if $y \geq 6$.

Chapter 10

Other Statistical Tests

10.1 Likelihood Ratio Tests

The notion of using the magnitude of the ratio of two probability density functions as the basis of a best test or of a uniformly most powerful test can be modified, and made intuitively appealing, to provide a method of constructing a test of a composite hypothesis against an alternative composite hypothesis or of constructing a test of a simple hypothesis against an alternative composite hypothesis when a uniformly most powerful test does not exist. This method leads to tests that are called "likelihood ratio tests." A likelihood ratio test, as just remarked, is not necessarily a uniformly most powerful test, but it has been proved in the literature that such a test often has desirable properties.

A certain terminology and notation will be introduced by means of an example.

Example 1. Let the random variable X be $n(\theta_1, \theta_2)$ and let the parameter space be $\Omega = \{(\theta_1, \theta_2); -\infty < \theta_1 < \infty, 0 < \theta_2 < \infty\}$. Let the composite hypothesis be H_0: $\theta_1 = 0$, $\theta_2 > 0$, and let the alternative composite hypothesis be H_1: $\theta_1 \neq 0$, $\theta_2 > 0$. The set $\omega = \{(\theta_1, \theta_2); \theta_1 = 0, 0 < \theta_2 < \infty\}$ is a subset of Ω and will be called the *subspace* specified by the hypothesis H_0. Then, for instance, the hypothesis H_0 may be described as H_0: $(\theta_1, \theta_2) \in \omega$. It is proposed that we test H_0 against all alternatives in H_1.

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size $n > 1$ from the

distribution of this example. The joint p.d.f. of $X_1, X_2, \ldots, X_n$ is, at each point in Ω,

$$L(\theta_1, \theta_2; x_1, \ldots, x_n) = \left(\frac{1}{2\pi\theta_2}\right)^{n/2} \exp\left[-\frac{\sum_1^n (x_i - \theta_1)^2}{2\theta_2}\right] = L(\Omega).$$

At each point $(\theta_1, \theta_2) \in \omega$, the joint p.d.f. of $X_1, X_2, \ldots, X_n$ is

$$L(0, \theta_2; x_1, \ldots, x_n) = \left(\frac{1}{2\pi\theta_2}\right)^{n/2} \exp\left[-\frac{\sum_1^n x_i^2}{2\theta_2}\right] = L(\omega).$$

The joint p.d.f., now denoted by $L(\omega)$, is not completely specified, since θ_2 may be any positive number; nor is the joint p.d.f., now denoted by $L(\Omega)$, completely specified, since θ_1 may be any real number and θ_2 any positive number. Thus the ratio of $L(\omega)$ to $L(\Omega)$ could not provide a basis for a test of H_0 against H_1. Suppose, however, we modify this ratio in the following manner: We shall find the maximum of $L(\omega)$ in ω, that is, the maximum of $L(\omega)$ with respect to θ_2. And we shall find the maximum of $L(\Omega)$ in Ω; that is, the maximum of $L(\Omega)$ with respect to θ_1 and θ_2. The ratio of these maxima will be taken as the criterion for a test of H_0 against H_1. Let the maximum of $L(\omega)$ in ω be denoted by $L(\hat{\omega})$ and let the maximum of $L(\Omega)$ in Ω be denoted by $L(\hat{\Omega})$. Then the criterion for the test of H_0 against H_1 is the likelihood ratio

$$\lambda(x_1, x_2, \ldots, x_n) = \lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})}.$$

Since $L(\omega)$ and $L(\Omega)$ are probability density functions, $\lambda \geq 0$; and since ω is a subset of Ω, $\lambda \leq 1$.

In our example, the maximum, $L(\hat{\omega})$, of $L(\omega)$ is obtained by first setting

$$\frac{d \ln L(\omega)}{d\theta_2} = -\frac{n}{2\theta_2} + \frac{\sum_1^n x_i^2}{2\theta_2^2}$$

equal to zero and solving for θ_2. The solution for θ_2 is $\sum_1^n x_i^2/n$, and this number maximizes $L(\omega)$. Thus the maximum is

$$L(\hat{\omega}) = \left(\frac{1}{2\pi \sum_1^n x_i^2/n}\right)^{n/2} \exp\left(-\frac{\sum_1^n x_i^2}{2\sum_1^n x_i^2/n}\right)$$

$$= \left(\frac{ne^{-1}}{2\pi \sum_1^n x_i^2}\right)^{n/2}.$$

On the other hand, by using Example 4, p. 257, the maximum, $L(\hat{\Omega})$, of $L(\Omega)$ is obtained by replacing θ_1 and θ_2 by $\sum_1^n x_i/n = \bar{x}$ and $\sum_1^n (x_i - \bar{x})^2/n$, respectively. That is,

$$L(\hat{\Omega}) = \left[\frac{1}{2\pi \sum_1^n (x_i - \bar{x})^2/n}\right]^{n/2} \exp\left[-\frac{\sum_1^n (x_i - \bar{x})^2}{2\sum_1^n (x_i - \bar{x})^2/n}\right]$$
$$= \left[\frac{ne^{-1}}{2\pi \sum_1^n (x_i - \bar{x})^2}\right]^{n/2}.$$

Thus, here

$$\lambda = \left[\frac{\sum_1^n (x_i - \bar{x})^2}{\sum_1^n x_i^2}\right]^{n/2}.$$

Because $\sum_1^n x_i^2 = \sum_1^n (x_i - \bar{x})^2 + n\bar{x}^2$, λ may be written

$$\lambda = \frac{1}{\left\{1 + \left[n\bar{x}^2/\sum_1^n (x_i - \bar{x})^2\right]\right\}^{n/2}}.$$

Now the hypothesis H_0 is $\theta_1 = 0$, $\theta_2 > 0$. If the observed number $\bar{x}$ were zero, the experiment tends to confirm H_0. But if $\bar{x} = 0$ and $\sum_1^n x_i^2 > 0$, then $\lambda = 1$. On the other hand, if $\bar{x}$ and $n\bar{x}^2/\sum_1^n (x_i - \bar{x})^2$ deviate considerably from zero, the experiment tends to negate H_0. Now the greater the deviation of $n\bar{x}^2/\sum_1^n (x_i - \bar{x})^2$ from zero, the smaller λ becomes. That is, if λ is used as a test criterion, then an intuitively appealing critical region for testing H_0 is a set defined by $0 \leq \lambda \leq \lambda_0$, where λ_0 is a positive proper fraction. Thus we reject H_0 if $\lambda \leq \lambda_0$. A test that has the critical region $\lambda \leq \lambda_0$ is a *likelihood ratio test.* In this example, $\lambda \leq \lambda_0$ when and only when

$$\frac{\sqrt{n}\,|\bar{x}|}{\sqrt{\sum_1^n (x_i - \bar{x})^2/(n-1)}} \geq \sqrt{(n-1)(\lambda_0^{-2/n} - 1)} = c.$$

If H_0: $\theta_1 = 0$ is true, the results on p. 195 show that the statistic

$$t(X_1, X_2, \ldots, X_n) = \frac{\sqrt{n}\,(\bar{X} - 0)}{\sqrt{\sum_1^n (X_i - \bar{X})^2/(n-1)}}$$

has a t distribution with $n - 1$ degrees of freedom. Accordingly, in this example the likelihood ratio test of H_0 against H_1 may be based on a T

statistic. For a given positive integer n, Table IV in the Appendix may be used (with $n - 1$ degrees of freedom) to detetermine the number c such that $\alpha = \Pr[|t(X_1, X_2, \ldots, X_n)| \geq c; H_0]$ is the desired significance level of the test. If the experimental values of $X_1, X_2, \ldots, X_n$ are, respectively, $x_1, x_2, \ldots, x_n$, then we reject H_0 if and only if $|t(x_1, x_2, \ldots, x_n)| \geq c$. If, for instance, $n = 6$ and $\alpha = 0.05$, then from Table IV, $c = 2.571$.

The preceding example should make the following generalization easier to read: Let $X_1, X_2, \ldots, X_n$ denote n mutually stochastically independent random variables having, respectively, the probability density functions $f_i(x_i; \theta_1, \theta_2, \ldots, \theta_m)$, $i = 1, 2, \ldots, n$. The set that consists of all parameter points $(\theta_1, \theta_2, \ldots, \theta_m)$ is denoted by Ω, which we have called the parameter space. Let ω be a subset of the parameter space Ω. We wish to test the (simple or composite) hypothesis H_0: $(\theta_1, \theta_2, \ldots, \theta_m) \in \omega$ against all alternative hypotheses. Define the likelihood functions

$$L(\omega) = \prod_{i=1}^{n} f_i(x_i; \theta_1, \theta_2, \ldots, \theta_m), \qquad (\theta_1, \theta_2, \ldots, \theta_m) \in \omega,$$

and

$$L(\Omega) = \prod_{i=1}^{n} f_i(x_i; \theta_1, \theta_2, \ldots, \theta_m), \qquad (\theta_1, \theta_2, \ldots, \theta_m) \in \Omega.$$

Let $L(\hat{\omega})$ and $L(\hat{\Omega})$ be the maxima, which we assume to exist, of these two likelihood functions. The ratio of $L(\hat{\omega})$ to $L(\hat{\Omega})$ is called the *likelihood ratio* and is denoted by

$$\lambda(x_1, x_2, \ldots, x_n) = \lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})}.$$

Let λ_0 be a positive proper fraction. The *likelihood ratio test principle* states that the hypothesis H_0: $(\theta_1, \theta_2, \ldots, \theta_m) \in \omega$ is rejected if and only if

$$\lambda(x_1, x_2, \ldots, x_n) = \lambda \leq \lambda_0.$$

The function λ defines a random variable $\lambda(X_1, X_2, \ldots, X_n)$, and the significance level of the test is given by

$$\alpha = \Pr[\lambda(X_1, X_2, \ldots, X_n) \leq \lambda_0; H_0].$$

The likelihood ratio test principle is an intuitive one. However, the principle does lead to the same test, when testing a simple hypothesis H_0 against an alternative simple hypothesis H_1, as that given by the Neyman-Pearson theorem (Exercise 10.3). Thus it might be expected that a test based on this principle has some desirable properties.

An example of the preceding generalization will be given.

Example 2. Let the stochastically independent random variables X and Y have distributions that are $n(\theta_1, \theta_3)$ and $n(\theta_2, \theta_3)$, where the means θ_1 and θ_2 and common variance θ_3 are unknown. Then $\Omega = \{(\theta_1, \theta_2, \theta_3); -\infty < \theta_1 < \infty, -\infty < \theta_2 < \infty, 0 < \theta_3 < \infty\}$. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ denote independent random samples from these distributions. The hypothesis H_0: $\theta_1 = \theta_2$, unspecified, and θ_3 unspecified, is to be tested against all alternatives. Then $\omega = \{(\theta_1, \theta_2, \theta_3); -\infty < \theta_1 = \theta_2 < \infty, 0 < \theta_3 < \infty\}$. Here $X_1, X_2, \ldots, X_n$, $Y_1, Y_2, \ldots, Y_m$ are $n + m > 2$ mutually stochastically independent random variables having the likelihood functions

$$L(\omega) = \left(\frac{1}{2\pi\theta_3}\right)^{(n+m)/2} \exp\left[-\frac{\sum_1^n (x_i - \theta_1)^2 + \sum_1^m (y_i - \theta_1)^2}{2\theta_3}\right]$$

and

$$L(\Omega) = \left(\frac{1}{2\pi\theta_3}\right)^{(n+m)/2} \exp\left[-\frac{\sum_1^n (x_i - \theta_1)^2 + \sum_1^m (y_i - \theta_2)^2}{2\theta_3}\right].$$

If

$$\frac{\partial \ln L(\omega)}{\partial \theta_1} \quad \text{and} \quad \frac{\partial \ln L(\omega)}{\partial \theta_3}$$

are equated to zero, then (Exercise 10.4)

$$\sum_1^n (x_i - \theta_1) + \sum_1^m (y_i - \theta_1) = 0,$$

(1)

$$-(n + m) + \frac{1}{\theta_3}\left[\sum_1^n (x_i - \theta_1)^2 + \sum_1^m (y_i - \theta_1)^2\right] = 0.$$

The solutions for θ_1 and θ_3 are, respectively,

$$u = \frac{\sum_1^n x_i + \sum_1^m y_i}{n + m}$$

and

$$w = \frac{\sum_1^n (x_i - u)^2 + \sum_1^m (y_i - u)^2}{n + m},$$

and u and w maximize $L(\omega)$. The maximum is

$$L(\hat{\omega}) = \left(\frac{e^{-1}}{2\pi w}\right)^{(n+m)/2}.$$

In like manner, if

$$\frac{\partial \ln L(\Omega)}{\partial \theta_1}, \quad \frac{\partial \ln L(\Omega)}{\partial \theta_2}, \quad \frac{\partial \ln L(\Omega)}{\partial \theta_3}$$

are equated to zero, then (Exercise 10.5)

$$\sum_1^n (x_i - \theta_1) = 0,$$

$$\sum_1^m (y_i - \theta_2) = 0, \tag{2}$$

$$-(n+m) + \frac{1}{\theta_3}\left[\sum_1^n (x_i - \theta_1)^2 + \sum_1^m (y_i - \theta_2)^2\right] = 0.$$

The solutions for θ_1, θ_2, and θ_3 are, respectively,

$$u_1 = \frac{\sum_1^n x_i}{n},$$

$$u_2 = \frac{\sum_1^m y_i}{m},$$

$$w' = \frac{\sum_1^n (x_i - u_1)^2 + \sum_1^m (y_i - u_2)^2}{n+m},$$

and u_1, u_2, and w' maximize $L(\Omega)$. The maximum is

$$L(\hat{\Omega}) = \left(\frac{e^{-1}}{2\pi w'}\right)^{(n+m)/2},$$

so that

$$\lambda(x_1, \ldots, x_n, y_1, \ldots, y_m) = \lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \left(\frac{w'}{w}\right)^{(n+m)/2}.$$

The random variable defined by $\lambda^{2/(n+m)}$ is

$$\frac{\sum_1^n (X_i - \overline{X})^2 + \sum_1^m (Y_i - \overline{Y})^2}{\sum_1^n \{X_i - [(n\overline{X} + m\overline{Y})/(n+m)]\}^2 + \sum_1^m \{Y_i - [(n\overline{X} + m\overline{Y})/(n+m)]\}^2}.$$

Now

$$\sum_1^n \left(X_i - \frac{n\overline{X} + m\overline{Y}}{n+m}\right)^2 = \sum_1^n \left[(X_i - \overline{X}) + \left(\overline{X} - \frac{n\overline{X} + m\overline{Y}}{n+m}\right)\right]^2$$

$$= \sum_1^n (X_i - \overline{X})^2 + n\left(\overline{X} - \frac{n\overline{X} + m\overline{Y}}{n+m}\right)^2$$

and

$$\sum_1^m \left(Y_i - \frac{n\overline{X} + m\overline{Y}}{n+m}\right)^2 = \sum_1^m \left[(Y_i - \overline{Y}) + \left(\overline{Y} - \frac{n\overline{X} + m\overline{Y}}{n+m}\right)\right]^2$$

$$= \sum_1^m (Y_i - \overline{Y})^2 + m\left(\overline{Y} - \frac{n\overline{X} + m\overline{Y}}{n+m}\right)^2.$$

But

$$n\left(\bar{X} - \frac{n\bar{X} + m\bar{Y}}{n+m}\right)^2 = \frac{m^2 n}{(n+m)^2}(\bar{X} - \bar{Y})^2$$

and

$$m\left(\bar{Y} - \frac{n\bar{X} + m\bar{Y}}{n+m}\right)^2 = \frac{n^2 m}{(n+m)^2}(\bar{X} - \bar{Y})^2.$$

Hence the random variable defined by $\lambda^{2/(n+m)}$ may be written

$$\frac{\sum_1^n (X_i - \bar{X})^2 + \sum_1^m (Y_i - \bar{Y})^2}{\sum_1^n (X_i - \bar{X})^2 + \sum_1^m (Y_i - \bar{Y})^2 + [nm/(n+m)](\bar{X} - \bar{Y})^2}$$

$$= \frac{1}{1 + \dfrac{[nm/(n+m)](\bar{X} - \bar{Y})^2}{\sum_1^n (X_i - \bar{X})^2 + \sum_1^m (Y_i - \bar{Y})^2}}.$$

If the hypothesis H_0: $\theta_1 = \theta_2$ is true, the random variable

$$T = \frac{\sqrt{\dfrac{nm}{n+m}}(\bar{X} - \bar{Y})}{\sqrt{\dfrac{\sum_1^n (X_i - \bar{X})^2 + \sum_1^m (Y_i - \bar{Y})^2}{n+m-2}}}$$

has, in accordance with p. 200, a t distribution with $n + m - 2$ degrees of freedom. Thus, the random variable defined by $\lambda^{2/(n+m)}$ is

$$\frac{n+m-2}{(n+m-2) + T^2}.$$

The test of H_0 against all alternatives may then be based on a t distribution with $n + m - 2$ degrees of freedom.

The likelihood ratio principle calls for the rejection of H_0 if and only if $\lambda \leq \lambda_0 < 1$. Thus, the significance level of the test is

$$\alpha = \Pr[\lambda(X_1, \ldots, X_n, Y_1, \ldots, Y_m) \leq \lambda_0; H_0].$$

However, $\lambda(X_1, \ldots, X_n, Y_1, \ldots, Y_m) \leq \lambda_0$ is equivalent to $|T| \geq c$, and so

$$\alpha = \Pr(|T| \geq c; H_0).$$

For given values of n and m, the number c is determined from Table IV in the Appendix (with $n + m - 2$ degrees of freedom) in such a manner as to yield a desired α. Then H_0 is rejected at a significance level α if and only if $|t| \geq c$, where t is the experimental value of T. If, for instance, $n = 10$, $m = 6$, and $\alpha = 0.05$, then $c = 2.145$.

In each of the two examples of this section, it was found that the likelihood ratio test could be based on a statistic which, when the

hypothesis H_0 is true, has a t distribution. To help us compute the powers of these tests at parameter points other than those described by the hypothesis H_0, we turn to the following definition.

Definition 1. Let the random variable W be $n(\delta, 1)$; let the random variable V be $\chi^2(r)$, and let W and V be stochastically independent. The quotient

$$T = \frac{W}{\sqrt{V/r}}$$

is said to have a *noncentral t distribution* with r degrees of freedom and noncentrality parameter δ. If $\delta = 0$, we say that T has a central t distribution.

In the light of this definition, let us reexamine the statistics of the examples of this section. In Example 1 we had

$$t(X_1, \ldots, X_n) = \frac{\sqrt{n}\bar{X}}{\sqrt{\sum_1^n (X_i - \bar{X})^2/(n-1)}} = \frac{\sqrt{n}\bar{X}/\sigma}{\sqrt{\sum_1^n (X_i - \bar{X})^2/[\sigma^2(n-1)]}}.$$

Here $W_1 = \sqrt{n}\bar{X}/\sigma$ is $n(\sqrt{n}\theta_1/\sigma, 1)$, $V_1 = \sum_1^n (X_i - \bar{X})^2/\sigma^2$ is $\chi^2(n-1)$, and W_1 and V_1 are stochastically independent. Thus, if $\theta_1 \neq 0$, we see, in accordance with the definition, that $t(X_1, \ldots, X_n)$ has a noncentral t distribution with $n - 1$ degrees of freedom and noncentrality parameter $\delta_1 = \sqrt{n}\theta_1/\sigma$. In Example 2 we had

$$T = \frac{W_2}{\sqrt{V_2/(n+m-2)}},$$

where

$$W_2 = \sqrt{\frac{nm}{n+m}}\,(\bar{X} - \bar{Y})/\sigma$$

and

$$V_2 = \left[\sum_1^n (X_i - \bar{X})^2 + \sum_1^m (Y_i - \bar{Y})^2\right]/\sigma^2.$$

Here W_2 is $n[\sqrt{nm/(n+m)}(\theta_1 - \theta_2)/\sigma, 1]$, V_2 is $\chi^2(n+m-2)$, and W_2 and V_2 are stochastically independent. Accordingly, if $\theta_1 \neq \theta_2$, T has a noncentral t distribution with $n + m - 2$ degrees of freedom and noncentrality parameter $\delta_2 = \sqrt{nm/(n+m)}(\theta_1 - \theta_2)/\sigma$. It is interest-

ing to note that $\delta_1 = \sqrt{n}\theta_1/\sigma$ measures the deviation of θ_1 from $\theta_1 = 0$ in units of the standard deviation $\sigma/\sqrt{n}$ of $\bar{X}$. The noncentrality parameter $\delta_2 = \sqrt{nm/(n + m)}(\theta_1 - \theta_2)/\sigma$ is equal to the deviation of $\theta_1 - \theta_2$ from $\theta_1 - \theta_2 = 0$ in units of the standard deviation $\sigma\sqrt{(n + m)/nm}$ of $\bar{X} - \bar{Y}$.

There are various tables of the noncentral t distribution, but they are much too cumbersome to be included in this book. However, with the aid of such tables, we can determine the power functions of these tests as functions of the noncentrality parameters.

In Example 2, in testing the equality of the means of two independent normal distributions, it was assumed that the unknown variances of the distributions were equal. Let us now consider the problem of testing the equality of these two unknown variances.

Example 3. We are given the stochastically independent random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ from the independent distributions which are $n(\theta_1, \theta_3)$ and $n(\theta_2, \theta_4)$, respectively. We have

$$\Omega = \{(\theta_1, \theta_2, \theta_3, \theta_4);\ -\infty < \theta_1, \theta_2 < \infty,\ 0 < \theta_3, \theta_4 < \infty\}.$$

The hypothesis H_0: $\theta_3 = \theta_4$, unspecified, with θ_1 and θ_2 also unspecified, is to be tested against all alternatives. Then

$$\omega = \{(\theta_1, \theta_2, \theta_3, \theta_4);\ -\infty < \theta_1, \theta_2 < \infty,\ 0 < \theta_3 = \theta_4 < \infty\}.$$

It is easy to show (see Exercise 10.10) that the statistic defined by $\lambda = L(\hat{\omega})/L(\hat{\Omega})$ is a function of the statistic

$$F = \frac{\sum_1^n (X_i - \bar{X})^2/(n - 1)}{\sum_1^m (Y_i - \bar{Y})^2/(m - 1)}.$$

If $\theta_3 = \theta_4$, this statistic F has an F distribution with $n - 1$ and $m - 1$ degrees of freedom. The hypothesis that $(\theta_1, \theta_2, \theta_3, \theta_4) \in \omega$ is rejected if the computed $F \leq c_1$ or if the computed $F \geq c_2$. The constants c_1 and c_2 are usually selected so that, if $\theta_3 = \theta_4$,

$$\Pr(F \leq c_1) = \Pr(F \geq c_2) = \frac{\alpha_1}{2},$$

where α_1 is the desired significance level of this test.

Recall that in Example 2 a statistic, which was denoted by T, was used to test the hypothesis that $\theta_1 = \theta_2$, provided the unknown variances θ_3 and θ_4 were equal. The hypothesis that $\theta_1 = \theta_2$ is rejected if the computed $|T| \geq c$, where the constant c is selected so that

$\alpha_2 = \Pr(|T| \geq c; \theta_1 = \theta_2, \theta_3 = \theta_4)$ is the assigned significance level of the test. We will show that, if $\theta_3 = \theta_4$, F and T are stochastically independent. Among other things, this means that if these two tests are performed sequentially, with respective significance levels α_1 and α_2, the probability of accepting both these hypotheses, when they are true, is $(1 - \alpha_1)(1 - \alpha_2)$. Thus the significance level of this joint test is $\alpha = 1 - (1 - \alpha_1)(1 - \alpha_2)$.

The stochastic independence of F and T, when $\theta_3 = \theta_4$, can be established by an appeal to sufficiency and completeness. The three statistics $\bar{X}$, $\bar{Y}$, and $\sum_1^n (X_i - \bar{X})^2 + \sum_1^m (Y_i - \bar{Y})^2$ are joint complete sufficient statistics for the three parameters θ_1, θ_2, and $\theta_3 = \theta_4$. Obviously, the distribution of F does not depend upon θ_1, θ_2, and $\theta_3 = \theta_4$ and hence F is stochastically independent of the three joint complete sufficient statistics. However, T is a function of these three joint complete sufficient statistics alone, and, accordingly, T is stochastically independent of F. It is important to note that these two statistics are stochastically independent whether $\theta_1 = \theta_2$ or $\theta_1 \neq \theta_2$, that is, whether T is or is not central. This permits us to calculate probabilities other than the significance level of the test. For example, if $\theta_3 = \theta_4$ and $\theta_1 \neq \theta_2$, then

$$\Pr(c_1 < F < c_2, |T| \geq c) = \Pr(c_1 < F < c_2) \Pr(|T| \geq c).$$

The second factor in the right-hand member is evaluated by using the probabilities for a noncentral t distribution. Of course, if $\theta_3 = \theta_4$ and the difference $\theta_1 - \theta_2$ is large, we would want the preceding probability to be close to one because the event $\{c_1 < F < c_2, |T| \geq c\}$ leads to a correct decision, namely, accept $\theta_3 = \theta_4$ and reject $\theta_1 = \theta_2$.

EXERCISES

10.1. In Example 1, let $n = 10$, and let the experimental values of the random variables yield $\bar{x} = 0.6$ and $\sum_1^{10} (x_i - \bar{x})^2 = 3.6$. If the test derived in that example is used, do we accept or reject H_0: $\theta_1 = 0$ at the 5 per cent significance level?

10.2. In Example 2, let $n = m = 8$, $\bar{x} = 75.2$, $\bar{y} = 78.6$, $\sum_1^8 (x_i - \bar{x})^2 = 71.2$, $\sum_1^8 (y_i - \bar{y})^2 = 54.8$. If we use the test derived in that example, do we accept or reject H_0: $\theta_1 = \theta_2$ at the 5 per cent significance level?

10.3. Show that the likelihood ratio principle leads to the same test, when testing a simple hypothesis H_0 against an alternative simple hypothesis H_1, as that given by the Neyman-Pearson theorem. Note that there are only two points in Ω.

10.4. Verify Equations (1) of Example 2 of this section.

10.5. Verify Equations (2) of Example 2 of this section.

10.6. Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution $n(\theta, 1)$. Show that the likelihood ratio principle for testing H_0: $\theta = \theta'$, where θ' is specified, against H_1: $\theta \neq \theta'$ leads to the inequality $|\bar{x} - \theta'| \geq c$. Is this a uniformly most powerful test of H_0 against H_1?

10.7. Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution $n(\theta_1, \theta_2)$. Show that the likelihood ratio principle for testing H_0: $\theta_2 = \theta_2'$ specified, and θ_1 unspecified, against H_1: $\theta_2 > \theta_2'$, θ_1 unspecified, leads to the inequality $\sum_1^n (x_i - \bar{x})^2 \geq c$. Here $\Omega = \{(\theta_1, \theta_2); -\infty < \theta_1 < \infty, \theta_2' \leq \theta_2 < \infty\}$.

10.8. Let $Y_1, Y_2, \ldots, Y_n$ be the order statistics of a random sample of size n from a distribution which has the p.d.f. $f(x; \theta_1, \theta_2) = (1/\theta_2) \exp[-(x - \theta_1)/\theta_2]$, $\theta_1 \leq x < \infty$, zero elsewhere. Show that the likelihood ratio principle for testing H_0: $\theta_2 = \theta_2'$, where θ_2' is a specified positive number and where θ_1 is an unspecified real number, against H_1: $\theta_2 < \theta_2'$, θ_1 unspecified, leads to the inequality $\sum_1^n (y_i - y_1) \leq c$. Here $\Omega = \{(\theta_1, \theta_2); -\infty < \theta_1 < \infty, 0 < \theta_2 \leq \theta_2'\}$. If $\theta_2 = \theta_2'$, show that $2\sum_1^n (Y_i - Y_1)/\theta_2'$ is $\chi^2(2n - 2)$. *Hint.* See Exercise 7.47.

10.9. Use the background of the previous exercise and show that the likelihood ratio principle for testing H_0: $\theta_1 = \theta_1'$, where θ_1' is a specified real number and where θ_2 is an unspecified positive number, against H_1: $\theta_1 > \theta_1'$, θ_2 unspecified, leads to the inequality $\left[n(y_1 - \theta_1')/\sum_1^n (y_i - y_1)\right] \geq c$. Here $\Omega = \{(\theta_1, \theta_2); \theta_1' \leq \theta_1 < \infty, 0 < \theta_2 < \infty\}$. If $\theta_1 = \theta_1'$, show that the statistic defined by the left-hand member of the inequality has an F distribution with 2 and $2n - 2$ degrees of freedom.

10.10. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be random samples from the independent distributions $n(\theta_1, \theta_3)$ and $n(\theta_2, \theta_4)$, respectively. (a) Show that the likelihood ratio for testing H_0: $\theta_1 = \theta_2$, $\theta_3 = \theta_4$ against all alternatives is given by

$$\frac{\left[\sum_1^n (x_i - \bar{x})^2/n\right]^{n/2}\left[\sum_1^m (y_i - \bar{y})^2/m\right]^{m/2}}{\left\{\left[\sum_1^n (x_i - u)^2 + \sum_1^m (y_i - u)^2\right]/(m + n)\right\}^{(n+m)/2}},$$

where $u = (n\bar{x} + m\bar{y})/(n + m)$. (b) Show that the likelihood ratio test for testing H_0: $\theta_3 = \theta_4$, θ_1 and θ_2 unspecified, against H_1: $\theta_3 \neq \theta_4$, θ_1 and θ_2 unspecified, can be based on the random variable

$$F = \frac{\sum_1^n (X_i - \bar{X})^2/(n-1)}{\sum_1^m (Y_i - \bar{Y})^2/(m-1)}.$$

10.2 Chi-Square Tests

In this section we introduce some tests of statistical hypotheses that are called *chi-square tests*. A test of this sort was originally proposed by Karl Pearson in 1900, and it provided one of the earlier methods of statistical inference.

Let the random variable X_i be $n(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, n$, and let $X_1, X_2, \ldots, X_n$ be mutually stochastically independent. Thus the joint p.d.f. of these variables is

$$\frac{1}{\sigma_1\sigma_2\cdots\sigma_n(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_1^n \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right], \qquad -\infty < x_i < \infty.$$

The random variable which is defined by the exponent (apart from the coefficient $-\frac{1}{2}$) is $\sum_1^n (X_i - \mu_i)^2/\sigma_i^2$, and this random variable is $\chi^2(n)$. In Chapter 12 we shall generalize this joint normal distribution of probability to n random variables that are stochastically *dependent*, and we shall call the distribution a *multivariate normal distribution*. It will then be shown that a certain exponent in the joint p.d.f. (apart from a coefficient of $-\frac{1}{2}$) defines a random variable which is $\chi^2(n)$. This fact is the mathematical basis of the chi-square tests.

Let us now discuss some random variables that have approximate chi-square distributions. Let X_1 be $b(n, p_1)$. Since the random variable $Y = (X_1 - np_1)/\sqrt{np_1(1-p_1)}$ has, as $n \to \infty$, a limiting distribution which is $n(0, 1)$, we would strongly suspect that the limiting distribution of $Z = Y^2$ is $\chi^2(1)$. This is, in fact, the case, as will now be shown. If $G_n(y)$ represents the distribution function of Y, we know that

$$\lim_{n\to\infty} G_n(y) = N(y), \qquad -\infty < y < \infty,$$

where $N(y)$ is the distribution function of a distribution which is $n(0, 1)$. Let $H_n(z)$ represent, for each positive integer n, the distribution function of $Z = Y^2$. Thus, if $z \geq 0$,

$$\begin{aligned} H_n(z) &= \Pr(Z \leq z) = \Pr(-\sqrt{z} \leq Y \leq \sqrt{z}) \\ &= G_n(\sqrt{z}) - G_n[(-\sqrt{z})-]. \end{aligned}$$

Accordingly, since $N(y)$ is everywhere continuous,

$$\lim_{n\to\infty} H_n(z) = N(\sqrt{z}) - N(-\sqrt{z})$$
$$= 2\int_0^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw.$$

If we change the variable of integration in this last integral by writing $w^2 = v$, then

$$\lim_{n\to\infty} H_n(z) = \int_0^z \frac{1}{\Gamma(\frac{1}{2})2^{1/2}} v^{1/2-1} e^{-v/2}\, dv,$$

provided $z \geq 0$. If $z < 0$, then $\lim_{n\to\infty} H_n(z) = 0$. Thus $\lim_{n\to\infty} H_n(z)$ is equal to the distribution function of a random variable which is $\chi^2(1)$. This is the desired result.

Let us now return to the random variable X_1 which is $b(n, p_1)$. Let $X_2 = n - X_1$ and let $p_2 = 1 - p_1$. If we denote Y^2 by Q_1 instead of Z, we see that Q_1 may be written as

$$Q_1 = \frac{(X_1 - np_1)^2}{np_1(1-p_1)} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{n(1-p_1)}$$
$$= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2},$$

because $(X_1 - np_1)^2 = (n - X_2 - n + np_2)^2 = (X_2 - np_2)^2$. Since Q_1 has a limiting chi-square distribution with one degree of freedom, we say, when n is a positive integer, that Q_1 has an approximate chi-square distribution with one degree of freedom. This result can be generalized as follows:

Let $X_1, X_2, \ldots, X_{k-1}$ have a multinomial distribution with parameters $n, p_1, \ldots, p_{k-1}$, as in Section 3.1. As a convenience, let $X_k = n - (X_1 + \cdots + X_{k-1})$ and let $p_k = 1 - (p_1 + \cdots + p_{k-1})$. Define Q_{k-1} by

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}.$$

It is proved in a more advanced course that, as $n \to \infty$, Q_{k-1} has a limiting distribution which is $\chi^2(k-1)$. If we accept this fact, we can say that Q_{k-1} has an approximate chi-square distribution with $k-1$ degrees of freedom when n is a positive integer. Some writers caution the user of this approximation to be certain that n is large enough so that each np_i, $i = 1, 2, \ldots, k$, is at least equal to five. In any case, it is important to realize that Q_{k-1} does not have a chi-square distribution, but only an approximate chi-square distribution.

The random variable Q_{k-1} may serve as the basis of the tests of certain statistical hypotheses which we now discuss. Let the sample space $\mathcal{A}$ of a random experiment be the union of a finite number k of mutually disjoint sets $A_1, A_2, \ldots, A_k$. Furthermore, let $P(A_i) = p_i$, $i = 1, 2, \ldots, k$, where $p_k = 1 - p_1 - \cdots - p_{k-1}$, so that p_i is the probability that the outcome of the random experiment is an element of the set A_i. The random experiment is to be repeated n independent times and X_i will represent the number of times the outcome is an element of the set A_i. That is, $X_1, X_2, \ldots, X_k = n - X_1 - \cdots - X_{k-1}$ are the frequencies with which the outcome is, respectively, an element of $A_1, A_2, \ldots, A_k$. Then the joint p.d.f. of $X_1, X_2, \ldots, X_{k-1}$ is the multinomial p.d.f. with the parameters $n, p_1, \ldots, p_{k-1}$. Consider the simple hypothesis (concerning this multinomial p.d.f.) H_0: $p_1 = p_{10}$, $p_2 = p_{20}, \ldots, p_{k-1} = p_{k-1,0}$, $(p_k = p_{k0} = 1 - p_{10} - \cdots - p_{k-1,0})$, where $p_{10}, \ldots, p_{k-1,0}$ are specified numbers. It is desired to test H_0 against all alternatives.

If the hypothesis H_0 is true, the random variable

$$Q_{k-1} = \sum_1^k \left[\frac{(X_i - np_{i0})^2}{np_{i0}}\right]$$

has an approximate chi-square distribution with $k - 1$ degrees of freedom. Since, when H_0 is true, np_{i0} is the expected value of X_i, one would feel intuitively that experimental values of Q_{k-1} should not be too large if H_0 is true. With this in mind, we may use Table II of the Appendix, with $k - 1$ degrees of freedom, and find c so that $\Pr(Q_{k-1} \geq c) = \alpha$, where α is the desired significance level of the test. If, then, the hypothesis H_0 is rejected when the observed value of Q_{k-1} is at least as great as c, the test of H_0 will have a significance level that is approximately equal to α.

Some illustrative examples follow.

Example 1. One of the first six positive integers is to be chosen by a random experiment (perhaps by the cast of a die). Let $A_i = \{x; x = i\}$, $i = 1, 2, \ldots, 6$. The hypothesis H_0: $P(A_i) = p_{i0} = \frac{1}{6}$, $i = 1, 2, \ldots, 6$, will be tested, at the approximate 5 per cent significance level, against all alternatives. To make the test, the random experiment will be repeated, under the same conditions, 60 independent times. In this example $k = 6$ and $np_{i0} = 60(\frac{1}{6}) = 10$, $i = 1, 2, \ldots, 6$. Let X_i denote the frequency with which the random experiment terminates with the outcome in A_i, $i = 1, 2, \ldots, 6$, and let $Q_5 = \sum_1^6 (X_i - 10)^2/10$. If H_0 is true, Table II, with $k - 1 = 6 - 1 = 5$ degrees of freedom, shows that we have $\Pr(Q_5 \geq 11.1) = 0.05$. Now

suppose the experimental frequencies of $A_1, A_2, \ldots, A_6$ are, respectively, 13, 19, 11, 8, 5, and 4. The observed value of Q_5 is

$$\frac{(13-10)^2}{10} + \frac{(19-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(5-10)^2}{10} + \frac{(4-10)^2}{10} = 15.6.$$

Since $15.6 > 11.1$, the hypothesis $P(A_i) = \frac{1}{6}$, $i = 1, 2, \ldots, 6$, is rejected at the (approximate) 5 per cent significance level.

Example 2. A point is to be selected from the unit interval $\{x; 0 < x < 1\}$ by a random process. Let $A_1 = \{x; 0 < x \leq \frac{1}{4}\}$, $A_2 = \{x; \frac{1}{4} < x \leq \frac{1}{2}\}$, $A_3 = \{x; \frac{1}{2} < x \leq \frac{3}{4}\}$, and $A_4 = \{x; \frac{3}{4} < x < 1\}$. Let the probabilities p_i, $i = 1, 2, 3, 4$, assigned to these sets under the hypothesis be determined by the p.d.f. $2x$, $0 < x < 1$, zero elsewhere. Then these probabilities are, respectively,

$$p_{10} = \int_0^{1/4} 2x\,dx = \frac{1}{16}, \quad p_{20} = \frac{3}{16}, \quad p_{30} = \frac{5}{16}, \quad \text{and} \quad p_{40} = \frac{7}{16}.$$

Thus the hypothesis to be tested is that p_1, p_2, p_3, and $p_4 = 1 - p_1 - p_2 - p_3$ have the preceding values in a multinomial distribution with $k = 4$. This hypothesis is to be tested at an approximate 0.025 significance level by repeating the random experiment $n = 80$ independent times under the same conditions. Here the np_{i0}, $i = 1, 2, 3, 4$, are, respectively, 5, 15, 25, and 35. Suppose the observed frequencies of A_1, A_2, A_3, and A_4 to be 6, 18, 20, and 36, respectively. Then the observed value of $Q_3 = \sum_1^4 (X_i - np_{i0})^2/(np_{i0})$ is

$$\frac{(6-5)^2}{5} + \frac{(18-15)^2}{15} + \frac{(20-25)^2}{25} + \frac{(36-35)^2}{35} = \frac{64}{35} = 1.83,$$

approximately. From Table II, with $4 - 1 = 3$ degrees of freedom, the value corresponding to a 0.025 significance level is $c = 9.35$. Since the observed value of Q_3 is less than 9.35, the hypothesis is accepted at the (approximate) 0.025 level of significance.

Thus far we have used the chi-square test when the hypothesis H_0 is a simple hypothesis. More often we encounter hypotheses H_0 in which the multinomial probabilities $p_1, p_2, \ldots, p_k$ are not completely specified by the hypothesis H_0. That is, under H_0, these probabilities are functions of unknown parameters. For illustration, suppose a certain random variable Y can take on any real value. Let us partition the space $\{y; -\infty < y < \infty\}$ into k mutually disjoint sets $A_1, A_2, \ldots, A_k$ so that the events $A_1, A_2, \ldots, A_k$ are mutually exclusive and exhaustive.

Let H_0 be the hypothesis that Y is $n(\mu, \sigma^2)$ with μ and σ^2 unspecified. Then each

$$p_i = \int_{A_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-(y - \mu)^2/2\sigma^2\right] dy, \qquad i = 1, 2, \ldots, k,$$

is a function of the unknown parameters μ and σ^2. Suppose we take a random sample $Y_1, \ldots, Y_n$ of size n from this distribution. If we let X_i denote the frequency of A_i, $i = 1, 2, \ldots, k$, so that $X_1 + \cdots + X_k = n$, the random variable

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

cannot be computed once $X_1, \ldots, X_k$ have been observed, since each p_i, and hence Q_{k-1}, is a function of the unknown parameters μ and σ^2.

There is a way out of our trouble, however. We can accept, without proof at this time, the following fact. We are entitled to use the observed values of the items of the random sample $Y_1, Y_2, \ldots, Y_n$ to obtain point estimates of μ and σ^2. These point estimates of μ and σ^2 enable us to compute numerically the estimates of each p_i. Accordingly, if these values are used, Q_{k-1} can be computed once $Y_1, Y_2, \ldots, Y_n$, and hence $X_1, X_2, \ldots, X_k$, are observed. However, a very important aspect of the fact which we accept without proof is that now Q_{k-1} is approximately $\chi^2(k - 3)$. That is, the number of degrees of freedom of the limiting chi-square distribution of Q_{k-1} is reduced by one for each parameter estimated by the experimental data. This statement applies not only to the problem at hand but also to more general situations. Two examples will now be given. The first of these examples will deal with the test of the hypothesis that two multinomial distributions are the same.

Example 3. Let us consider two independent multinomial distributions with parameters $n_j, p_{1j}, p_{2j}, \ldots, p_{kj}$, $j = 1, 2$, respectively. Let X_{ij}, $i = 1, 2, \ldots, k$, $j = 1, 2$, represent the corresponding frequencies. If n_1 and n_2 are large, the random variable

$$\sum_{j=1}^{2} \sum_{i=1}^{k} \frac{(X_{ij} - n_j p_{ij})^2}{n_j p_{ij}}$$

is the sum of two stochastically independent random variables, each of which we treat as though it were $\chi^2(k - 1)$; that is, the random variable is approximately $\chi^2[2k - 2]$. Consider the hypothesis

$$H_0: p_{11} = p_{12}, p_{21} = p_{22}, \ldots, p_{k1} = p_{k2},$$

where each $p_{i1} = p_{i2}$, $i = 1, 2, \ldots, k$, is unspecified. Thus we need point estimates of these parameters. The maximum likelihood statistic for $p_{i1} = p_{i2}$ is $(X_{i1} + X_{i2})/(n_1 + n_2)$, $i = 1, 2, \ldots, k$. Note that we need only $k - 1$ point estimates because we have a point estimate of $p_{k1} = p_{k2}$ once we have point estimates of the first $k - 1$ probabilities. In accordance with the fact which has been stated, the random variable

$$\sum_{j=1}^{2} \sum_{i=1}^{k} \frac{\{X_{ij} - n_j[(X_{i1} + X_{i2})/(n_1 + n_2)]\}^2}{n_j[(X_{i1} + X_{i2})/(n_1 + n_2)]}$$

has an approximate χ^2 distribution with $2k - 2 - (k - 1) = k - 1$ degrees of freedom. Thus we are able to test the hypothesis that two multinomial distributions are the same; this hypothesis is rejected when the computed value of this random variable is at least as great as an appropriate number from Table II of the Appendix, with $k - 1$ degrees of freedom.

The second example deals with the subject of *contingency tables*.

Example 4. Let the result of a random experiment be classified by two attributes (such as the color of the hair and the color of the eyes). That is, one attribute of the outcome is one and only one of certain mutually exclusive and exhaustive events, say $A_1, A_2, \ldots, A_a$; and the other attribute of the outcome is also one and only one of certain mutually exclusive and exhaustive events, say $B_1, B_2, \ldots, B_b$. Let $p_{ij} = P(A_i \cap B_j)$, $i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$. The random experiment is to be repeated n independent times and X_{ij} will denote the frequency of the event $A_i \cap B_j$. Since there are $k = ab$ such events as $A_i \cap B_j$, the random variable

$$Q_{ab-1} = \sum_{j=1}^{b} \sum_{i=1}^{a} \frac{(X_{ij} - np_{ij})^2}{np_{ij}}$$

has an approximate chi-square distribution with $ab - 1$ degrees of freedom, provided n is large. Suppose that we wish to test the independence of the A attribute and the B attribute; that is, we wish to test the hypothesis H_0: $P(A_i \cap B_j) = P(A_i)P(B_j)$, $i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$. Let us denote $P(A_i)$ by $p_{i\cdot}$ and $P(B_j)$ by $p_{\cdot j}$; thus

$$p_{i\cdot} = \sum_{j=1}^{b} p_{ij}, \qquad p_{\cdot j} = \sum_{i=1}^{a} p_{ij},$$

and

$$1 = \sum_{j=1}^{b} \sum_{i=1}^{a} p_{ij} = \sum_{j=1}^{b} p_{\cdot j} = \sum_{i=1}^{a} p_{i\cdot}.$$

Then the hypothesis can be formulated as H_0: $p_{ij} = p_{i\cdot}p_{\cdot j}$, $i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$. To test H_0, we can use Q_{ab-1} with p_{ij} replaced by $p_{i\cdot}p_{\cdot j}$. But if $p_{i\cdot}$, $i = 1, 2, \ldots, a$, and $p_{\cdot j}$, $j = 1, 2, \ldots, b$, are unknown, as they frequently are in the applications, we cannot compute Q_{ab-1} once the

frequencies are observed. In such a case we estimate these unknown parameters by

$$\hat{p}_{i\cdot} = \frac{X_{i\cdot}}{n}, \qquad \text{where} \qquad X_{i\cdot} = \sum_{j=1}^{b} X_{ij}, \qquad i = 1, 2, \ldots, a,$$

and

$$\hat{p}_{\cdot j} = \frac{X_{\cdot j}}{n}, \qquad \text{where} \qquad X_{\cdot j} = \sum_{i=1}^{a} X_{ij}, \qquad j = 1, 2, \ldots, b.$$

Since $\sum_i \hat{p}_{i\cdot} = \sum_j \hat{p}_{\cdot j} = 1$, we have estimated only $a - 1 + b - 1 = a + b - 2$ parameters. So if these estimates are used in Q_{ab-1}, with $\hat{p}_{ij} = \hat{p}_{i\cdot}\hat{p}_{\cdot j}$, then, according to the rule which has been stated in this section, the random variable

$$\sum_{j=1}^{b} \sum_{i=1}^{a} \frac{[X_{ij} - n(X_{i\cdot}/n)(X_{\cdot j}/n)]^2}{n(X_{i\cdot}/n)(X_{\cdot j}/n)}$$

has an approximate chi-square distribution with $ab - 1 - (a + b - 2) = (a - 1)(b - 1)$ degrees of freedom, provided H_0 is true. The hypothesis H_0 is then rejected if the computed value of this statistic exceeds the constant c, where c is selected from Table II so that the test has the desired significance level α.

In each of the four examples of this section, we have indicated that the statistic used to test the hypothesis H_0 has an approximate chi-square distribution provided n is sufficiently large and H_0 is true. To compute the power of any of these tests for values of the parameters not described by H_0, we need the distribution of the statistic when H_0 is not true. In each of these cases, the statistic has an approximate distribution which is called a *noncentral chi-square distribution*. The noncentral chi-square distribution will be discussed in Section 10.5.

EXERCISES

10.11. A number is to be selected from the interval $\{x; 0 < x < 2\}$ by a random process. Let $A_i = \{x; (i - 1)/2 < x \leq i/2\}$, $i = 1, 2, 3$, and let $A_4 = \{x; \frac{3}{2} < x < 2\}$. A certain hypothesis assigns probabilities p_{i0} to these sets in accordance with $p_{i0} = \int_{A_i} (\frac{1}{2})(2 - x)\,dx$, $i = 1, 2, 3, 4$. This hypothesis (concerning the multinomial p.d.f. with $k = 4$) is to be tested, at the 5 per cent level of significance, by a chi-square test. If the observed frequencies of the sets A_i, $i = 1, 2, 3, 4$, are, respectively, 30, 30, 10, 10, would H_0 be accepted at the (approximate) 5 per cent level of significance?

10.12. Let the following sets be defined. $A_1 = \{x; -\infty < x \leq 0\}$, $A_i = \{x; i - 2 < x \leq i - 1\}$, $i = 2, \ldots, 7$, and $A_8 = \{x; 6 < x < \infty\}$. A certain hypothesis assigns probabilities p_{i0} to these sets A_i in accordance with

$$p_{i0} = \int_{A_i} \frac{1}{2\sqrt{2\pi}} \exp\left[-\frac{(x - 3)^2}{2(4)}\right] dx, \qquad i = 1, 2, \ldots, 7, 8.$$

This hypothesis (concerning the multinomial p.d.f. with $k = 8$) is to be tested, at the 5 per cent level of significance, by a chi-square test. If the observed frequencies of the sets A_i, $i = 1, 2, \ldots, 8$, are, respectively, 60, 96, 140, 210, 172, 160, 88, and 74, would H_0 be accepted at the (approximate) 5 per cent level of significance?

10.13. Consider the problem from genetics of crossing two types of peas. The Mendelian theory states that the probabilities of the classifications (a) round and yellow, (b) wrinkled and yellow, (c) round and green, and (d) wrinkled and green are $\frac{9}{16}$, $\frac{3}{16}$, $\frac{3}{16}$, and $\frac{1}{16}$, respectively. If, from 160 independent observations, the observed frequencies of these respective classifications are 86, 35, 26, and 13, are these data consistent with the Mendelian theory? That is, test, with $\alpha = 0.01$, the hypothesis that the respective probabilities are $\frac{9}{16}$, $\frac{3}{16}$, $\frac{3}{16}$, and $\frac{1}{16}$.

10.14. Two different teaching procedures were used on two different groups of students. Each group contained 100 students of about the same ability. At the end of the term, an evaluating team assigned a letter grade to each student. The results were tabulated as follows.

Grade	A	B	C	D	F	Total
Group I	15	25	32	17	11	100
Group II	9	18	29	28	16	100

If we consider these data to be observations from two independent multinomial distributions with $k = 5$, test, at the 5 per cent significance level, the hypothesis that the two distributions are the same (and hence the two teaching procedures are equally effective).

10.15. Families were selected at random from a certain population in which the parents had been married for at least 10 years but less than 15 years. The data were tabled by two attributes in the following manner.

Family Income in Dollars	Number of Children			
	0	1	2	More Than 2
Under 6000	11	24	49	44
6000–12,000	9	28	39	31
Over 12,000	6	15	16	15

Test, at the 0.01 significance level, the hypothesis that the attributes—family income and number of children—are independent in this population.

10.16. A certain genetic model suggests that the probabilities of a particular trinomial distribution are, respectively, $p_1 = p^2$, $p_2 = 2p(1 - p)$, and $p_3 = (1 - p)^2$, where $0 < p < 1$. If X_1, X_2, X_3 represent the respective

frequencies in n independent trials, explain how we could check on the adequacy of the genetic model.

10.3 The Distributions of Certain Quadratic Forms

A homogeneous polynomial of degree two in n variables is called a *quadratic* form in those variables. If both the variables and the coefficients are real, the form is called a *real quadratic* form. Only real quadratic forms will be considered in this book. To illustrate, the form $X_1^2 + X_1X_2 + X_2^2$ is a quadratic form in the two variables X_1 and X_2; the form $X_1^2 + X_2^2 + X_3^2 - 2X_1X_2$ is a quadratic form in the three variables X_1, X_2, and X_3; but the form $(X_1 - 1)^2 + (X_2 - 2)^2 = X_1^2 + X_2^2 - 2X_1 - 4X_2 + 5$ is not a quadratic form in X_1 and X_2, although it is a quadratic form in the variables $X_1 - 1$ and $X_2 - 2$.

Let $\bar{X}$ and S^2 denote, respectively, the mean and the variance of a random sample $X_1, X_2, \ldots, X_n$ from an arbitrary distribution. Thus

$$nS^2 = \sum_1^n (X_i - \bar{X})^2 = \sum_1^n \left[X_i - \frac{(X_1 + X_2 + \cdots + X_n)}{n}\right]^2$$

$$= \frac{n-1}{n}(X_1^2 + X_2^2 + \cdots + X_n^2) - \frac{2}{n}(X_1X_2 + \cdots + X_1X_n + \cdots + X_{n-1}X_n)$$

is a quadratic form in the n variables $X_1, X_2, \ldots, X_n$. If the sample arises from a distribution which is $n(\mu, \sigma^2)$, we know that the random variable nS^2/σ^2 is $\chi^2(n-1)$ regardless of the value of μ. This fact proved useful in our search for a confidence interval for σ^2 when μ is unknown.

It has been seen that tests of certain statistical hypotheses require a statistic which is a quadratic form. For instance, Example 2, p. 281, made use of the statistic $\sum_1^n X_i^2$, which is a quadratic form in the variables $X_1, X_2, \ldots, X_n$. Later in this chapter, tests of other statistical hypotheses will be investigated, and it will be seen that functions of statistics that are quadratic forms will be needed to carry out the tests in an expeditious manner. But first we shall make a study of the distributions of certain quadratic forms in normal and stochastically independent random variables.

The following theorem will be proved in Chapter 12.

Theorem 1. *Let* $Q = Q_1 + Q_2 + \cdots + Q_{k-1} + Q_k$, *where* $Q, Q_1, \ldots, Q_k$ *are* $k + 1$ *random variables that are real quadratic forms in* n *mutually stochastically independent random variables which are normally*

distributed with the means $\mu_1, \mu_2, \ldots, \mu_n$ *and the same variance* σ^2. *Let* $Q/\sigma^2, Q_1/\sigma^2, \ldots, Q_{k-1}/\sigma^2$ *have chi-square distributions with degrees of freedom* $r, r_1, \ldots, r_{k-1}$, *respectively. Let* Q_k *be nonnegative. Then:*

(a) $Q_1, \ldots, Q_k$ *are mutually stochastically independent, and hence*

(b) Q_k/σ^2 *has a chi-square distribution with* $r - (r_1 + \cdots + r_{k-1}) = r_k$ *degrees of freedom.*

Three examples illustrative of the theorem will follow. Each of these examples will deal with a distribution problem which is based on the remarks made in the subsequent paragraph.

Let the random variable X have a distribution which is $n(\mu, \sigma^2)$. Let a and b denote positive integers greater than one and let $n = ab$. Consider a random sample of size $n = ab$ from this normal distribution. The items of the random sample will be denoted by the symbols

$$\begin{array}{ccccccc} X_{11}, & X_{12}, & \ldots, & X_{1j}, & \ldots, & X_{1b} \\ X_{21}, & X_{22}, & \ldots, & X_{2j}, & \ldots, & X_{2b} \\ \vdots & & & & & \\ X_{i1}, & X_{i2}, & \ldots, & X_{ij}, & \ldots, & X_{ib} \\ \vdots & & & & & \\ X_{a1}, & X_{a2}, & \ldots, & X_{aj}, & \ldots, & X_{ab}. \end{array}$$

In this notation, the first subscript indicates the row, and the second subscript indicates the column in which the item appears. Thus, X_{ij} is in row i and column j, $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, b$. By assumption these $n = ab$ random variables are mutually stochastically independent, and each has the same normal distribution with mean μ and variance σ^2. Thus, if we wish, we may consider each row as being a random sample of size b from the given distribution; and we may consider each column as being a random sample of size a from the given distribution. We now define $a + b + 1$ statistics. They are

$$\bar{X} = \frac{X_{11} + \cdots + X_{1b} + \cdots + X_{a1} + \cdots + X_{ab}}{ab} = \frac{\sum_{i=1}^{a} \sum_{j=1}^{b} X_{ij}}{ab},$$

$$\bar{X}_{i\cdot} = \frac{X_{i1} + X_{i2} + \cdots + X_{ib}}{b} = \frac{\sum_{j=1}^{b} X_{ij}}{b}, \qquad i = 1, 2, \ldots, a,$$

and

$$\bar{X}_{\cdot j} = \frac{X_{1j} + X_{2j} + \cdots + X_{aj}}{a} = \frac{\sum_{i=1}^{a} X_{ij}}{a}, \qquad j = 1, 2, \ldots, b.$$

Thus the statistic $\bar{X}$ is the mean of the random sample of size $n = ab$; the statistics $\bar{X}_{1\cdot}, \bar{X}_{2\cdot}, \ldots, \bar{X}_{a\cdot}$ are, respectively, the means of the rows; and the statistics $\bar{X}_{\cdot 1}, \bar{X}_{\cdot 2}, \ldots, \bar{X}_{\cdot b}$ are, respectively, the means of the columns. The examples illustrative of the theorem follow.

Example 1. Consider the variance S^2 of the random sample of size $n = ab$. We have the algebraic identity

$$\begin{aligned} abS^2 &= \sum_{i=1}^{a} \sum_{j=1}^{b} (X_{ij} - \bar{X})^2 \\ &= \sum_{i=1}^{a} \sum_{j=1}^{b} [(X_{ij} - \bar{X}_{i\cdot}) + (\bar{X}_{i\cdot} - \bar{X})]^2 \\ &= \sum_{i=1}^{a} \sum_{j=1}^{b} (X_{ij} - \bar{X}_{i\cdot})^2 + \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{X}_{i\cdot} - \bar{X})^2 \\ &\quad + 2 \sum_{i=1}^{a} \sum_{j=1}^{b} (X_{ij} - \bar{X}_{i\cdot})(\bar{X}_{i\cdot} - \bar{X}). \end{aligned}$$

The last term of the right-hand member of this identity may be written

$$2 \sum_{i=1}^{a} \left[(\bar{X}_{i\cdot} - \bar{X}) \sum_{j=1}^{b} (X_{ij} - \bar{X}_{i\cdot}) \right] = 2 \sum_{i=1}^{a} [(\bar{X}_{i\cdot} - \bar{X})(b\bar{X}_{i\cdot} - b\bar{X}_{i\cdot})] = 0,$$

and the term

$$\sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{X}_{i\cdot} - \bar{X})^2$$

may be written

$$b \sum_{i=1}^{a} (\bar{X}_{i\cdot} - \bar{X})^2.$$

Thus,

$$abS^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} (X_{ij} - \bar{X}_{i\cdot})^2 + b \sum_{i=1}^{a} (\bar{X}_{i\cdot} - \bar{X})^2,$$

or, for brevity,

$$Q = Q_1 + Q_2.$$

Clearly, Q, Q_1, and Q_2 are quadratic forms in the $n = ab$ variables X_{ij}. We shall use the theorem with $k = 2$ to show that Q_1 and Q_2 are stochastically independent. Since S^2 is the variance of a random sample of size $n = ab$ from the given normal distribution, then abS^2/σ^2 has a chi-square distribution with $ab - 1$ degrees of freedom. Now

$$\frac{Q_1}{\sigma^2} = \sum_{i=1}^{a} \left[\frac{\sum_{j=1}^{b} (X_{ij} - \bar{X}_{i\cdot})^2}{\sigma^2} \right].$$

For each fixed value of i, $\sum_{j=1}^{b} (X_{ij} - \bar{X}_{i\cdot})^2/b$ is the variance of a random sample of size b from the given normal distribution, and, accordingly, $\sum_{j=1}^{b} (X_{ij} - \bar{X}_{i\cdot})^2/\sigma^2$ has a chi-square distribution with $b - 1$ degrees of

freedom. Because the X_{ij} are mutually stochastically independent, Q_1/σ^2 is the sum of a mutually stochastically independent random variables, each having a chi-square distribution with $b - 1$ degrees of freedom. Hence Q_1/σ^2 has a chi-square distribution with $a(b - 1)$ degrees of freedom. Now $Q_2 = b \sum_{i=1}^{a} (\bar{X}_{i\cdot} - \bar{X})^2 \geq 0$. In accordance with the theorem, Q_1 and Q_2 are stochastically independent, and Q_2/σ^2 has a chi-square distribution with $ab - 1 - a(b - 1) = a - 1$ degrees of freedom.

Example 2. In abS^2, replace $X_{ij} - \bar{X}$ by $(X_{ij} - \bar{X}_{\cdot j}) + (\bar{X}_{\cdot j} - \bar{X})$ to obtain

$$abS^2 = \sum_{j=1}^{b} \sum_{i=1}^{a} [(X_{ij} - \bar{X}_{\cdot j}) + (\bar{X}_{\cdot j} - \bar{X})]^2,$$

or

$$abS^2 = \sum_{j=1}^{b} \sum_{i=1}^{a} (X_{ij} - \bar{X}_{\cdot j})^2 + a \sum_{j=1}^{b} (\bar{X}_{\cdot j} - \bar{X})^2,$$

or, for brevity,

$$Q = Q_3 + Q_4.$$

It is easy to show (Exercise 10.17) that Q_3/σ^2 has a chi-square distribution with $b(a - 1)$ degrees of freedom. Since $Q_4 = a \sum_{j=1}^{b} (\bar{X}_{\cdot j} - \bar{X})^2 \geq 0$, the theorem enables us to assert that Q_3 and Q_4 are stochastically independent and that Q_4/σ^2 has a chi-square distribution with $ab - 1 - b(a - 1) = b - 1$ degrees of freedom.

Example 3. In abS^2, replace $X_{ij} - \bar{X}$ by $(\bar{X}_{i\cdot} - \bar{X}) + (\bar{X}_{\cdot j} - \bar{X}) + (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})$ to obtain (Exercise 10.18)

$$abS^2 = b \sum_{i=1}^{a} (\bar{X}_{i\cdot} - \bar{X})^2 + a \sum_{j=1}^{b} (\bar{X}_{\cdot j} - \bar{X})^2 + \sum_{j=1}^{b} \sum_{i=1}^{a} (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2,$$

or, for brevity,

$$Q = Q_2 + Q_4 + Q_5,$$

where Q_2 and Q_4 are as defined in Examples 1 and 2. From Examples 1 and 2, Q/σ^2, Q_2/σ^2, and Q_4/σ^2 have chi-square distributions with $ab - 1$, $a - 1$, and $b - 1$ degrees of freedom, respectively. Since $Q_5 \geq 0$, the theorem asserts that Q_2, Q_4, and Q_5 are mutually stochastically independent and that Q_5/σ^2 has a chi-square distribution with $ab - 1 - (a - 1) - (b - 1) = (a - 1)(b - 1)$ degrees of freedom.

Once these quadratic form statistics have been shown to be stochastically independent, a multiplicity of F statistics can be defined. For instance,

$$\frac{Q_4/[\sigma^2(b - 1)]}{Q_3/[\sigma^2 b(a - 1)]} = \frac{Q_4/(b - 1)}{Q_3/[b(a - 1)]}$$

has an F distribution with $b - 1$ and $b(a - 1)$ degrees of freedom; and

$$\frac{Q_4/[\sigma^2(b-1)]}{Q_5/[\sigma^2(a-1)(b-1)]} = \frac{Q_4}{Q_5/(a-1)}$$

has an F distribution with $b - 1$ and $(a - 1)(b - 1)$ degrees of freedom. In the subsequent sections it will be seen that some likelihood ratio tests of certain statistical hypotheses can be based on these F statistics.

EXERCISES

10.17. In Example 2, verify that $Q = Q_3 + Q_4$ and that Q_3/σ^2 has a chi-square distribution with $b(a - 1)$ degrees of freedom.

10.18. In Example 3, verify that $Q = Q_2 + Q_4 + Q_5$.

10.19. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $n(\mu, \sigma^2)$. Show that

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=2}^{n} (X_i - \bar{X}')^2 + \frac{n-1}{n}(X_1 - \bar{X}')^2,$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n$ and $\bar{X}' = \sum_{i=2}^{n} X_i/(n-1)$. *Hint.* Replace $X_i - \bar{X}$ by $(X_i - \bar{X}') - (X_1 - \bar{X}')/n$. Show that $\sum_{i=2}^{n} (X_i - \bar{X}')^2/\sigma^2$ has a chi-square distribution with $n - 2$ degrees of freedom. Prove that the two terms in the right-hand member are stochastically independent. What then is the distribution of

$$\frac{[(n-1)/n](X_1 - \bar{X}')^2}{\sigma^2}?$$

10.20. Let X_{ijk}, $i = 1, \ldots, a$; $j = 1, \ldots, b$; $k = 1, \ldots, c$, be a random sample of size $n = abc$ from a normal distribution $n(\mu, \sigma^2)$. Let $\bar{X} = \sum_{k=1}^{c}\sum_{j=1}^{b}\sum_{i=1}^{a} X_{ijk}/n$ and $\bar{X}_{i\cdot\cdot} = \sum_{k=1}^{c}\sum_{j=1}^{b} X_{ijk}/(bc)$. Show that

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \bar{X})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \bar{X}_{i\cdot\cdot})^2 + bc\sum_{i=1}^{a}(\bar{X}_{i\cdot\cdot} - \bar{X})^2.$$

Show that $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \bar{X}_{i\cdot\cdot})^2/\sigma^2$ has a chi-square distribution with $a(bc - 1)$ degrees of freedom. Prove that the two terms in the right-hand member are stochastically independent. What then is the distribution of $bc\sum_{i=1}^{a}(\bar{X}_{i\cdot\cdot} - \bar{X})^2/\sigma^2$? Furthermore, let $\bar{X}_{\cdot j\cdot} = \sum_{k=1}^{c}\sum_{i=1}^{a} X_{ijk}/(ac)$ and $\bar{X}_{ij\cdot} = \sum_{k=1}^{c} X_{ijk}/c$. Show that

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \bar{X})^2$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \bar{X}_{ij\cdot})^2 + bc\sum_{i=1}^{a}(\bar{X}_{i\cdot\cdot} - \bar{X})^2 + ac\sum_{j=1}^{b}(\bar{X}_{\cdot j\cdot} - \bar{X})^2$$

$$+ c\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X})^2.$$

Show that the four terms in the right-hand member, when divided by σ^2, are mutually stochastically independent chi-square variables with $ab(c-1)$, $a-1$, $b-1$, and $(a-1)(b-1)$ degrees of freedom, respectively.

10.4 A Test of the Equality of Several Means

Consider b mutually stochastically independent random variables that have normal distributions with unknown means $\mu_1, \mu_2, \ldots, \mu_b$, respectively, and unknown but common variance σ^2. Let $X_{1j}, X_{2j}, \ldots, X_{aj}$ represent a random sample of size a from the normal distribution with mean μ_j and variance σ^2, $j = 1, 2, \ldots, b$. It is desired to test the composite hypothesis $H_0: \mu_1 = \mu_2 = \cdots = \mu_b = \mu$, μ unspecified, against all possible alternative hypotheses H_1. A likelihood ratio test will be used. Here the total parameter space is

$$\Omega = \{(\mu_1, \mu_2, \ldots, \mu_b, \sigma^2);\ -\infty < \mu_j < \infty, 0 < \sigma^2 < \infty\}$$

and

$$\omega = \{(\mu_1, \mu_2, \ldots, \mu_b, \sigma^2);\ -\infty < \mu_1 = \mu_2 = \cdots = \mu_b = \mu < \infty, 0 < \sigma^2 < \infty\}.$$

The likelihood functions, denoted by $L(\omega)$ and $L(\Omega)$ are, respectively,

$$L(\omega) = \left(\frac{1}{2\pi\sigma^2}\right)^{ab/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{b}\sum_{i=1}^{a}(x_{ij}-\mu)^2\right]$$

and

$$L(\Omega) = \left(\frac{1}{2\pi\sigma^2}\right)^{ab/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{b}\sum_{i=1}^{a}(x_{ij}-\mu_j)^2\right].$$

Now

$$\frac{\partial \ln L(\omega)}{\partial \mu} = \frac{\sum_{j=1}^{b}\sum_{i=1}^{a}(x_{ij}-\mu)}{\sigma^2}$$

and

$$\frac{\partial \ln L(\omega)}{\partial(\sigma^2)} = -\frac{ab}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{j=1}^{b}\sum_{i=1}^{a}(x_{ij}-\mu)^2.$$

If we equate these partial derivatives to zero, the solutions for μ and σ^2 are, respectively, in ω,

$$(1) \qquad \frac{\sum_{j=1}^{b}\sum_{i=1}^{a}x_{ij}}{ab} = \bar{x},$$

$$\frac{\sum_{j=1}^{b}\sum_{i=1}^{a}(x_{ij}-\bar{x})^2}{ab} = v,$$

and these numbers maximize $L(\omega)$. Furthermore,

$$\frac{\partial \ln L(\Omega)}{\partial \mu_j} = \frac{\sum_{i=1}^{a} (x_{ij} - \mu_j)}{\sigma^2}, \qquad j = 1, 2, \ldots, b,$$

and

$$\frac{\partial \ln L(\Omega)}{\partial (\sigma^2)} = -\frac{ab}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \mu_j)^2.$$

If we equate these partial derivatives to zero, the solutions for $\mu_1, \mu_2, \ldots, \mu_b$, and σ^2 are, respectively, in Ω,

$$(2) \qquad \begin{aligned} \frac{\sum_{i=1}^{a} x_{ij}}{a} &= \bar{x}_{.j}, \qquad j = 1, 2, \ldots, b, \\ \frac{\sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \bar{x}_{.j})^2}{ab} &= w, \end{aligned}$$

and these numbers maximize $L(\Omega)$. These maxima are, respectively,

$$\begin{aligned} L(\hat{\omega}) &= \left[\frac{ab}{2\pi \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \bar{x})^2} \right]^{ab/2} \exp \left[-\frac{ab \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \bar{x})^2}{2 \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \bar{x})^2} \right] \\ &= \left[\frac{ab}{2\pi \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \bar{x})^2} \right]^{ab/2} e^{-ab/2} \end{aligned}$$

and

$$L(\hat{\Omega}) = \left[\frac{ab}{2\pi \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \bar{x}_{.j})^2} \right]^{ab/2} e^{-ab/2}.$$

Finally,

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \left[\frac{\sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \bar{x}_{.j})^2}{\sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \bar{x})^2} \right]^{ab/2}.$$

In the notation of Section 10.3, the statistics defined by the functions $\bar{x}$ and v given by Equations (1) of this section are

$$\bar{X} = \sum_{j=1}^{b} \sum_{i=1}^{a} X_{ij}/ab \quad \text{and} \quad S^2 = \sum_{j=1}^{b} \sum_{i=1}^{a} (X_{ij} - \bar{X})^2/ab = Q/ab;$$

while the statistics defined by the functions $\bar{x}_{\cdot 1}, \bar{x}_{\cdot 2}, \ldots, \bar{x}_{\cdot b}$ and w given by Equations (2) in this section are, respectively, $\bar{X}_{\cdot j} = \sum_{i=1}^{a} X_{ij}/a$, $j = 1, 2, \ldots, b$, and $Q_3/ab = \sum_{j=1}^{b} \sum_{i=1}^{a} (X_{ij} - \bar{X}_{\cdot j})^2/ab$. Thus, in the notation of Section 10.3, $\lambda^{2/ab}$ defines the statistic Q_3/Q.

We reject the hypothesis H_0 if $\lambda \leq \lambda_0$. To find λ_0 so that we have a desired significance level α, we must assume that the hypothesis H_0 is true. If the hypothesis H_0 is true, the random variables X_{ij} constitute a random sample of size $n = ab$ from a distribution which is normal with mean μ and variance σ^2. This being the case, it was shown in Example 2, p. 319, that $Q = Q_3 + Q_4$, where $Q_4 = a \sum_{j=1}^{b} (\bar{X}_{\cdot j} - \bar{X})^2$; that Q_3 and Q_4 are stochastically independent; and that Q_3/σ^2 and Q_4/σ^2 have chi-square distributions with $b(a - 1)$ and $b - 1$ degrees of freedom, respectively. Thus the statistic defined by $\lambda^{2/ab}$ may be written

$$\frac{Q_3}{Q_3 + Q_4} = \frac{1}{1 + Q_4/Q_3}.$$

The significance level of the test of H_0 is

$$\begin{aligned} \alpha &= \Pr\left[\frac{1}{1 + Q_4/Q_3} \leq \lambda_0^{2/ab}; H_0\right] \\ &= \Pr\left[\frac{Q_4/(b - 1)}{Q_3/[b(a - 1)]} \geq c; H_0\right], \end{aligned}$$

where

$$c = \frac{b(a - 1)}{b - 1}(\lambda_0^{-2/ab} - 1).$$

But

$$F = \frac{Q_4/[\sigma^2(b - 1)]}{Q_3/[\sigma^2 b(a - 1)]} = \frac{Q_4/(b - 1)}{Q_3/[b(a - 1)]}$$

has an F distribution with $b - 1$ and $b(a - 1)$ degrees of freedom. Hence the test of the composite hypothesis $H_0: \mu_1 = \mu_2 = \cdots = \mu_b = \mu$, μ unspecified, against all possible alternatives may be based on an F statistic. The constant c is so selected as to yield the desired value of α.

Remark. It should be pointed out that a test of the equality of the b means μ_j, $j = 1, 2, \ldots, b$, does not require that we take a random sample of

size a from each of the b normal distributions. That is, the samples may be of different sizes, say $a_1, a_2, \ldots, a_b$. A consideration of this procedure is left to Exercise 10.22.

Suppose now we wish to compute the power of the test of H_0 against H_1 when H_0 is false, that is, when we do not have $\mu_1 = \mu_2 = \cdots = \mu_b = \mu$. It will be seen in Section 10.5 that, when H_1 is true, no longer is Q_4/σ^2 a random variable which is $\chi^2(b-1)$. Thus we cannot use an F statistic to compute the power of the test when H_1 is true. This problem is discussed in Section 10.5.

An observation should be made in connection with maximizing a likelihood function with respect to certain parameters. Sometimes it is easier to avoid the use of the calculus. For example, $L(\Omega)$ of this section can be maximized with respect to μ_j, for every fixed positive σ^2, by minimizing

$$z = \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \mu_j)^2$$

with respect to μ_j, $j = 1, 2, \ldots, b$. Now z can be written as

$$\begin{aligned} z &= \sum_{j=1}^{b} \sum_{i=1}^{a} [(x_{ij} - \bar{x}_{\cdot j}) + (\bar{x}_{\cdot j} - \mu_j)]^2 \\ &= \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \bar{x}_{\cdot j})^2 + a \sum_{j=1}^{b} (\bar{x}_{\cdot j} - \mu_j)^2. \end{aligned}$$

Since each term in the right-hand member of the preceding equation is nonnegative, clearly z is a minimum, with respect to μ_j, if we take $\mu_j = \bar{x}_{\cdot j}$, $j = 1, 2, \ldots, b$.

EXERCISES

10.21. Let μ_1, μ_2, μ_3 be, respectively, the means of three independent normal distributions having common but unknown variance σ^2. In order to test, at the 5 per cent significance level, the hypothesis H_0: $\mu_1 = \mu_2 = \mu_3$ against all possible alternative hypotheses, we take a random sample of size 5 from each of these distributions. By using the F statistic considered in this section, determine whether we accept or reject H_0 if the observed values from the three distributions are respectively

3	2	4
0	5	3
−1	1	6
0	3	8
2	5	5.

10.22. Let $X_{1j}, X_{2j}, \ldots, X_{a_j j}$ represent independent random samples of sizes a_j from normal distributions with means μ_j and variances σ^2, $j = 1, 2, \ldots, b$. Show that

$$\sum_{j=1}^{b} \sum_{i=1}^{a_j} (X_{ij} - \bar{X})^2 = \sum_{j=1}^{b} \sum_{i=1}^{a_j} (X_{ij} - \bar{X}_{\cdot j})^2 + \sum_{j=1}^{b} a_j(\bar{X}_{\cdot j} - \bar{X})^2,$$

or $Q' = Q_3' + Q_4'$. Here $\bar{X} = \sum_{j=1}^{b} \sum_{i=1}^{a_j} X_{ij} / \sum_{j=1}^{b} a_j$ and $\bar{X}_{\cdot j} = \sum_{i=1}^{a_j} X_{ij}/a_j$. If $\mu_1 = \mu_2 = \cdots = \mu_b$, show that Q'/σ^2 and Q_3'/σ^2 have chi-square distributions. Prove that Q_3' and Q_4' are stochastically independent, and hence Q_4'/σ^2 also has a chi-square distribution. If the likelihood ratio λ is used to test $H_0: \mu_1 = \mu_2 = \cdots = \mu_b = \mu$, μ unspecified and σ^2 unknown, against all possible alternatives, show that $\lambda \leq \lambda_0$ is equivalent to the computed $F \geq c$, where

$$F = \frac{\left(\sum_{j=1}^{b} a_j - b\right) Q_4'}{(b-1)Q_3'}.$$

What is the distribution of F when H_0 is true?

10.5 Noncentral χ^2 and Noncentral F

Let $X_1, X_2, \ldots, X_n$ denote mutually stochastically independent random variables that are $n(\mu_i, \sigma^2)$, $i = 1, 2, \ldots, n$ and let $Y = \sum_{1}^{n} X_i^2/\sigma^2$. If each μ_i is zero, we know that Y is $\chi^2(n)$. We shall now investigate the distribution of Y when each μ_i is not zero. The moment-generating function of Y is given by

$$\begin{aligned} M(t) &= E\left[\exp\left(t \sum_{i=1}^{n} \frac{X_i^2}{\sigma^2}\right)\right] \\ &= \prod_{i=1}^{n} E\left[\exp\left(t \frac{X_i^2}{\sigma^2}\right)\right]. \end{aligned}$$

Consider

$$E\left[\exp\left(\frac{tX_i^2}{\sigma^2}\right)\right] = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{tx_i^2}{\sigma^2} - \frac{(x_i - \mu_i)^2}{2\sigma^2}\right] dx_i.$$

The integral exists if $t < \frac{1}{2}$. To evaluate the integral, note that

$$\begin{aligned} \frac{tx_i^2}{\sigma^2} - \frac{(x_i - \mu_i)^2}{2\sigma^2} &= -\frac{x_i^2(1-2t)}{2\sigma^2} + \frac{2\mu_i x_i}{2\sigma^2} - \frac{\mu_i^2}{2\sigma^2} \\ &= \frac{t\mu_i^2}{\sigma^2(1-2t)} - \frac{1-2t}{2\sigma^2}\left(x_i - \frac{\mu_i}{1-2t}\right)^2. \end{aligned}$$

Accordingly, with $t < \frac{1}{2}$, we have

$$E\left[\exp\left(\frac{tX_i^2}{\sigma^2}\right)\right] = \exp\left[\frac{t\mu_i^2}{\sigma^2(1-2t)}\right]\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1-2t}{2\sigma^2}\left(x_i - \frac{\mu_i}{1-2t}\right)^2\right]dx_i.$$

If we multiply the integrand by $\sqrt{1-2t}$, $t < \frac{1}{2}$, we have the integral of a normal p.d.f. with mean $\mu_i/(1-2t)$ and variance $\sigma^2/(1-2t)$. Thus

$$E\left[\exp\left(\frac{tX_i^2}{\sigma^2}\right)\right] = \frac{1}{\sqrt{1-2t}}\exp\left[\frac{t\mu_i^2}{\sigma^2(1-2t)}\right],$$

and the moment-generating function of $Y = \sum_1^n X_i^2/\sigma^2$ is given by

$$M(t) = \frac{1}{(1-2t)^{n/2}}\exp\left[\frac{t\sum_1^n \mu_i^2}{\sigma^2(1-2t)}\right], \qquad t < \tfrac{1}{2}.$$

A random variable which has a moment-generating function of the functional form

$$M(t) = \frac{1}{(1-2t)^{r/2}}e^{t\theta/(1-2t)},$$

where $t < \frac{1}{2}$, $0 < \theta$, and r is a positive integer, is said to have a *noncentral chi-square distribution* with r degrees of freedom and noncentrality parameter θ. If one sets the noncentrality parameter $\theta = 0$, one has $M(t) = (1-2t)^{-r/2}$, which is the moment-generating function of a random variable which is $\chi^2(r)$. Such a random variable can appropriately be called a *central chi-square variable*. We shall use the symbol $\chi^2(r, \theta)$ to denote a noncentral chi-square distribution which has the parameters r and θ; and we shall say that a random variable is $\chi^2(r, \theta)$ to mean that the random variable has this kind of distribution. The symbol $\chi^2(r, 0)$ is equivalent to $\chi^2(r)$. Thus our random variable $Y = \sum_1^n X_i^2/\sigma^2$ of this section is $\chi^2\left(n, \sum_1^n \mu_i^2/\sigma^2\right)$. If each μ_i is equal to zero, then Y is $\chi^2(n, 0)$ or, more simply, Y is $\chi^2(n)$.

The noncentral chi-square variables in which we have interest are certain quadratic forms, in normally distributed variables, divided by a variance σ^2. In our example, it is worth noting that the noncentrality parameter of $\sum_1^n X_i^2/\sigma^2$, which is $\sum_1^n \mu_i^2/\sigma^2$, may be computed by replacing each X_i in the quadratic form by its mean μ_i, $i = 1, 2, \ldots, n$. This is

no fortuitous circumstance; any quadratic form $Q = Q(X_1, \ldots, X_n)$ in normally distributed variables, which is such that Q/σ^2 is $\chi^2(r, \theta)$, has $\theta = Q(\mu_1, \mu_2, \ldots, \mu_n)/\sigma^2$; and if Q/σ^2 is a chi-square variable (central or noncentral) for certain real values of $\mu_1, \mu_2, \ldots, \mu_n$, it is chi-square (central or noncentral) for *all* real values of these means.

It should be pointed out that Theorem 1 of this chapter (p. 316) is valid whether the random variables are central or noncentral chi-square variables.

We next discuss a noncentral F variable. If U and V are stochastically independent and are, respectively, $\chi^2(r_1)$ and $\chi^2(r_2)$, the random variable F has been defined by $F = r_2U/r_1V$. Now suppose, in particular, that U is $\chi^2(r_1, \theta)$, V is $\chi^2(r_2)$, and that U and V are stochastically independent. The random variable r_2U/r_1V is called a *noncentral* F variable with r_1 and r_2 degrees of freedom and with noncentrality parameter θ. Note that the noncentrality parameter of F is precisely the noncentrality parameter of the random variable U which is $\chi^2(r_1, \theta)$.

Tables of noncentral chi-square and noncentral F are available in the literature. However, like those of noncentral t, they are too bulky to be put in this book.

EXERCISES

10.23. Let Y_i, $i = 1, 2, \ldots, n$, denote mutually stochastically independent random variables that are, respectively, $\chi^2(r_i, \theta_i)$, $i = 1, 2, \ldots, n$. Prove that $Z = \sum_1^n Y_i$ is $\chi^2\left(\sum_1^n r_i, \sum_1^n \theta_i\right)$.

10.24. Compute the mean and the variance of a random variable which is $\chi^2(r, \theta)$.

10.25. Compute the mean of a random variable which has a noncentral F distribution with degrees of freedom r_1 and $r_2 > 2$ and noncentrality parameter θ.

10.26. Show that the square of a noncentral T random variable is a noncentral F random variable.

10.27. Let X_1 and X_2 be two stochastically independent random variables. Let X_1 and $Y = X_1 + X_2$ be $\chi^2(r_1, \theta_1)$ and $\chi^2(r, \theta)$, respectively. Here $r_1 < r$ and $\theta_1 \leq \theta$. Show that X_2 is $\chi^2(r - r_1, \theta - \theta_1)$.

10.6 The Analysis of Variance

The problem considered in Section 10.4 is an example of a method of statistical inference called the *analysis of variance*. This method

derives its name from the fact that the quadratic form abS^2, which is a total sum of squares, is resolved into several component parts. In this section other problems in the analysis of variance will be investigated.

Let X_{ij}, $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, b$, denote $n = ab$ random variables which are mutually stochastically independent and have normal distributions with common variance σ^2. The means of these normal distributions are $\mu_{ij} = \mu + \alpha_i + \beta_j$, where $\sum_1^a \alpha_i = 0$ and $\sum_1^b \beta_j = 0$. For example, take $a = 2$, $b = 3$, $\mu = 5$, $\alpha_1 = 1$, $\alpha_2 = -1$, $\beta_1 = 1$, $\beta_2 = 0$, and $\beta_3 = -1$. Then the $ab = 6$ random variables have means

$$\begin{aligned} \mu_{11} &= 7, & \mu_{12} &= 6, & \mu_{13} &= 5, \\ \mu_{21} &= 5, & \mu_{22} &= 4, & \mu_{23} &= 3. \end{aligned}$$

Had we taken $\beta_1 = \beta_2 = \beta_3 = 0$, then the six random variables would have had means

$$\begin{aligned} \mu_{11} &= 6, & \mu_{12} &= 6, & \mu_{13} &= 6, \\ \mu_{21} &= 4, & \mu_{22} &= 4, & \mu_{23} &= 4. \end{aligned}$$

Thus, if we wish to test the composite hypothesis that

$$\begin{aligned} &\mu_{11} = \mu_{12} = \cdots = \mu_{1b}, \\ &\mu_{21} = \mu_{22} = \cdots = \mu_{2b}, \\ &\quad\vdots \\ &\mu_{a1} = \mu_{a2} = \cdots = \mu_{ab}, \end{aligned}$$

we could say that we are testing the composite hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_b$ (and hence each $\beta_j = 0$, since their sum is zero). On the other hand, the composite hypothesis

$$\begin{aligned} &\mu_{11} = \mu_{21} = \cdots = \mu_{a1}, \\ &\mu_{12} = \mu_{22} = \cdots = \mu_{a2}, \\ &\quad\vdots \\ &\mu_{1b} = \mu_{2b} = \cdots = \mu_{ab}, \end{aligned}$$

is the same as the composite hypothesis that $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$.

Remarks. The model just described, and others similar to it, are widely used in statistical applications. Consider a situation in which it is desirable to investigate the effects of two factors which influence an outcome. Thus the variety of a grain and the type of fertilizer used influence the yield; or the teacher and the size of a class may influence the score on a standard test.

Let X_{ij} denote the yield from the use of variety i of a grain and type j of fertilizer. A test of the hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_b = 0$ would then be a test of the hypothesis that the mean yield of each variety of grain is the same regardless of the type of fertilizer used.

There is no loss of generality in assuming that $\sum_1^a \alpha_i = \sum_1^b \beta_j = 0$. To see this, let $\mu_{ij} = \mu' + \alpha_i' + \beta_j'$. Write $\bar{\alpha}' = \sum \alpha_i'/a$ and $\bar{\beta}' = \sum \beta_j'/b$. We have $\mu_{ij} = (\mu' + \bar{\alpha}' + \bar{\beta}') + (\alpha_i' - \bar{\alpha}') + (\beta_j' - \bar{\beta}') = \mu + \alpha_i + \beta_j$, where $\sum \alpha_i = \sum \beta_j = 0$.

A likelihood ratio test will be used to test the composite hypothesis $H_0: \beta_1 = \beta_2 = \cdots = \beta_b = 0$ against all possible alternative hypotheses. Here the parameter space is

$$\Omega = \left\{ (\mu, \alpha_1, \ldots, \alpha_a, \beta_1, \ldots, \beta_b, \sigma^2); \begin{array}{l} -\infty < \mu < \infty, \\ -\infty < \alpha_i < \infty, \ \sum_1^a \alpha_i = 0 \\ -\infty < \beta_j < \infty, \ \sum_1^b \beta_j = 0 \\ 0 < \sigma^2 < \infty \end{array} \right\}$$

and

$$\omega = \left\{ (\mu, \alpha_1, \ldots, \alpha_a, \beta_1, \ldots, \beta_b, \sigma^2); \begin{array}{l} -\infty < \mu < \infty \\ -\infty < \alpha_i < \infty, \ \sum_1^a \alpha_i = 0 \\ \beta_1 = \beta_2 = \cdots = \beta_b = 0 \\ 0 < \sigma^2 < \infty \end{array} \right\}.$$

The likelihood functions, denoted by $L(\omega)$ and $L(\Omega)$, are, respectively,

$$L(\omega) = \left(\frac{1}{2\pi\sigma^2}\right)^{ab/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \mu - \alpha_i)^2\right]$$

and

$$L(\Omega) = \left(\frac{1}{2\pi\sigma^2}\right)^{ab/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \mu - \alpha_i - \beta_j)^2\right].$$

Consider first the problem of maximizing $L(\Omega)$. To accomplish this, we wish, for each σ^2, to minimize the sum of squares

$$z = \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \mu - \alpha_i - \beta_j)^2$$

with respect to $\mu, \alpha_1, \ldots, \alpha_a, \beta_1, \ldots, \beta_b$. If we let

$$\bar{x} = \frac{\sum_{j=1}^{b} \sum_{i=1}^{a} x_{ij}}{ab}, \qquad \bar{x}_{i\cdot} = \frac{\sum_{j=1}^{b} x_{ij}}{b}, \qquad \bar{x}_{\cdot j} = \frac{\sum_{i=1}^{a} x_{ij}}{a},$$

the sum of squares z can be rewritten as

$$z = \sum_{j=1}^{b} \sum_{i=1}^{a} [(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}) + (\bar{x} - \mu) + (\bar{x}_{i\cdot} - \bar{x} - \alpha_i) + (\bar{x}_{\cdot j} - \bar{x} - \beta_j)]^2.$$

Since $\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0$, we obtain

$$(1) \quad z = \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2 + ab(\bar{x} - \mu)^2 + b \sum_{i=1}^{a} (\bar{x}_{i\cdot} - \bar{x} - \alpha_i)^2 + a \sum_{j=1}^{b} (\bar{x}_{\cdot j} - \bar{x} - \beta_j)^2,$$

because each of the other terms sums to zero. The values of μ, α_i, and β_j which minimize z are, respectively, $\bar{x}$, $\bar{x}_{i\cdot} - \bar{x}$, and $\bar{x}_{\cdot j} - \bar{x}$, because these values make each of the last three nonnegative terms of Equation (1) equal to zero. Note that these values satisfy the conditions $\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0$. To find the value of σ^2 which maximizes $L(\Omega)$, equate the partial derivative

$$(2) \quad \frac{\partial \ln L(\Omega)}{\partial(\sigma^2)} = -\frac{ab}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \mu - \alpha_i - \beta_j)^2$$

to zero and replace μ, α_i, and β_j by $\bar{x}$, $\bar{x}_{i\cdot} - \bar{x}$, and $\bar{x}_{\cdot j} - \bar{x}$, respectively. The solution for σ^2 is

$$w = \frac{1}{ab} \sum_{j=1}^{b} \sum_{i=1}^{a} [x_{ij} - \bar{x} - (\bar{x}_{i\cdot} - \bar{x}) - (\bar{x}_{\cdot j} - \bar{x})]^2$$

$$= \frac{1}{ab} \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2.$$

These values maximize $L(\Omega)$ and the maximum is

$$L(\hat{\Omega}) = \left(\frac{1}{2\pi w}\right)^{ab/2} e^{-ab/2}.$$

Now consider the problem of maximizing $L(\omega)$. That is, we wish to minimize z when $\beta_1 = \cdots = \beta_b = 0$. Accordingly, if in Equation (1) we set $\beta_1 = \cdots = \beta_b = 0$, we see immediately that $\bar{x}$ and $\bar{x}_{i\cdot} - \bar{x}$ are the respective values of μ and α_i which minimize z under the conditions that $\beta_1 = \cdots = \beta_b = \sum_{i=1}^{a} \alpha_i = 0$. If the partial derivative (2) is equated to zero, when $\beta_1 = \cdots = \beta_b = 0$, and if μ and α_i are replaced, respectively, by $\bar{x}$ and $\bar{x}_{i\cdot} - \bar{x}$, we see that the solution for σ^2 is

$$w' = \frac{1}{ab} \sum_{j=1}^{b} \sum_{i=1}^{a} (x_{ij} - \bar{x}_{i\cdot})^2.$$

These values maximize $L(\omega)$ and the maximum is

$$L(\hat{\omega}) = \left(\frac{1}{2\pi w'}\right)^{ab/2} e^{-ab/2}.$$

Finally, the likelihood ratio λ is given by

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \left(\frac{w}{w'}\right)^{ab/2}.$$

In the notation of Section 10.3, the statistic defined by $\lambda^{2/ab}$ is

$$\frac{\sum_{j=1}^{b}\sum_{i=1}^{a}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2}{\sum_{j=1}^{b}\sum_{i=1}^{a}(X_{ij} - \bar{X}_{i\cdot})^2}.$$

The denominator in the preceding ratio may be written (Exercise 10.29) in the form

$$\sum_{j=1}^{b}\sum_{i=1}^{a}[(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}) + (\bar{X}_{\cdot j} - \bar{X})]^2$$
$$= \sum_{j=1}^{b}\sum_{i=1}^{a}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2 + a\sum_{j=1}^{b}(\bar{X}_{\cdot j} - \bar{X})^2.$$

Accordingly, in the notation of Section 10.3, the statistic defined by $\lambda^{2/ab}$ is

$$\left[1 + \frac{a\sum_{j=1}^{b}(\bar{X}_{\cdot j} - \bar{X})^2}{\sum_{j=1}^{b}\sum_{i=1}^{a}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2}\right]^{-1} = \frac{1}{1 + Q_4/Q_5}.$$

We reject the hypothesis H_0 if $\lambda \leq \lambda_0$. To find λ_0 so that the test of H_0 has a desired significance level α, we must assume that the hypothesis H_0 is true. But here, unlike the situation in Section 10.3, when the hypothesis is true, the $n = ab$ mutually stochastically independent normally distributed random variables X_{ij} do not have the same means. Thus the results of Examples 1, 2, and 3 of Section 10.3 may not be valid; that is, in particular, Q_4 and Q_5 may not be stochastically independent, and Q_4/σ^2 and Q_5/σ^2 may not have their chi-square distributions. Fortunately, however, these fears are groundless, as may be seen by the following argument. With $\beta_1 = \cdots = \beta_b = 0$, then certainly if $\alpha_1 = \cdots = \alpha_a = 0$ the ab random variables X_{ij} have the same mean μ. We know from Section 10.3 (p. 319) that Q_4/σ^2 and Q_5/σ^2 are stochastically independent chi-square variables. Thus, in accordance with the discussion in Section 10.5, Q_4/σ^2 and

Q_5/σ^2 are stochastically independent and chi-square (possibly noncentral), even though the mean of X_{ij} is $\mu + \alpha_i$. We shall compute the noncentrality parameters of Q_4/σ^2 and Q_5/σ^2 when H_0 is true. We have $E(X_{ij}) = \mu + \alpha_i$, $E(\bar{X}_{i\cdot}) = \mu + \alpha_i$, $E(\bar{X}_{\cdot j}) = \mu$, and $E(\bar{X}) = \mu$. Thus the noncentrality parameter of Q_4/σ^2 is

$$\frac{a \sum_{j=1}^{b} (\mu - \mu)^2}{\sigma^2} = 0,$$

and the noncentrality parameter of Q_5/σ^2 is

$$\frac{\sum_{j=1}^{b} \sum_{i=1}^{a} (\mu + \alpha_i - \mu - \alpha_i - \mu + \mu)^2}{\sigma^2} = 0.$$

Thus, not only are Q_4/σ^2 and Q_5/σ^2 stochastically independent but also each, when H_0 is true, is a central chi-square variable.

The significance level of the test of H_0 is

$$\alpha = \Pr\left[\frac{1}{1 + Q_4/Q_5} \leq \lambda_0^{2/ab}; H_0\right]$$

$$= \Pr\left[\frac{(a-1)Q_4}{Q_5} \geq c; H_0\right],$$

where

$$c = (a-1)(\lambda_0^{-2/ab} - 1).$$

But

$$F = \frac{Q_4/[\sigma^2(b-1)]}{Q_5/[\sigma^2(a-1)(b-1)]} = \frac{(a-1)Q_4}{Q_5}$$

has an F distribution with $b - 1$ and $(a - 1)(b - 1)$ degrees of freedom. Hence the test of the composite hypothesis H_0: $\beta_1 = \beta_2 = \cdots = \beta_b = 0$, or H_0: $\mu_{ij} = \mu + \alpha_i$, $\sum_1^a \alpha_i = 0$, against the alternative composite hypothesis H_1: $\mu_{ij} = \mu + \alpha_i + \beta_j$, $\sum_1^a \alpha_i = \sum_1^b \beta_j = 0$, may be based on an F statistic. The constant c is so selected as to yield the desired value of α.

If we are to compute the power function of the test, we need the distribution of F when H_0 is not true. From Section 10.5 we know, when H_1 is true, that Q_4/σ^2 and Q_5/σ^2 are stochastically independent (central or noncentral) chi-square variables. We shall compute the noncentrality parameters of Q_4/σ^2 and Q_5/σ^2 when H_1 is true. We have

$E(X_{ij}) = \mu + \alpha_i + \beta_j$, $E(\bar{X}_{i\cdot}) = \mu + \alpha_i$, $E(\bar{X}_{\cdot j}) = \mu + \beta_j$ and $E(\bar{X}) = \mu$. Accordingly, the noncentrality parameter of Q_4/σ^2 is

$$\frac{a\sum_{j=1}^{b}(\mu + \beta_j - \mu)^2}{\sigma^2} = \frac{a\sum_{j=1}^{b}\beta_j^2}{\sigma^2}$$

and that of Q_5/σ^2 is

$$\frac{\sum_{j=1}^{b}\sum_{i=1}^{a}(\mu + \alpha_i + \beta_j - \mu - \alpha_i - \mu - \beta_j + \mu)^2}{\sigma^2} = 0.$$

Thus, if the hypothesis H_0 is not true, F has a noncentral F distribution with $b - 1$ and $(a - 1)(b - 1)$ degrees of freedom and noncentrality parameter $a\sum_{j=1}^{k}\beta_j^2/\sigma^2$. The desired probabilities can then be found in tables of the noncentral F distribution.

The analysis of variance problem which has just been discussed is usually referred to as a *two-way classification with one observation per cell.* Each combination of i and j determines a cell; thus there is a total of ab cells in this model. Let us now investigate another two-way classification problem, but in this case we take $c > 1$ stochastically independent observations per cell.

Let X_{ijk}, $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$, and $k = 1, 2, \ldots, c$, denote $n = abc$ random variables which are mutually stochastically independent and which have normal distributions with common, but unknown, variance σ^2. The mean of each X_{ijk}, $k = 1, 2, \ldots, c$, is $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, where $\sum_{i=1}^{a}\alpha_i = 0$, $\sum_{j=1}^{b}\beta_j = 0$, $\sum_{i=1}^{a}\gamma_{ij} = 0$, and $\sum_{j=1}^{b}\gamma_{ij} = 0$. For example, take $a = 2$, $b = 3$, $\mu = 5$, $\alpha_1 = 1$, $\alpha_2 = -1$, $\beta_1 = 1$, $\beta_2 = 0$, $\beta_3 = -1$, $\gamma_{11} = 1$, $\gamma_{12} = 1$, $\gamma_{13} = -2$, $\gamma_{21} = -1$, $\gamma_{22} = -1$, and $\gamma_{23} = 2$. Then the means are

$$\begin{aligned} \mu_{11} &= 8, & \mu_{12} &= 7, & \mu_{13} &= 3, \\ \mu_{21} &= 4, & \mu_{22} &= 3, & \mu_{23} &= 5. \end{aligned}$$

Note that, if each $\gamma_{ij} = 0$, then

$$\begin{aligned} \mu_{11} &= 7, & \mu_{12} &= 6, & \mu_{13} &= 5, \\ \mu_{21} &= 5, & \mu_{22} &= 4, & \mu_{23} &= 3. \end{aligned}$$

That is, if $\gamma_{ij} = 0$, each of the means in the first row is two greater than the corresponding mean in the second row. In general, if each $\gamma_{ij} = 0$, the means of row i_1 differ from the corresponding means of row i_2 by a

constant. This constant may be different for different choices of i_1 and i_2. A similar statement can be made about the means of columns j_1 and j_2. The parameter γ_{ij} is called the *interaction* associated with cell (i, j). That is, the interaction between the ith level of one classification and the jth level of the other classification is γ_{ij}. One interesting hypothesis to test is that each interaction is equal to zero. This will now be investigated.

A likelihood ratio test will be used to test the hypothesis

$$H_0: \gamma_{ij} = 0, \qquad i = 1, 2, \ldots, a, \; j = 1, 2, \ldots, b,$$

against all possible alternatives. In Exercise 10.34, the student is asked to show that the statistic defined by the likelihood ratio λ is, in terms of the notation of Exercise 10.20, a function of the statistic

$$(3) \quad F = \frac{\left[c \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X})^2\right]/[(a-1)(b-1)]}{[\sum \sum \sum (X_{ijk} - \bar{X}_{ij.})^2]/[ab(c-1)]}.$$

That same exercise, with the statement made in Section 10.5 (p. 327) can be used to show that this F has an F distribution with $(a-1)(b-1)$ and $ab(c-1)$ degrees of freedom. The student should verify that the noncentrality parameter of this F distribution is equal to $c \sum_{j=1}^{b} \sum_{i=1}^{a} \gamma_{ij}^2/\sigma^2$. Thus F is central when $H_0: \gamma_{ij} = 0$, $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$, is true.

EXERCISES

10.28. If, in the two-way classification with one observation per cell, $a = 4$ and $b = 3$, test at the 5 per cent significance level the hypothesis $\beta_1 = \beta_2 = \beta_3 = 0$ if the observed values of X_{ij} are

$$\begin{array}{ccc} 3 & 5 & 7 \\ -1 & 2 & 2 \\ 0 & 2 & 5 \\ 6 & 6 & 10. \end{array}$$

Hint. The computed value of Q_5 can be determined by $Q_5 = Q - Q_2 - Q_4$, once the values of terms in the right-hand member are found.

10.29. Show that

$$\sum_{j=1}^{b} \sum_{i=1}^{a} (X_{ij} - \bar{X}_{i.})^2 = \sum_{j=1}^{b} \sum_{i=1}^{a} (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})^2 + a \sum_{j=1}^{b} (\bar{X}_{.j} - \bar{X})^2.$$

10.30. Let X_{ij} denote $n = ab$ random variables, mutually stochastically independent and having normal distributions with means $\mu_{ij} = \mu + \alpha_i + \beta_j$, $\sum_1^a \alpha_i = 0$, $\sum_1^b \beta_j = 0$, and common but unknown variance σ^2. Consider the hypothesis H_0: $\mu_{ij} = \mu + \beta_j$ (or H_0: $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$). Show that a likelihood ratio test of H_0 against all possible alternatives may be based, in the notation of Section 10.3, on the $F = (b-1)(Q_2/Q_6)$ statistic having $(a-1)$ and $(a-1)(b-1)$ degrees of freedom.

10.31. Use the data of Exercise 10.28 to test at the 5 per cent significance level the hypothesis in Exercise 10.30.

10.32. Let the random variables X_{ij} be defined as in Exercise 10.30. Consider a likelihood ratio test of the composite hypothesis H_0: $\mu_{ij} = \mu$ (or H_0: $\alpha_i = \beta_j = 0$, $i = 1, \ldots, a$, $j = 1, \ldots, b$) against H_1: $\mu_{ij} = \mu + \alpha_i + \beta_j$, $\sum_{i=1}^a \alpha_i = \sum_{j=1}^b \beta_j = 0$. Show that this test can be based on an F statistic having $a + b - 2$ and $(a-1)(b-1)$ degrees of freedom.

10.33. Let X_{ijk}, $i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$; $k = 1, 2, \ldots, c$, denote $n = abc$ mutually stochastically independent random variables which have normal distributions with common unknown variance σ^2. Let $E(X_{ijk}) = \mu + \alpha_i + \beta_j + \delta_k$, where $\sum_1^a \alpha_i = \sum_1^b \beta_j = \sum_1^c \delta_k = 0$.

(a) Show that the maximum likelihood statistics for μ, α_i, β_j, and δ_k are $\bar{X}$, $\bar{X}_{i..} - \bar{X}$, $\bar{X}_{.j.} - \bar{X}$, and $\bar{X}_{..k} - \bar{X}$, respectively.

(b) Show that the likelihood ratio test of H_0: $\delta_1 = \delta_2 = \cdots = \delta_c = 0$ against all possible alternatives may be based on the statistic

$$F = \frac{ab \sum_1^c (\bar{X}_{..k} - \bar{X})^2/(c-1)}{\sum_1^c \sum_1^b \sum_1^a (X_{ijk} - \bar{X}_{i..} - \bar{X}_{.j.} - \bar{X}_{..k} + 2\bar{X})^2/(abc - a - b - c + 2)}.$$

(c) Prove, under the alternative hypothesis, that F has a noncentral F distribution with $c - 1$ and $abc - a - b - c + 2$ degrees of freedom, and with noncentrality parameter $ab \sum_1^c \delta_k^2/\sigma^2$.

10.34. Show that the statistic defined by the likelihood ratio λ which is used to test H_0: $\gamma_{ij} = 0$ is a function of the F statistic given by Display (3). Demonstrate that this F has an F distribution with $(a-1)(b-1)$ and $ab(c-1)$ degrees of freedom. If at least one $\gamma_{ij} \neq 0$, show that the noncentrality parameter is equal to $c \sum_{j=1}^b \sum_{i=1}^a \gamma_{ij}^2/\sigma^2$.

10.7 A Regression Problem

Consider a laboratory experiment the outcome of which depends upon the temperature; that is, the technician first sets a temperature

dial at a fixed point c and subsequently observes the outcome of the experiment for that dial setting. From past experience, the technician knows that if he repeats the experiment with the temperature dial set at the same point c, he is not likely to observe precisely the same outcome. He then assumes that the outcome of his experiment is a random variable X whose distribution depends not only upon certain unknown parameters but also upon a nonrandom variable c which he can choose more or less at pleasure. Let $c_1, c_2, \ldots, c_n$ denote n arbitrarily selected values of c (but not all equal) and let X_i denote the outcome of the experiment when $c = c_i$, $i = 1, 2, \ldots, n$. We then have the n pairs $(X_1, c_1), \ldots, (X_n, c_n)$ in which the X_i are random variables but the c_i are known numbers and $i = 1, 2, \ldots, n$. Once the n experiments have been performed (the first with $c = c_1$, the second with $c = c_2$ and so on) and the outcome of each recorded, we have the n pairs of known numbers $(x_1, c_1), \ldots, (x_n, c_n)$. These numbers are to be used to make statistical inferences about the unknown parameters in the distribution of the random variable X. Certain problems of this sort are called *regression* problems and we shall study a particular one in some detail.

Let $c_1, c_2, \ldots, c_n$ be n given numbers, not all equal, and let $\bar{c} = \sum_1^n c_i/n$. Let $X_1, X_2, \ldots, X_n$ be n mutually stochastically independent random variables with joint p.d.f.

$$L(\alpha, \beta, \sigma^2; x_1, x_2, \ldots, x_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_1^n [x_i - \alpha - \beta(c_i - \bar{c})]^2\right\}.$$

Thus each X_i has a normal distribution with the same variance σ^2, but the means of these distributions are $\alpha + \beta(c_i - \bar{c})$. Since the c_i are not all equal, in this regression problem the means of the normal distributions depend upon the choice of $c_1, c_2, \ldots, c_n$. We shall investigate ways of making statistical inferences about the parameters α, β, and σ^2.

It is easy to show (see Exercise 10.35) that the maximum likelihood statistics for α, β, and σ^2 are

$$\hat{\alpha} = \frac{\sum_1^n X_i}{n} = \bar{X},$$

$$\hat{\beta} = \frac{\sum_1^n (c_i - \bar{c})(X_i - \bar{X})}{\sum_1^n (c_i - \bar{c})^2} = \frac{\sum_1^n (c_i - \bar{c})X_i}{\sum_1^n (c_i - \bar{c})^2},$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_1^n [X_i - \hat{\alpha} - \hat{\beta}(c_i - \bar{c})]^2.$$

Since $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of $X_1, X_2, \ldots, X_n$, each is normally distributed (Theorem 1, p. 158). It is easy to show (Exercise 10.36) that their respective means are α and β and their respective variances are σ^2/n and $\sigma^2/\sum_1^n (c_i - \bar{c})^2$.

Consider next the algebraic identity (Exercise 10.37)

$$\begin{aligned}\sum_1^n [X_i - \alpha - \beta(c_i - \bar{c})]^2 &= \sum_1^n \{(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)(c_i - \bar{c}) \\ &\quad + [X_i - \hat{\alpha} - \hat{\beta}(c_i - \bar{c})]\}^2 \\ &= n(\hat{\alpha} - \alpha)^2 + (\hat{\beta} - \beta)^2 \sum_1^n (c_i - \bar{c})^2 \\ &\quad + \sum_1^n [X_i - \hat{\alpha} - \hat{\beta}(c_i - \bar{c})]^2,\end{aligned}$$

or

$$\sum_1^n [X_i - \alpha - \beta(c_i - \bar{c})]^2 = n(\hat{\alpha} - \alpha)^2 + (\hat{\beta} - \beta)^2 \sum_1^n (c_i - \bar{c})^2 + n\hat{\sigma}^2,$$

or, for brevity,

$$Q = Q_1 + Q_2 + Q_3.$$

Here Q, Q_1, Q_2, and Q_3 are real quadratic forms in the variables

$$X_i - \alpha - \beta(c_i - \bar{c}), \qquad i = 1, 2, \ldots, n.$$

In this equation, Q represents the sum of the squares of n mutually stochastically independent random variables that have normal distributions with means zero and variances σ^2. Thus Q/σ^2 has a chi-square distribution with n degrees of freedom. Each of the random variables $\sqrt{n}(\hat{\alpha} - \alpha)/\sigma$ and $\sqrt{\sum_1^n (c_i - \bar{c})^2}(\hat{\beta} - \beta)/\sigma$ has a normal distribution with zero mean and unit variance; thus, each of Q_1/σ^2 and Q_2/σ^2 has a chi-square distribution with one degree of freedom. Since Q_3 is nonnegative, we have, in accordance with the theorem of Section 10.3, that Q_1, Q_2, and Q_3 are mutually stochastically independent, so that Q_3/σ^2 has a chi-square distribution with $n - 1 - 1 = n - 2$ degrees of freedom. Then each of the random variables

$$T_1 = \frac{[\sqrt{n}(\hat{\alpha} - \alpha)]/\sigma}{\sqrt{Q_3/[\sigma^2(n-2)]}} = \frac{\hat{\alpha} - \alpha}{\sqrt{\hat{\sigma}^2/(n-2)}}$$

and

$$T_2 = \frac{\left[\sqrt{\sum_1^n (c_i - \bar{c})^2}(\hat{\beta} - \beta)\right]/\sigma}{\sqrt{Q_3/[\sigma^2(n-2)]}} = \frac{\hat{\beta} - \beta}{\sqrt{n\hat{\sigma}^2/\left[(n-2)\sum_1^n (c_i - \bar{c})^2\right]}}$$

has a t distribution with $n - 2$ degrees of freedom. These facts enable us to obtain confidence intervals for α and β. The fact that $n\hat{\sigma}^2/\sigma^2$ has a chi-square distribution with $n - 2$ degrees of freedom provides a means of determining a confidence interval for σ^2. These are some of the statistical inferences about the parameters to which reference was made in the introductory remarks of this section.

Remark. The more discerning student should quite properly question our constructions of T_1 and T_2 immediately above. We know that the *squares* of the linear forms are stochastically independent of $Q_3 = n\hat{\sigma}^2$, but we do not know, at this time, that the linear forms themselves enjoy this independence. This problem arises again in Section 10.8. In Exercise 12.15, a more general problem is proposed of which the present case is a special instance.

EXERCISES

10.35. Verify that the maximum likelihood statistics for α, β, and σ^2 are the $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma}^2$ given in this section.

10.36. Show that $\hat{\alpha}$ and $\hat{\beta}$ have the respective means α and β and the respective variances σ^2/n and $\sigma^2/\sum_1^n (c_i - \bar{c})^2$.

10.37. Verify that $\sum_1^n [X_i - \alpha - \beta(c_i - \bar{c})]^2 = Q_1 + Q_2 + Q_3$, as stated in the text.

10.38. Let the mutually stochastically independent random variables $X_1, X_2, \ldots, X_n$ have, respectively, the probability density functions $n(\beta c_i, \gamma^2 c_i^2)$, $i = 1, 2, \ldots, n$, where the given numbers $c_1, c_2, \ldots, c_n$ are not all equal and no one is zero. Find the maximum likelihood statistics for β and γ^2.

10.39. Let the mutually stochastically independent random variables $X_1, \ldots, X_n$ have the joint p.d.f.

$$L(\alpha, \beta, \sigma^2; x_1, \ldots, x_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_1^n [x_i - \alpha - \beta(c_i - \bar{c})]^2\right\},$$

where the given numbers $c_1, c_2, \ldots, c_n$ are not all equal. Let $H_0: \beta = 0$ (α and σ^2 unspecified). It is desired to use a likelihood ratio test to test H_0

against all possible alternatives. Find λ and see whether the test can be based on a familiar statistic. *Hint.* In the notation of this section show that

$$\sum_1^n (X_i - \hat{\alpha})^2 = Q_3 + \hat{\beta}^2 \sum_1^n (c_i - \bar{c})^2.$$

10.8 A Test of Stochastic Independence

Let X and Y have a bivariate normal distribution with means μ_1 and μ_2, positive variances σ_1^2 and σ_2^2, and correlation coefficient ρ. We wish to test the hypothesis that X and Y are stochastically independent. Because two jointly normally distributed random variables are stochastically independent if and only if $\rho = 0$, we test the hypothesis H_0: $\rho = 0$ against the hypothesis H_1: $\rho \neq 0$. A likelihood ratio test will be used. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ denote a random sample of size $n > 2$ from the bivariate normal distribution. Although it is fairly difficult to show, the statistic which is defined by the likelihood ratio λ is a function of the statistic

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

This statistic R is called the *correlation coefficient* of the random sample. The likelihood ratio principle, which calls for the rejection of H_0 if $\lambda \leq \lambda_0$, is equivalent to the computed value of $|R| \geq c$. That is, if the absolute value of the correlation coefficient of the sample is too large, we reject the hypothesis that the correlation coefficient of the distribution is equal to zero. To determine a value of c for a satisfactory significance level, it will be necessary to obtain the distribution of R, or a function of R, when H_0 is true. This will now be done.

Let $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, $n > 2$, where $x_1, x_2, \ldots, x_n$ and $\bar{x} = \sum_1^n x_i/n$ are fixed numbers such that $\sum_1^n (x_i - \bar{x})^2 > 0$. Consider the conditional p.d.f. of $Y_1, Y_2, \ldots, Y_n$, given that $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. Because $Y_1, Y_2, \ldots, Y_n$ are mutually stochastically independent and, with $\rho = 0$, are also mutually stochastically independent of $X_1, X_2, \ldots, X_n$, this conditional p.d.f. is given by

$$\left(\frac{1}{\sqrt{2\pi}\,\sigma_2}\right)^n \exp\left[-\frac{\sum_1^n (y_i - \mu_2)^2}{2\sigma_2^2}\right].$$

The conditional distribution of $\sum_1^n (Y_i - \bar{Y})^2/\sigma_2^2$, given $X_1 = x_1, \ldots, X_n = x_n$, is $\chi^2(n-1)$. Moreover, the conditional distribution of the linear function W of $Y_1, Y_2, \ldots, Y_n$, namely,

$$(1) \qquad W = \frac{\sum_1^n [(x_i - \bar{x})(Y_i - \bar{Y})]}{\sqrt{\sum_1^n (x_i - \bar{x})^2}} = \frac{\sum_1^n [(x_i - \bar{x})Y_i]}{\sqrt{\sum_1^n (x_i - \bar{x})^2}}.$$

is $n(0, \sigma_2^2)$ (see Exercise 10.44). Thus the conditional distribution of W^2/σ_2^2, given $X_1 = x_1, \ldots, X_n = x_n$, is $\chi^2(1)$. We have the algebraic identity (see Exercise 10.45)

$$(2) \qquad \sum_1^n (Y_i - \bar{Y})^2 = W^2 + \sum_1^n \left\{ Y_i - \bar{Y} - \frac{\sum_1^n [(x_i - \bar{x})(Y_i - \bar{Y})]}{\sum_1^n (x_i - \bar{x})^2} (x_i - \bar{x}) \right\}^2.$$

The left-hand member of this equation and the first term of the right-hand member are, when divided by σ_2^2, respectively, conditionally $\chi^2(n-1)$ and conditionally $\chi^2(1)$. In accordance with Theorem 1, the nonnegative quadratic form, say U, which is the second term of the right-hand member of Equation (2) is conditionally stochastically independent of W^2, and, when divided by σ_2^2, is conditionally $\chi^2(n-2)$. Now W/σ_2 is $n(0, 1)$. Then (Remark, p. 338)

$$\frac{W/\sigma_2}{\sqrt{U/[\sigma_2^2(n-2)]}} = \frac{W\sqrt{(n-2)}}{\sqrt{U}}$$

has a conditional t distribution with $n-2$ degrees of freedom. Let

$$R_c = \frac{\sum_1^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_1^n (x_i - \bar{x})^2 \sum_1^n (Y_i - \bar{Y})^2}}.$$

Then we have (Exercise 10.46)

$$(3) \qquad \frac{W\sqrt{n-2}}{\sqrt{U}} = \frac{R_c\sqrt{n-2}}{\sqrt{1 - R_c^2}};$$

this ratio has, given $X_1 = x_1, \ldots, X_n = x_n$, a conditional t distribution with $n - 2$ degrees of freedom. Note that the p.d.f., say $g(t)$, of this t distribution does not depend upon $x_1, x_2, \ldots, x_n$. Now the joint p.d.f. of $X_1, X_2, \ldots, X_n$ and $R\sqrt{n-2}/\sqrt{1-R^2}$, where

$$R = \frac{\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_1^n (X_i - \bar{X})^2 \sum_1^n (Y_i - \bar{Y})^2}},$$

is the product of $g(t)$ and the joint p.d.f. of $X_1, X_2, \ldots, X_n$. Integration on $x_1, x_2, \ldots, x_n$ yields the marginal p.d.f. of $R\sqrt{n-2}/\sqrt{1-R^2}$; because $g(t)$ does not depend upon $x_1, x_2, \ldots, x_n$, it is obvious that this marginal p.d.f. is $g(t)$, the conditional p.d.f. of $R_c\sqrt{n-2}/\sqrt{1-R_c^2}$. The change of variable technique can now be used to find the p.d.f. of R.

Remarks. Since R has, when $\rho = 0$, a conditional distribution which does not depend upon $x_1, x_2, \ldots, x_n$ (and hence that conditional distribution is, in fact, the marginal distribution of R), we have the remarkable fact that R is stochastically independent of $X_1, X_2, \ldots, X_n$. It follows that R is stochastically independent of *every function* of $X_1, X_2, \ldots, X_n$ alone; that is, a function that does not depend upon any Y_i. In like manner, R is stochastically independent of every function of $Y_1, Y_2, \ldots, Y_n$ alone. Moreover, a careful review of the argument reveals that nowhere did we use the fact that X has a marginal distribution which is normal. Thus if X and Y are stochastically independent, and if Y has a normal distribution, then R has the same conditional distribution whatever be the distribution of X, subject to the condition $\sum_1^n (x_i - \bar{x})^2 > 0$. Moreover, if $\Pr\left[\sum_1^n (X_i - \bar{X})^2 > 0\right] = 1$, then R has the same marginal distribution whatever be the distribution of X.

If we write $T = R\sqrt{n-2}/\sqrt{1-R^2}$, where T has a t distribution with $n - 2 > 0$ degrees of freedom, it is easy to show, by the change of variable technique (Exercise 10.47), that the p.d.f. of R is given by

$$(4) \qquad g(r) = \frac{\Gamma[(n-1)/2]}{\Gamma(\frac{1}{2})\Gamma[(n-2)/2]}(1 - r^2)^{(n-4)/2}, \qquad -1 < r < 1,$$
$$= 0 \text{ elsewhere.}$$

We have now solved the problem of the distribution of R, when $\rho = 0$ and $n > 2$, or, perhaps more conveniently, that of $R\sqrt{n-2}/\sqrt{1-R^2}$. The likelihood ratio test of the hypothesis H_0: $\rho = 0$ against all alternatives H_1: $\rho \neq 0$ may be based either on the statistic R or on

the statistic $R\sqrt{n-2}/\sqrt{1-R^2} = T$. In either case, the significance level of the test is

$$\alpha = \Pr(|R| \geq c_1; H_0) = \Pr(|T| \geq c_2; H_0),$$

where the constants c_1 and c_2 are chosen so as to give the desired value of α.

Remark. To obtain the power of the test when the hypothesis H_0: $\rho = 0$ is false, we need the distribution of R when $\rho \neq 0$. One way to obtain this distribution is to use the general formulation of Exercise 8.10 (p. 258). In that exercise, let θ denote the vector $[\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho]$, for which there exist the five joint sufficient statistics $\bar{X}$, $\bar{Y}$, $V_1 = \sum (X_i - \bar{X})^2$, $V_2 = \sum (Y_i - \bar{Y})^2$, and R; and let θ_0 denote the vector $[0, 0, 1, 1, 0]$. In the notation of that exercise, we would have

$$g_1(\bar{x}, \bar{y}, v_1, v_2, r; \theta) = g_1(\bar{x}, \bar{y}, v_1, v_2, r; \theta_0) \frac{L(\theta)}{L(\theta_0)}, \tag{5}$$

where $L(\theta)$ is the joint p.d.f. of $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. Since $\bar{X}$, $\bar{Y}$, V_1, V_2, and R are mutually stochastically independent provided $\theta = \theta_0$, the p.d.f. $g_1(\bar{x}, \bar{y}, v_1, v_2, r; \theta_0)$ can be determined easily. From the $g_1(\bar{x}, \bar{y}, v_1, v_2, r; \theta)$, the marginal p.d.f. of R can then be obtained by integrating on $\bar{x}$, $\bar{y}$, v_1, and v_2.

EXERCISES

10.40. Show that

$$R = \frac{\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_1^n (X_i - \bar{X})^2 \sum_1^n (Y_i - \bar{Y})^2}} = \frac{\sum_1^n X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_1^n X_i^2 - n\bar{X}^2\right)\left(\sum_1^n Y_i^2 - n\bar{Y}^2\right)}}.$$

10.41. A random sample of size $n = 6$ from a bivariate normal distribution yields the value of the correlation coefficient to be 0.89. Would we accept or reject, at the 5 per cent significance level, the hypothesis that $\rho = 0$?

10.42. If, with $\rho = 0$, the bivariate normal p.d.f. is written in the exponential form, what are the four joint complete sufficient statistics $\sum_1^n K_1(X_i, Y_i)$, $\sum_1^n K_2(X_i, Y_i)$, $\sum_1^n K_3(X_i, Y_i)$, and $\sum_1^n K_4(X_i, Y_i)$ for μ_1, μ_2, σ_1^2, and σ_2^2? Why then are $\bar{X}$, $\bar{Y}$, V_1, and V_2 also joint complete sufficient statistics for these parameters?

10.43. If $n = 2$ and $\rho = 0$, show that R is a random variable of the discrete type and find the p.d.f. of R.

10.44. Verify that W of Equation (1) of this section is $n(0, \sigma_2^2)$.

10.45. Verify the algebraic identity (2) of this section.

10.46. Verify Equation (3) of this section.

10.47. Verify the p.d.f. (4) of this section.

10.48. From Equation (5) of this section show that the pair $(\bar{X}, \bar{Y})$ is stochastically independent of (V_1, V_2, R), and show that $\bar{X}$ and $\bar{Y}$ have a bivariate normal distribution with parameters μ_1, μ_2, σ_1^2/n, σ_2^2/n, and ρ. *Hint.* Show that the exponent in $L(\theta_0)$ is $-(v_1 + n\bar{x}^2 + v_2 + n\bar{y}^2)/2$ and that the exponent in $L(\theta)$ is the product of $-\frac{1}{2}(1 - \rho^2)^{-1}$ and

$$\frac{v_1 + n(\bar{x} - \mu_1)^2}{\sigma_1^2} - \frac{2\rho[r\sqrt{v_1 v_2} + n(\bar{x} - \mu_1)(\bar{y} - \mu_2)]}{\sigma_1\sigma_2} + \frac{v_2 + n(\bar{y} - \mu_2)^2}{\sigma_2^2}.$$

10.9 Multiple Comparisons

Consider b mutually stochastically independent random variables which have normal distributions with unknown means $\mu_1, \mu_2, \ldots, \mu_b$, respectively, and with unknown but common variance σ^2. Let $k_1, k_2, \ldots, k_b$ represent b known real constants which are not all zero. We want to find a confidence interval for $\sum_1^b k_j\mu_j$, a linear function of the means $\mu_1, \mu_2, \ldots, \mu_b$. To do this, we take a random sample $X_{1j}, X_{2j}, \ldots, X_{aj}$ of size a from the distribution $n(\mu_j, \sigma^2)$, $j = 1, 2, \ldots, b$. If we denote $\sum_{i=1}^a X_{ij}/a$ by $\bar{X}_{\cdot j}$, then we know that $\bar{X}_{\cdot j}$ is $n(\mu_j, \sigma^2/a)$, that $\sum_{i=1}^a (X_{ij} - \bar{X}_{\cdot j})^2/\sigma^2$ is $\chi^2(a - 1)$, and that the two random variables are stochastically independent. Since the random samples are taken from mutually independent distributions, the $2b$ random variables $\bar{X}_{\cdot j}$, $\sum_{i=1}^a (X_{ij} - \bar{X}_{\cdot j})^2/\sigma^2$, $j = 1, 2, \ldots, b$, are mutually stochastically independent Moreover, $\bar{X}_{\cdot 1}, \bar{X}_{\cdot 2}, \ldots, \bar{X}_{\cdot b}$ and

$$\sum_{j=1}^b \sum_{i=1}^a \frac{(X_{ij} - \bar{X}_{\cdot j})^2}{\sigma^2}$$

are mutually stochastically independent and the latter is $\chi^2[b(a - 1)]$. Let $Z = \sum_1^b k_j\bar{X}_{\cdot j}$. Then Z is normal with mean $\sum_1^b k_j\mu_j$ and variance $\left(\sum_1^b k_j^2\right)\sigma^2/a$, and Z is stochastically independent of

$$V = \frac{1}{b(a - 1)} \sum_{j=1}^b \sum_{i=1}^a (X_{ij} - \bar{X}_{\cdot j})^2.$$

Hence the random variable

$$T = \frac{\sum_1^b k_j \bar{X}_{.j} - \sum_1^b k_j \mu_j}{\sqrt{\left(\sum_1^b k_i^2\right)\sigma^2/a}} \Big/ \sqrt{V/\sigma^2} = \frac{\sum_1^b k_j \bar{X}_{.j} - \sum_1^b k_j \mu_j}{\sqrt{\left(\sum_1^b k_j^2\right)V/a}}$$

has a t distribution with $b(a-1)$ degrees of freedom. A positive number c can be found in Table IV in the Appendix, for certain values of α, $0 < \alpha < 1$, such that $\Pr(-c \leq T \leq c) = 1 - \alpha$. It follows that the probability is $1 - \alpha$ that

$$\sum_1^b k_j \bar{X}_{.j} - c\sqrt{\left(\sum_1^b k_j^2\right)\frac{V}{a}} \leq \sum_1^b k_j \mu_j \leq \sum_1^b k_j \bar{X}_{.j} + c\sqrt{\left(\sum_1^b k_j^2\right)\frac{V}{a}}.$$

The experimental values of $\bar{X}_{.j}$, $j = 1, 2, \ldots, b$, and V will provide a $100(1-\alpha)$ per cent confidence interval for $\sum_1^b k_j \mu_j$.

It should be observed that the confidence interval for $\sum_1^b k_j \mu_j$ depends upon the particular choice of $k_1, k_2, \ldots, k_b$. It is conceivable that we may be interested in more than one linear function of $\mu_1, \mu_2, \ldots, \mu_b$, such as $\mu_2 - \mu_1$, $\mu_3 - (\mu_1 + \mu_2)/2$, or $\mu_1 + \cdots + \mu_b$. We can, of course, find for each $\sum_1^b k_j \mu_j$ a random interval that has a preassigned probability of including that particular $\sum_1^b k_j \mu_j$. But how can we compute the probability that *simultaneously* these random intervals include their respective linear functions of $\mu_1, \mu_2, \ldots, \mu_b$? The following procedure of multiple comparisons, due to Scheffé, is one solution to this problem.

The random variable

$$\frac{\sum_{j=1}^b (\bar{X}_{.j} - \mu_j)^2}{\sigma^2/a}$$

is $\chi^2(b)$ and, because it is a function of $\bar{X}_{.1}, \ldots, \bar{X}_{.b}$ alone, it is stochastically independent of the random variable

$$V = \frac{1}{b(a-1)} \sum_{j=1}^b \sum_{i=1}^a (X_{ij} - \bar{X}_{.j})^2.$$

Hence the random variable

$$F = \frac{a\sum_{j=1}^b (\bar{X}_{.j} - \mu_j)^2/b}{V}$$

has an F distribution with b and $b(a-1)$ degrees of freedom. From Table V in the Appendix, for certain values of α, we can find a constant d such that $\Pr(F \leq d) = 1 - \alpha$ or

$$\Pr\left[\sum_{j=1}^{b} (\bar{X}_{.j} - \mu_j)^2 \leq (bd)\frac{V}{a}\right] = 1 - \alpha.$$

Note that $\sum_{j=1}^{b} (\bar{X}_{.j} - \mu_j)^2$ is the square of the distance, in b-dimensional space, from the point $(\mu_1, \mu_2, \ldots, \mu_b)$ to the random point $(\bar{X}_{.1}, \bar{X}_{.2}, \ldots, \bar{X}_{.b})$. Consider a space of dimension b and let $(t_1, t_2, \ldots, t_b)$ denote the coordinates of a point in that space. An equation of a hyperplane that passes through the point $(\mu_1, \mu_2, \ldots, \mu_b)$ is given by

$$k_1(t_1 - \mu_1) + k_2(t_2 - \mu_2) + \cdots + k_b(t_b - \mu_b) = 0, \tag{1}$$

where not all of the real numbers k_j, $j = 1, 2, \ldots, b$, are equal to zero. The square of the distance from this hyperplane to the point $(t_1 = \bar{X}_{.1}, t_2 = \bar{X}_{.2}, \ldots, t_b = \bar{X}_{.b})$ is

$$\frac{[k_1(\bar{X}_{.1} - \mu_1) + k_2(\bar{X}_{.2} - \mu_2) + \cdots + k_b(\bar{X}_{.b} - \mu_b)]^2}{k_1^2 + k_2^2 + \cdots + k_b^2}. \tag{2}$$

From the geometry of the situation, it follows that $\sum_{1}^{b} (\bar{X}_{.j} - \mu_j)^2$ is equal to the maximum of expression (2) with respect to $k_1, k_2, \ldots, k_b$. Thus the inequality $\sum_{1}^{b} (\bar{X}_{.j} - \mu_j)^2 \leq (bd)(V/a)$ holds if and only if

$$\frac{\left[\sum_{j=1}^{b} k_j(\bar{X}_{.j} - \mu_j)\right]^2}{\sum_{j=1}^{b} k_j^2} \leq (bd)\frac{V}{a}, \tag{3}$$

for every real $k_1, k_2, \ldots, k_b$, not all zero. Accordingly, these two equivalent events have the same probability $1 - \alpha$. However, Inequality (3) may be written in the form

$$\left|\sum_{1}^{b} k_j\bar{X}_{.j} - \sum_{1}^{b} k_j\mu_j\right| \leq \sqrt{bd\left(\sum_{1}^{b} k_j^2\right)\frac{V}{a}}.$$

Thus the probability is $1 - \alpha$ that simultaneously, for *all* real $k_1, k_2, \ldots, k_b$, not all zero,

$$\sum_{1}^{b} k_j\bar{X}_{.j} - \sqrt{bd\left(\sum_{1}^{b} k_j^2\right)\frac{V}{a}} \leq \sum_{1}^{b} k_j\mu_j \leq \sum_{1}^{b} k_j\bar{X}_{.j} + \sqrt{bd\left(\sum_{1}^{b} k_j^2\right)\frac{V}{a}}. \tag{4}$$

Denote by A the event where Inequality (4) is true for all real $k_1, \ldots, k_b$, and denote by B the event where that inequality is true

for a finite number of b-tuples $(k_1, \ldots, k_b)$. If the event A occurs, certainly the event B occurs. Hence $P(A) \leq P(B)$. In the applications, one is often interested only in a finite number of linear functions $\sum_1^b k_j\mu_j$. Once the experimental values are available, we obtain from (4) a confidence interval for each of these linear functions. Since $P(B) \geq P(A) = 1 - \alpha$, we have a confidence coefficient of at least $100(1 - \alpha)$ per cent that the linear functions are in these respective confidence intervals.

Remarks. If the sample sizes, say $a_1, a_2, \ldots, a_b$, are unequal, Inequality (4) becomes

$$(4') \quad \sum_1^b k_j\bar{X}_{\cdot j} - \sqrt{bd\sum_1^b \frac{k_j^2}{a_j} V} \leq \sum_1^b k_j\mu_j \leq \sum_1^b k_j\bar{X}_{\cdot j} + \sqrt{bd\sum_1^b \frac{k_j^2}{a_j} V},$$

where

$$\bar{X}_{\cdot j} = \frac{\sum_{i=1}^{a_j} X_{ij}}{a_j}, \qquad V = \frac{\sum_{j=1}^b \sum_{i=1}^{a_j} (X_{ij} - \bar{X}_{\cdot j})^2}{\sum_1^b (a_j - 1)},$$

and d is selected from Table V with b and $\sum_1^b (a_j - 1)$ degrees of freedom. Inequality (4') reduces to Inequality (4) when $a_1 = a_2 = \cdots = a_b$.

Moreover, if we restrict our attention to linear functions of the form $\sum_1^b k_j\mu_j$ with $\sum_1^b k_j = 0$ (such linear functions are called *contrasts*), the radical in Inequality (4') is replaced by

$$\sqrt{d(b-1)\sum_1^b \frac{k_j^2}{a_j} V},$$

where now d is found in Table V with $b - 1$ and $\sum_1^b (a_j - 1)$ degrees of freedom.

In these multiple comparisons, one often finds that the length of a confidence interval is much greater than the length of a $100(1 - \alpha)$ per cent confidence interval for a particular linear function $\sum_1^b k_j\mu_j$. But this is to be expected because in one case the probability $1 - \alpha$ applies to just one event and in the other it applies to the simultaneous occurrence of many events. One reasonable way to reduce the length of these intervals is to take a larger value of α, say 0.25, instead of 0.05. After all, it is still a very strong statement to say that the probability is 0.75 that *all* of these events occur.

EXERCISES

10.49. If $A_1, A_2, \ldots, A_k$ are events, prove, by induction, Boole's inequality $P(A_1 \cup A_2 \cup \cdots \cup A_k) \leq \sum_1^k P(A_i)$. Then show that

$$P(A_1^* \cap A_2^* \cap \cdots \cap A_k^*) \geq 1 - \sum_1^k P(A_i).$$

10.50. In the notation of this section, let $(k_{i1}, k_{i2}, \ldots, k_{ib})$, $i = 1, 2, \ldots, m$, represent a finite number of b-tuples. The problem is to find simultaneous confidence intervals for $\sum_{j=1}^b k_{ij}\mu_j$, $i = 1, 2, \ldots, m$, by a method different from that of Scheffé. Define the random variable T_i by

$$\left(\sum_{j=1}^b k_{ij}\bar{X}_{\cdot j} - \sum_{j=1}^b k_{ij}\mu_j\right) \Big/ \sqrt{\left(\sum_{j=1}^b k_{ij}^2\right) V/a}, \qquad i = 1, 2, \ldots, m.$$

(a) Let the event A_i^* be given by $-c_i \leq T_i \leq c_i$, $i = 1, 2, \ldots, m$. Find the random variables U_i and W_i such that $U_i \leq \sum_{j=1}^b k_{ij}\mu_j \leq W_i$ is equivalent to A_i^*.

(b) Select c_i such that $P(A_i^*) = 1 - \alpha/m$; that is, $P(A_i) = \alpha/m$. Use the results of Exercise 10.49 to determine a lower bound on the probability that simultaneously the random intervals $(U_1, W_1), \ldots, (U_m, W_m)$ include $\sum_{j=1}^b k_{1j}\mu_j, \ldots, \sum_{j=1}^b k_{mj}\mu_j$, respectively.

(c) Let $a = 3$, $b = 6$, and $\alpha = 0.05$. Consider the linear functions $\mu_1 - \mu_2$, $\mu_2 - \mu_3$, $\mu_3 - \mu_4$, $\mu_4 - (\mu_5 + \mu_6)/2$, and $(\mu_1 + \mu_2 + \cdots + \mu_6)/6$. Here $m = 5$. Show that the lengths of the confidence intervals given by the results of part (b) are shorter than the corresponding ones given by the method of Scheffé, as described in the text. If m becomes sufficiently large, however, this is not the case.

Chapter II

Nonparametric Methods

11.1 Confidence Intervals for Distribution Quantiles

We shall first define the concept of a quantile of a distribution of a random variable of the continuous type. Let X be a random variable of the continuous type with p.d.f. $f(x)$ and distribution function $F(x)$. Let p denote a positive proper fraction and assume that the equation $F(x) = p$ has a unique solution for x. This unique root is denoted by the symbol ξ_p and it is called the *quantile* (of the distribution) *of order* p. Thus, $\Pr(X \leq \xi_p) = F(\xi_p) = p$. For example, the quantile of order $\frac{1}{2}$ is the median of the distribution and $\Pr(X \leq \xi_{0.5}) = F(\xi_{0.5}) = \frac{1}{2}$.

In Chapter 6, we computed the probability that a certain random interval includes some special point. Frequently this special point was a parameter of the distribution of probability under consideration. Thus we were led to the notion of an interval estimate of a parameter. If the parameter happens to be a quantile of the distribution, and if we work with certain functions of the order statistics, it will be seen that this method of statistical inference is applicable to all distributions of the continuous type. We call these methods *distribution-free* or *nonparametric* methods of inference.

We now give an example which is illustrative of a theorem which is basic in this kind of statistical inference.

Example 1. Let the p.d.f. of a random variable X be given by

$$f(x) = 3x^2, \qquad 0 < x < 1,$$
$$= 0 \text{ elsewhere.}$$

Define the random variable Z by $Z = X^3$. It will be shown that the p.d.f. of Z is

$$\begin{aligned} h(z) &= 1, \qquad 0 < z < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

The transformation $z = x^3$ or $x = z^{1/3}$ maps the set $\{x; 0 < x < 1\}$ onto the set $\{z; 0 < z < 1\}$. The Jacobian of the transformation is

$$J = \frac{dx}{dz} = \frac{1}{3z^{2/3}}.$$

Thus the p.d.f. of z is

$$\begin{aligned} h(z) = f(z^{1/3})|J| &= 3z^{2/3}\frac{1}{3z^{2/3}} = 1, \qquad 0 < z < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

That is, $Z = X^3$ has a uniform distribution having this p.d.f. This result could have been obtained somewhat more easily by observing three things:

(i) $F(x) = x^3$ is the distribution function of X when $0 < x < 1$.
(ii) $z = x^3 = F(x)$, so $dz/dx = 3x^2 = f(x)$ when $0 < x < 1$.
(iii) $dx/dz = 1/(dz/dx)$ if $dz/dx \neq 0$.

It follows that here

$$f(x)\left|\frac{dx}{dz}\right| = f(x)\frac{dx}{dz} = \left(\frac{dz}{dx}\right)\left(\frac{1}{dz/dx}\right) = 1.$$

When x is replaced in the preceding expression by $z^{1/3}$, it is seen that $h(z) = 1$ on the set $\{z; 0 < z < 1\}$.

The preceding example is illustrative of the following theorem.

Theorem 1. *Let X be a random variable of the continuous type having a p.d.f. $f(x)$ and distribution function $F(x)$. Then the random variable $Z = F(X)$ has a uniform distribution with p.d.f.*

$$\begin{aligned} h(z) &= 1, \qquad 0 < z < 1, \\ &= 0 \textit{ elsewhere.} \end{aligned}$$

Proof. We prove the theorem under the assumption that the p.d.f. $f(x)$ is positive and continuous, provided $a < x < b$, and is zero elsewhere. The distribution function of X may be written

$$\begin{aligned} F(x) &= 0, \qquad x \leq a, \\ &= \int_a^x f(w)\,dw, \qquad a < x < b, \\ &= 1, \qquad b \leq x. \end{aligned}$$

The transformation $z = F(x)$ maps the set $\{x; a < x < b\}$ onto the set $\{z; 0 < z < 1\}$. Since $dz/dx = f(x)$, $a < x < b$, then

$$f(x)\left|\frac{dx}{dz}\right| = f(x)\frac{dx}{dz} = f(x)\frac{1}{dz/dx} = f(x)\frac{1}{f(x)} = 1, \qquad a < x < b.$$

That is, the p.d.f. of $Z = F(X)$ is

$$\begin{aligned} h(z) &= 1, \qquad 0 < z < 1, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

as was to be shown.

The random variable $Z = F(X)$ is an important random variable. Theorem 1 describes the distribution of Z. It is our purpose now to make an interpretation. Since $Z = F(X)$ has the p.d.f.

$$\begin{aligned} h(z) &= 1, \qquad 0 < z < 1, \\ &= 0 \text{ elsewhere,} \end{aligned}$$

then, if $0 < p < 1$, we have

$$\Pr[F(X) \leq p] = \int_0^p dz = p.$$

Now $F(x) = \Pr(X \leq x)$. Since $\Pr(X = x) = 0$, then $F(x)$ is the fractional part of the probability for the distribution of X which is between $-\infty$ and x. If $F(x) \leq p$, then no more than $100p$ per cent of the probability for the distribution of X is between $-\infty$ and x. But recall $\Pr[F(X) \leq p] = p$. That is, the probability that the random variable $Z = F(X)$ is less than or equal to p is precisely the probability that the random interval $(-\infty, X)$ contains no more than $100p$ per cent of the probability for the distribution. For example, the probability that the random interval $(-\infty, X)$ contains no more than 70 per cent of the probability for the distribution is 0.70; and the probability that the random interval $(-\infty, X)$ contains more than 70 per cent of the probability for the distribution is $1 - 0.70 = 0.30$.

We now consider certain functions of the order statistics. Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution that has a positive and continuous p.d.f. $f(x)$ if and only if $a < x < b$; and let $F(x)$ denote the associated distribution function. Consider the random variables $F(X_1), F(X_2), \ldots, F(X_n)$. These random variables are mutually stochastically independent and each, in accordance with Theorem 1, has a uniform distribution on the interval (0, 1). Thus $F(X_1), F(X_2), \ldots, F(X_n)$ is a random sample of size n from a uniform distribution on the interval (0, 1). Consider the order statistics of this

random sample $F(X_1), F(X_2), \ldots, F(X_n)$. Let Z_1 be the smallest of these $F(X_i)$, Z_2 the next $F(X_i)$ in order of magnitude, ..., and Z_n the largest $F(X_i)$. If $Y_1, Y_2, \ldots, Y_n$ are the order statistics of the initial random sample $X_1, X_2, \ldots, X_n$, the fact that $F(x)$ is a nondecreasing (here, strictly increasing) function of x implies that $Z_1 = F(Y_1)$, $Z_2 = F(Y_2), \ldots, Z_n = F(Y_n)$. Thus, the joint p.d.f. of $Z_1, Z_2, \ldots, Z_n$ is given by

$$h(z_1, z_2, \ldots, z_n) = n!, \qquad 0 < z_1 < z_2 < \cdots < z_n < 1,$$
$$= 0 \text{ elsewhere.}$$

This proves a special case of the following theorem.

Theorem 2. *Let $Y_1, Y_2, \ldots, Y_n$ denote the order statistics of a random sample of size n from a distribution of the continuous type that has p.d.f. $f(x)$ and distribution function $F(x)$. The joint p.d.f. of the random variables $Z_i = F(Y_i)$, $i = 1, 2, \ldots, n$, is*

$$h(z_1, z_2, \ldots, z_n) = n!, \qquad 0 < z_1 < z_2 < \cdots < z_n < 1,$$
$$= 0 \textit{ elsewhere.}$$

Because the distribution function of $Z = F(X)$ is given by z, $0 < z < 1$, the marginal p.d.f. of $Z_k = F(Y_k)$ is the following beta p.d.f.:

$$(1) \quad h_k(z_k) = \frac{n!}{(k-1)!\,(n-k)!} z_k^{k-1}(1 - z_k)^{n-k}, \qquad 0 < z_k < 1,$$
$$= 0 \text{ elsewhere.}$$

Moreover, the joint p.d.f. of $Z_i = F(Y_i)$ and $Z_j = F(Y_j)$ is, with $i < j$, given by

$$(2) \quad h_{ij}(z_i, z_j) = \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!}$$
$$\times z_i^{i-1}(z_j - z_i)^{j-i-1}(1 - z_j)^{n-j}, \qquad 0 < z_i < z_j < 1,$$
$$= 0 \text{ elsewhere.}$$

Let $Y_1, Y_2, \ldots, Y_n$ denote the order statistics of a random sample of size n from a distribution of the continuous type with distribution function $F(x)$. We shall compute the probability that the kth order statistic is less than the quantile of order p. Now

$$\Pr(Y_k < \xi_p) = \Pr[F(Y_k) < F(\xi_p)]$$
$$= \Pr(Z_k < p),$$

since $F(\xi_p) = p$ and $Z_k = F(Y_k)$. The p.d.f. of Z_k is given by Equation (1) of this section. Accordingly,

$$(3) \qquad \Pr(Y_k < \xi_p) = \Pr(Z_k < p) = \int_0^p \frac{n!}{(k-1)!\,(n-k)!} z_k^{k-1}(1 - z_k)^{n-k}\, dz_k.$$

We shall show that the integral in the right-hand member of Equation (3) may be expressed in a form which is sometimes computationally more convenient. If we are to have $Y_k < \xi_p$, at least k items of the random sample must be less than ξ_p. Now $\Pr(X < \xi_p) = p$, where X is an item of the random sample. The sum

$$(4) \qquad \sum_{w=k}^{n} \frac{n!}{w!\,(n-w)!} p^w(1-p)^{n-w}$$

is precisely the probability of at least k "successes" throughout n independent trials, each trial with probability p of success. Thus, $\Pr(Y_k < \xi_p)$ is given by Display (4). Accordingly, the right-hand member of Equation (3) and Display (4) are equal. It is interesting to note that a "partial sum" of a binomial p.d.f. can be expressed exactly as the integral of a beta p.d.f. over a subset of $\{x; 0 < x < 1\}$.

Display (4) will be used to compute the probability that a certain random interval includes ξ_p, the quantile of order p. Take $Y_i < Y_j$. The event $Y_i < \xi_p$ is the union (exclusive of events of probability zero) of the two mutually exclusive events $Y_i < \xi_p < Y_j$ and $Y_i < Y_j < \xi_p$ (or $Y_j < \xi_p$). Consequently,

$$\Pr(Y_i < \xi_p) = \Pr(Y_i < \xi_p < Y_j) + \Pr(Y_j < \xi_p),$$

or

$$\Pr(Y_i < \xi_p < Y_j) = \Pr(Y_i < \xi_p) - \Pr(Y_j < \xi_p).$$

When particular values of n, i, and j are specified, each term of the right-hand member of the last equation can be computed by using either Equation (3) or Display (4). By this procedure, suppose it has been found that $\gamma = \Pr(Y_i < \xi_p < Y_j)$. Then the probability is γ that the random interval (Y_i, Y_j) includes the quantile of order p. If the experimental values of Y_i and Y_j are, respectively, y_i and y_j, then the interval (y_i, y_j) serves as a 100γ per cent confidence interval for ξ_p, the quantile of order p.

An illustrative example follows.

Example 2. Let $Y_1 < Y_2 < Y_3 < Y_4$ be the order statistics of a random sample of size 4 from a distribution of the continuous type. The probability

that the random interval (Y_1, Y_4) includes the median $\xi_{0.5}$ of the distribution will be computed. We have

$$\begin{aligned}\Pr(Y_1 < \xi_{0.5} < Y_4) &= \Pr(Y_1 < \xi_{0.5}) - \Pr(Y_4 < \xi_{0.5}) \\ &= \sum_{w=1}^{4} \frac{4!}{w!\,(4-w)!}\left(\frac{1}{2}\right)^4 - \sum_{w=4}^{4} \frac{4!}{w!\,(4-w)!}\left(\frac{1}{2}\right)^4 \\ &= \int_0^{0.5} \frac{4!}{0!\,3!}(1-z_1)^3\,dz_1 - \int_0^{0.5} \frac{4!}{3!\,0!} z_4^3\,dz_4 \\ &= \left[1 - \left(\frac{1}{2}\right)^4\right] - \left(\frac{1}{2}\right)^4 = 0.875.\end{aligned}$$

If Y_1 and Y_4 are observed to be $y_1 = 2.8$ and $y_4 = 4.2$, respectively, the interval (2.8, 4.2) is an 87.5 per cent confidence interval for the median $\xi_{0.5}$ of the distribution.

EXERCISES

11.1. Find the distribution of $Z = F(X)$ by computing the moment-generating function of Z, $M(t) = E(e^{tF(X)})$. Impose the same conditions of the proof of Theorem 1.

11.2. Let Y_n denote the nth order statistic of a random sample of size n from a distribution of the continuous type. Find the smallest value of n for which $\Pr(\xi_{0.9} < Y_n) \geq 0.75$.

11.3. Let $Y_1 < Y_2$ denote the order statistics of a random sample of size 2 from a distribution which is $n(\mu, \sigma^2)$, where σ^2 is known. (a) Show that $\Pr(Y_1 < \mu < Y_2) = \frac{1}{2}$ and compute the expected value of the random length $Y_2 - Y_1$. (b) If $\bar{X}$ is the mean of this sample, find the constant c such that $\Pr(\bar{X} - c\sigma < \mu < \bar{X} + c\sigma) = \frac{1}{2}$ and compare the length of this random interval with the expected value of that of part (a).

11.4. Let $Y_1 < Y_2 < Y_3 < Y_4 < Y_5$ denote the order statistics of a random sample of size 5 from a distribution of the continuous type. Compute (a) $\Pr(Y_1 < \xi_{0.5} < Y_5)$; (b) $\Pr(Y_1 < \xi_{0.25} < Y_3)$; (c) $\Pr(Y_4 < \xi_{0.80} < Y_5)$.

11.5. Compute $\Pr(Y_3 < \xi_{0.5} < Y_7)$ if $Y_1 < \cdots < Y_9$ are the order statistics of a random sample of size 9 from a distribution of the continuous type.

11.6. Find the smallest value of n for which $\Pr(Y_1 < \xi_{0.5} < Y_n) \geq 0.99$, where $Y_1 < \cdots < Y_n$ are the order statistics of a random sample of size n from a distribution of the continuous type.

11.2 Tolerance Limits for Distributions

We propose now to investigate a problem that has something of the same flavor as that treated in Section 11.1. Specifically, can we compute

the probability that a certain random interval includes (or *covers*) a preassigned percentage of the probability for the distribution that is under consideration? And, by appropriate selection of the random interval, can we be led to an additional distribution-free method of statistical inference?

On p. 350, after Theorem 1, it was remarked that the probability that the random variable $Z = F(X)$ is less than or equal to a positive proper fraction p is exactly the same as the probability that the random interval $(-\infty, X)$ contains no more than $100p$ per cent of the probability for the distribution of X. We now refer to Theorem 2. Consider the difference $Z_j - Z_i = F(Y_j) - F(Y_i)$, $i < j$. Now $F(y_j) = \Pr(X \leq y_j)$ and $F(y_i) = \Pr(X \leq y_i)$. Since $\Pr(X = y_i) = \Pr(X = y_j) = 0$, then the difference $F(y_j) - F(y_i)$ is that fractional part of the probability for the distribution of X that is between y_i and y_j. Let p denote a positive proper fraction. If $F(y_j) - F(y_i) \geq p$, then at least $100p$ per cent of the probability for the distribution of X is between y_i and y_j. Let it be given that $\gamma = \Pr[F(Y_j) - F(Y_i) \geq p]$. Then the random interval (Y_i, Y_j) has probability γ of containing at least $100p$ per cent of the probability for the distribution of X. If now y_i and y_j denote, respectively, experimental values of Y_i and Y_j, the interval (y_i, y_j) either does or does not contain at least $100p$ per cent of the probability for the distribution of X. However, we refer to the interval (y_i, y_j) as a 100γ per cent tolerance interval for $100p$ per cent of the probability for the distribution of X. In like vein, y_i and y_j are called 100γ per cent tolerance limits for $100p$ per cent of the probability for the distribution of X.

One way to compute the probability $\gamma = \Pr[F(Y_j) - F(Y_i) \geq p]$ is to use Equation (2) of Section 11.1, which gives the joint p.d.f. of $Z_i = F(Y_i)$ and $Z_j = F(Y_j)$. We have

$$h_{ij}(z_i, z_j) = \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!} z_i^{i-1}(z_j - z_i)^{j-i-1}(1 - z_j)^{n-j}$$

$$0 < z_i < z_j < 1,$$

$$= 0 \text{ elsewhere.}$$

The required probability is then given by

$$\gamma = \Pr(Z_j - Z_i \geq p) = \int_0^{1-p} \int_{p+z_i}^{1} h_{ij}(z_i, z_j)\, dz_j\, dz_i.$$

Sometimes this is a rather tedious computation. For this reason and for the reason that *coverages* are important in distribution-free statistical inference, we choose to introduce at this time the concept of a coverage.

Consider the random variables $W_1 = F(Y_1) = Z_1$, $W_2 = F(Y_2) - F(Y_1) = Z_2 - Z_1$, $W_3 = F(Y_3) - F(Y_2) = Z_3 - Z_2, \ldots, W_n = F(Y_n) - F(Y_{n-1}) = Z_n - Z_{n-1}$. The random variable W_1 is called a *coverage* of the random interval $\{x; -\infty < x < Y_1\}$ and the random variable W_i, $i = 2, 3, \ldots, n$, is called a *coverage* of the random interval $\{x; Y_{i-1} < x < Y_i\}$. We shall find the joint p.d.f. of the n coverages $W_1, W_2, \ldots, W_n$. First we note that the inverse functions of the associated transformation are given by

$$\begin{aligned} z_1 &= w_1, \\ z_2 &= w_1 + w_2, \\ z_3 &= w_1 + w_2 + w_3, \\ &\vdots \qquad\qquad \vdots \\ z_n &= w_1 + w_2 + w_3 + \cdots + w_n. \end{aligned}$$

We also note that the Jacobian is equal to one and that the space of positive probability density is

$$\{(w_1, w_2, \ldots, w_n); 0 < w_i, i = 1, 2, \ldots, n, w_1 + \cdots + w_n < 1\}.$$

Since the joint p.d.f. of $Z_1, Z_2, \ldots, Z_n$ is $n!$, $0 < z_1 < z_2 < \cdots < z_n < 1$, zero elsewhere, the joint p.d.f. of the n coverages is

$$\begin{aligned} k(w_1, \ldots, w_n) &= n!, \qquad 0 < w_i, i = 1, \ldots, n, w_1 + \cdots + w_n < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

A reexamination of Example 1 of Section 4.5 (p. 140) reveals that this is a Dirichlet p.d.f. with $k = n$ and $\alpha_1 = \alpha_2 = \cdots = \alpha_{n+1} = 1$.

Because the p.d.f. $k(w_1, \ldots, w_n)$ is symmetric in $w_1, w_2, \ldots, w_n$, it is evident that the distribution of every sum of r, $r < n$, of these coverages $W_1, \ldots, W_n$ is exactly the same for each fixed value of r. For instance, if $i < j$ and $r = j - i$, the distribution of $Z_j - Z_i = F(Y_j) - F(Y_i) = W_{i+1} + W_{i+2} + \cdots + W_j$ is exactly the same as that of $Z_{j-i} = F(Y_{j-i}) = W_1 + W_2 + \cdots + W_{j-i}$. But we know that the p.d.f. of Z_{j-i} is the beta p.d.f. of the form

$$\begin{aligned} h_{j-i}(v) &= \frac{\Gamma(n+1)}{\Gamma(j-i)\Gamma(n-j+i+1)} v^{j-i-1}(1-v)^{n-j+i}, \qquad 0 < v < 1, \\ &= 0 \text{ elsewhere.} \end{aligned}$$

Consequently, $F(Y_j) - F(Y_i)$ has this p.d.f. and

$$\Pr[F(Y_j) - F(Y_i) \geq p] = \int_p^1 h_{j-i}(v)\, dv.$$

Example 1. Let $Y_1 < Y_2 < \cdots < Y_6$ be the order statistics of a random sample of size 6 from a distribution of the continuous type. We want to use the observed interval (y_1, y_6) as a tolerance interval for 80 per cent of the distribution. Then

$$\begin{aligned}\gamma &= \Pr[F(Y_6) - F(Y_1) \geq 0.8] \\ &= 1 - \int_0^{0.8} 30v^4(1 - v)\, dv,\end{aligned}$$

because the integrand is the p.d.f. of $F(Y_6) - F(Y_1)$. Accordingly,

$$\gamma = 1 - 6(0.8)^5 + 5(0.8)^6 = 0.34,$$

approximately. That is, the observed values of Y_1 and Y_6 will define a 34 per cent tolerance interval for 80 per cent of the probability for the distribution.

Example 2. Each of the coverages W_i, $i = 1, 2, \ldots, n$, has the beta p.d.f.

$$\begin{aligned}k_1(w) &= n(1 - w)^{n-1}, \qquad 0 < w < 1, \\ &= 0 \text{ elsewhere,}\end{aligned}$$

because $W_1 = Z_1 = F(Y_1)$ has this p.d.f. Accordingly, the mathematical expectation of each W_i is

$$\int_0^1 nw(1 - w)^{n-1}\, dw = \frac{1}{n + 1}.$$

Now the coverage W_i can be thought of as the area under the graph of the p.d.f. $f(x)$, above the x-axis, and between the lines $x = Y_{i-1}$ and $x = Y_i$. (We take $Y_0 = -\infty$.) Thus the expected value of each of these random areas W_i, $i = 1, 2, \ldots, n$, is $1/(n + 1)$. That is, the order statistics partition the probability for the distribution into $n + 1$ parts and the expected value of each of these parts is $1/(n + 1)$. More generally, the expected value of $F(Y_j) - F(Y_i)$, $i < j$, is $(j - i)/(n + 1)$ since $F(Y_j) - F(Y_i)$ is the sum of $j - i$ of these coverages.

EXERCISES

11.7. Let Y_1 and Y_n be respectively the first and nth order statistics of a random sample of size n from a distribution of the continuous type having distribution function $F(x)$. Find the smallest value of n such that $\Pr[F(Y_n) - F(Y_1) \geq 0.5]$ is at least 0.95.

11.8. Let Y_2 and Y_{n-1} denote the second and the $(n - 1)$st order statistics of a random sample of size n from a distribution of the continuous type having distribution function $F(x)$. Compute $\Pr[F(Y_{n-1}) - F(Y_2) \geq p]$, where $0 < p < 1$.

11.9. Let $Y_1 < Y_2 < \cdots < Y_n$ be the order statistics of a random sample of size n from a distribution of the continuous type having distribution function $F(x)$. (a) What is the distribution of $U = 1 - F(Y_j)$? (b) Determine the distribution of $V = F(Y_n) - F(Y_j) + F(Y_i) - F(Y_1)$, where $i < j$.

11.3 The Sign Test

Some of the chi-square tests of Section 10.2 are illustrative of the type of tests which we investigate in the remainder of this chapter. Recall, in that section, we tested the hypothesis that the distribution of a certain random variable X is a specified distribution. We did this in the following manner. The space of X was partitioned into k mutually disjoint sets $A_1, A_2, \ldots, A_k$. The probability p_{i0} that $X \in A_i$ was computed under the assumption that the specified distribution is the correct distribution, $i = 1, 2, \ldots, k$. The original hypothesis was then replaced by the hypothesis

$$H_0: \Pr(X \in A_i) = p_{i0}, \qquad i = 1, 2, \ldots, k;$$

and a chi-square test, based upon a statistic which was denoted by Q_{k-1}, was used to test the hypothesis H_0 against all alternative hypotheses.

There is a certain subjective element in the use of this test, namely, the choice of k and of $A_1, A_2, \ldots, A_k$. But it is important to note that the limiting distribution of Q_{k-1}, under H_0, is $\chi^2(k-1)$; that is, the distribution of Q_{k-1} is *free* of $p_{10}, p_{20}, \ldots, p_{k0}$ and, accordingly, of the specified distribution of X. Here, and elsewhere, "under H_0" means when H_0 is true. A test of an hypothesis H_0 which is based upon a statistic whose distribution, under H_0, does not depend upon the specified distribution nor any parameters of that distribution is called a *distribution-free* or a *nonparametric* test.

Next, let $F(x)$ be the unknown distribution function of the random variable X. Let there be given two numbers ξ and p_0, where $0 < p_0 < 1$. We wish to test the hypothesis $H_0: F(\xi) = p_0$, that is, the hypothesis that $\xi = \xi_{p_0}$, the quantile of order p_0 of the distribution of X. We could use the statistic Q_{k-1}, with $k = 2$, to test H_0 against all alternatives. Suppose, however, that we are interested only in the alternative hypothesis, which is $H_1: F(\xi) > p_0$. One procedure is to base the test of H_0 against H_1 upon the random variable Y, which is the number of items less than or equal to ξ in a random sample of size n from the distribution. The statistic Y can be thought of as the number of "successes" throughout n independent trials. Then, if H_0 is true, Y is $b[n, p_0 = F(\xi)]$; whereas if H_0 is false, Y is $b[n, p = F(\xi)]$ whatever be the

distribution function $F(x)$. We reject H_0 and accept H_1 if and only if the observed value $y \geq c$, where c is an integer selected such that $\Pr(Y \geq c; H_0)$ is some reasonable significance level α. The power function of the test is given by

$$K(p) = \sum_{y=c}^{n} \binom{n}{y} p^y (1-p)^{n-y}, \qquad p_0 \leq p < 1,$$

where $p = F(\xi)$. In certain instances we may wish to approximate $K(p)$ by using an approximation to the binomial distribution.

Suppose that the alternative hypothesis to H_0: $F(\xi) = p_0$ is H_1: $F(\xi) < p_0$. Then the critical region is a set $\{y; y \leq c_1\}$. Finally, if the alternative hypothesis is H_1: $F(\xi) \neq p_0$, the critical region is a set $\{y; y \leq c_2 \text{ or } c_3 \leq y\}$.

Frequently, $p_0 = \frac{1}{2}$ and, in that case, the hypothesis is that the given number ξ is a median of the distribution. In the following example, this value of p_0 is used.

Example 1. Let $X_1, X_2, \ldots, X_{10}$ be a random sample of size 10 from a distribution with distribution function $F(x)$. We wish to test the hypothesis H_0: $F(72) = \frac{1}{2}$ against the alternative hypothesis H_1: $F(72) > \frac{1}{2}$. Let Y be the number of sample items which are less than or equal to 72. Let the observed value of Y be y, and let the test be defined by the critical region $\{y; y \geq 8\}$. The power function of the test is given by

$$K(p) = \sum_{y=8}^{10} \binom{10}{y} p^y (1-p)^{10-y}, \qquad \frac{1}{2} \leq p < 1,$$

where $p = F(72)$. In particular, the significance level is

$$\alpha = K(\tfrac{1}{2}) = \left[\binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right](\tfrac{1}{2})^{10} = \tfrac{7}{128}.$$

In many places in the literature the test which we have just described is called the *sign test.* The reason for this terminology is that the test is based upon a statistic Y which is equal to the number of nonpositive signs in the sequence $X_1 - \xi, X_2 - \xi, \ldots, X_n - \xi$. In the next section, a distribution-free test, which considers both the sign and the magnitude of each deviation $X_i - \xi$, is studied.

EXERCISES

11.10. Suggest a chi-square test of the hypothesis which states that a distribution is one of the beta type, with parameters $\alpha = 2$ and $\beta = 2$. Further, suppose that the test is to be based upon a random sample of size 100. In the solution, give k, define $A_1, A_2, \ldots, A_k$, and compute each p_{10}. If

possible, compare your proposal with those of other students. Are any of them the same?

11.11. Let $X_1, X_2, \ldots, X_{48}$ be a random sample of size 48 from a distribution that has the distribution function $F(x)$. To test H_0: $F(41) = \frac{1}{4}$ against H_1: $F(41) < \frac{1}{4}$, use the statistic Y, which is the number of sample items less than or equal to 41. If the observed value of Y is $y \leq 7$, reject H_0 and accept H_1. If $p = F(41)$, find the power function $K(p)$, $0 < p \leq \frac{1}{4}$, of the test. Approximate $\alpha = K(\frac{1}{4})$.

11.12. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution of the discrete type. Let $p_1 = \Pr(X_i < \xi)$, $p_2 = \Pr(X_i = \xi) > 0$, and $p_3 = \Pr(X_i > \xi)$. Let Y_1 be equal to the number of sample items less than ξ, Y_2 the number equal to ξ, and Y_3 the number greater than ξ. (a) Find the chi-square test of the hypothesis that $p_1 = p_3$. (b) Show that the conditional distribution of Y_1, given $Y_2 = y_2$, is one of the binomial type. Propose a test of the hypothesis that $p_1 = p_3$ which is based upon this conditional distribution.

11.4 A Test of Wilcoxon

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with distribution function $F(x)$. We have considered a test of the hypothesis $F(\xi) = \frac{1}{2}$, ξ given, which is based upon the signs of the deviations $X_1 - \xi, X_2 - \xi, \ldots, X_n - \xi$. In this section, a statistic is studied that takes into account, not only these signs, but also the magnitudes of the deviations.

In order that we can find such a statistic which is distribution-free, we must make two additional assumptions:

(a) $F(x)$ is the distribution function of a continuous type of random variable X;

(b) the p.d.f. $f(x)$ of X has a graph which is symmetric about the vertical axis through $\xi_{0.5}$, the median (which we assume to be unique) of the distribution.

Thus,

$$F(\xi_{0.5} - x) = 1 - F(\xi_{0.5} + x)$$

and

$$f(\xi_{0.5} - x) = f(\xi_{0.5} + x),$$

for all x. Moreover, the probability that any two items of a random sample are equal is zero, and in our discussion we shall assume that no two are equal.

The problem is to test the hypothesis that the median $\xi_{0.5}$ of the distribution is equal to some fixed number, say ξ. Thus we may, in all cases and without loss of generality, take $\xi = 0$. The reason for this is that if $\xi \neq 0$, then the fixed ξ can be subtracted from each sample item and the resulting variables can be used to test the hypothesis that their underlying distribution is symmetric about zero. Hence our conditions on $F(x)$ and $f(x)$ become $F(-x) = 1 - F(x)$ and $f(-x) = f(x)$, respectively.

To test the hypothesis H_0: $F(0) = \frac{1}{2}$, we proceed by first ranking $X_1, X_2, \ldots, X_n$ according to magnitude, disregarding their algebraic signs. Let R_i be the rank of $|X_i|$ among $|X_1|, |X_2|, \ldots, |X_n|$, $i = 1, 2, \ldots, n$. For example, if $n = 3$ and if we have $|X_2| < |X_3| < |X_1|$, then $R_1 = 3$, $R_2 = 1$, and $R_3 = 2$. Thus, $R_1, R_2, \ldots, R_n$ is an arrangement of the first n positive integers $1, 2, \ldots, n$. Further let Z_i, $i = 1, 2, \ldots, n$, be defined by

$$\begin{aligned} Z_i &= -1, \qquad \text{if } X_i < 0, \\ &= 1, \qquad \text{if } X_i > 0. \end{aligned}$$

If we recall that $\Pr(X_i = 0) = 0$, we see that it does not change the probabilities whether we associate $Z_i = 1$ or $Z_i = -1$ with the outcome $X_i = 0$.

The statistic $W = \sum_{i=1}^{n} Z_i R_i$ is the Wilcoxon statistic. Note that in computing this statistic we simply associate the sign of each X_i with the rank of its absolute value and sum the resulting n products.

If the alternative to the hypothesis H_0: $\xi_{0.5} = 0$ is H_1: $\xi_{0.5} > 0$, we reject H_0 if the observed value of W is an element of the set $\{w; w \geq c\}$. This is due to the fact that large positive values of W indicate that most of the large deviations from zero are positive. For alternatives $\xi_{0.5} < 0$ and $\xi_{0.5} \neq 0$ the critical regions are, respectively, the sets $\{w; w \leq c_1\}$ and $\{w; w \leq c_2 \text{ or } w \geq c_3\}$. To compute probabilities like $\Pr(W \geq c; H_0)$ we need to determine the distribution of W, under H_0.

To help us find the distribution of W, when H_0: $F(0) = \frac{1}{2}$ is true, we note the following facts:

(a) The assumption that $f(x) = f(-x)$ ensures that $\Pr[X_i < 0] = \Pr[X_i > 0] = \frac{1}{2}$, $i = 1, 2, \ldots, n$.

(b) Now $Z_i = -1$ if $X_i < 0$ and $Z_i = 1$ if $X_i > 0$, $i = 1, 2, \ldots, n$. Hence we have $\Pr(Z_i = -1) = \Pr(Z_i = 1) = \frac{1}{2}$, $i = 1, 2, \ldots, n$. Moreover, $Z_1, Z_2, \ldots, Z_n$ are mutually stochastically independent because $X_1, X_2, \ldots, X_n$ are mutually stochastically independent.

(c) A sum W is made up of the numbers $1, 2, \ldots, n$, each number with either a positive or a negative sign.

The preceding observations enable us to say that $W = \sum_1^n Z_i R_i$ has the same distribution as the random variable $V = \sum_1^n V_i$, where $V_1, V_2, \ldots, V_n$ are mutually stochastically independent and

$$\Pr(V_i = i) = \Pr(V_i = -i) = \tfrac{1}{2},$$

$i = 1, 2, \ldots, n$. That $V_1, V_2, \ldots, V_n$ are mutually stochastically independent follows from the fact that $Z_1, Z_2, \ldots, Z_n$ have that property; that is, the numbers $1, 2, \ldots, n$ always appear in a sum W and those numbers receive their algebraic signs by independent assignment. Thus each of $V_1, V_2, \ldots, V_n$ is like one and only one of $Z_1R_1, Z_2R_2, \ldots, Z_nR_n$.

Since W and V have the same distribution, the moment-generating function of W is that of V, namely,

$$M(t) = E\left\{\exp\left[t\sum_1^n V_i\right]\right\} = \prod_{i=1}^n E[e^{tV_i}]$$
$$= \prod_{i=1}^n \left(\frac{e^{-it} + e^{it}}{2}\right).$$

We can express $M(t)$ as the sum of terms of the form $(a_j/2^n)e^{b_jt}$. When $M(t)$ is written in this manner, we can determine by inspection the p.d.f. of the discrete-type random variable W. For example, the smallest value of W is found from the term $(1/2^n)e^{-t}e^{-2t}\cdots e^{-nt} = (1/2^n)e^{-n(n+1)t/2}$ and it is $-n(n+1)/2$. The probability of this value of W is the coefficient $1/2^n$. To make these statements more concrete, take $n = 3$. Then

$$M(t) = \left(\frac{e^{-t} + e^{t}}{2}\right)\left(\frac{e^{-2t} + e^{2t}}{2}\right)\left(\frac{e^{-3t} + e^{3t}}{2}\right)$$
$$= (\tfrac{1}{8})(e^{-6t} + e^{-4t} + e^{-2t} + 2 + e^{2t} + e^{4t} + e^{6t}).$$

Thus the p.d.f. of W, for $n = 3$, is given by

$$g(w) = \tfrac{1}{8}, \qquad w = -6, -4, -2, 2, 4, 6,$$
$$= \tfrac{2}{8}, \qquad w = 0,$$
$$= 0 \text{ elsewhere.}$$

The mean and the variance of W are more easily computed directly than by working with the moment-generating function $M(t)$. Because $V = \sum_1^n V_i$ and $W = \sum_1^n Z_i R_i$ have the same distribution, they have the

same mean and the same variance. When the hypothesis H_0: $F(0) = \frac{1}{2}$ is true, it is easy to determine the values of these two characteristics of the distribution of W. Since $E(V_i) = 0$, $i = 1, 2, \ldots, n$, we have

$$\mu_W = E(W) = \sum_1^n E(V_i) = 0.$$

The variance of V_i is $(-i)^2(\frac{1}{2}) + (i)^2(\frac{1}{2}) = i^2$. Thus the variance of W is

$$\sigma_W^2 = \sum_1^n i^2 = \frac{n(n+1)(2n+1)}{6}.$$

For large values of n, the determination of the exact distribution of W becomes tedious. Accordingly, one looks for an approximating distribution. Although W is distributed as is the sum of n random variables that are mutually stochastically independent, our form of the central limit theorem cannot be applied because the n random variables do not have identical distributions. However, a more general theorem, due to Liapounov, states that if U_i has mean μ_i and variance σ_i^2, $i = 1, 2, \ldots, n$, if $U_1, U_2, \ldots, U_n$ are mutually stochastically independent, if $E(|U_i - \mu_i|^3)$ is finite for every i, and if

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n E(|U_i - \mu_i|^3)}{\left(\sum_{i=1}^n \sigma_i^2\right)^{3/2}} = 0,$$

then

$$\frac{\sum_{i=1}^n U_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}$$

has a limiting distribution which is $n(0, 1)$. For our variables $V_1, V_2, \ldots, V_n$ we have

$$E(|V_i - \mu_i|^3) = i^3(\tfrac{1}{2}) + i^3(\tfrac{1}{2}) = i^3;$$

and it is known that

$$\sum_{i=1}^n i^3 = \frac{n^2(n+1)^2}{4}.$$

Now

$$\lim_{n\to\infty} \frac{n^2(n+1)^2/4}{[n(n+1)(2n+1)/6]^{3/2}} = 0$$

because the numerator is of order n^4 and the denominator is of order $n^{9/2}$. Thus,

$$\frac{W}{\sqrt{n(n+1)(2n+1)/6}}$$

is approximately $n(0, 1)$ when H_0 is true. This allows us to approximate probabilities like $\Pr(W \geq c; H_0)$ when the sample size n is large.

Example 1. Let $\xi_{0.5}$ be the median of a symmetric distribution which is of the continuous type. To test, with $\alpha = 0.01$, the hypothesis H_0: $\xi_{0.5} = 75$ against H_1: $\xi_{0.5} > 75$ we observed a random sample of size $n = 18$. Let it be given that the deviations of these 18 values from 75 are the following numbers:

$$1.5, -0.5, 1.6, 0.4, 2.3, -0.8, 3.2, 0.9, 2.9,$$
$$0.3, 1.8, -0.1, 1.2, 2.5, 0.6, -0.7, 1.9, 1.3.$$

The experimental value of the Wilcoxon statistic is equal to

$$w = 11 - 4 + 12 + 3 + 15 - 7 + 18 + 8 + 17 + 2 + 13 - 1 + 9 + 16 + 5 - 6 + 14 + 10 = 135.$$

Since, with $n = 18$ so that $\sqrt{n(n+1)(2n+1)/6} = 45.92$, we have that

$$0.01 = \Pr\left(\frac{W}{45.92} \geq 2.326\right) = \Pr(W \geq 106.8).$$

Because $w = 135 > 106.8$, we reject H_0 at the approximate 0.01 significance level.

There are many modifications and generalizations of the Wilcoxon statistic. One generalization is the following: Let $c_1 \leq c_2 \leq \cdots \leq c_n$ be nonnegative numbers. Then, in the Wilcoxon statistic, replace the ranks $1, 2, \ldots, n$ by $c_1, c_2, \ldots, c_n$, respectively. For example, if $n = 3$ and if we have $|X_2| < |X_3| < |X_1|$, then $R_1 = 3$ is replaced by c_3, $R_2 = 1$ by c_1, and $R_3 = 2$ by c_2. In this example the generalized statistic is given by $Z_1c_3 + Z_2c_1 + Z_3c_2$. Similar to the Wilcoxon statistic, this generalized statistic is distributed under H_0, as is the sum of n stochastically independent random variables, the ith of which takes each of the values $c_i \neq 0$ and $-c_i$ with probability $\frac{1}{2}$; if $c_i = 0$, that variable takes the value $c_i = 0$ with probability one. Some special cases of this statistic are proposed in the Exercises.

EXERCISES

11.13. The observed values of a random sample of size 10 from a distribution which is symmetric about $\xi_{0.5}$ are 10.2, 14.1, 9.2, 11.3, 7.2, 9.8, 6.5, 11.8, 8.7, 10.8. Use Wilcoxon's statistic to test the hypothesis H_0: $\xi_{0.5} = 8$ against H_1: $\xi_{0.5} > 8$ if $\alpha = 0.05$. Even though n is small, use the normal approximation.

11.14. Find the distribution of W for $n = 4$ and $n = 5$. *Hint.* Multiply the moment-generating function of W, with $n = 3$, by $(e^{-4t} + e^{4t})/2$ to get that of W, with $n = 4$.

11.15. Let $X_1, X_2, \ldots, X_n$ be mutually stochastically independent. If the p.d.f. of X_i is uniform over the interval $(-2^{1-i}, 2^{1-i})$, $i = 1, 2, 3, \ldots$, show that Liapounov's condition is not satisfied. The sum $\sum_{i=1}^{n} X_i$ does not have an approximate normal distribution because the first random variables in the sum tend to dominate it.

11.16. If $n = 4$ and, in the notation of the text, $c_1 = 1$, $c_2 = 2$, $c_3 = c_4 = 3$, find the distribution of the generalization of the Wilcoxon statistic, say W_g. For a general n, find the mean and the variance of W_g if $c_i = i$, $i \leq n/2$, and $c_i = [n/2] + 1, i > n/2$, where $[z]$ is the greatest integer function. Does Liapounov's condition hold here?

11.17. A modification of Wilcoxon's statistic which is frequently used is achieved by replacing R_i by $R_i - 1$; that is, use the modification $W_m = \sum_{1}^{n} Z_i(R_i - 1)$. Show that $W_m/\sqrt{(n-1)n(2n-1)/6}$ has a limiting distribution which is $n(0, 1)$.

11.18. If, in the discussion of the generalization of the Wilcoxon statistic, we let $c_1 = c_2 = \cdots = c_n = 1$, show that we obtain a statistic equivalent to that used in the sign test.

11.19. If $c_1, c_2, \ldots, c_n$ are selected so that $i/(n+1) = \int_0^{c_i} \sqrt{2/\pi}\, e^{-x^2/2}\, dx$, $i = 1, 2, \ldots, n$, the generalized Wilcoxon W_g is an example of a *normal scores* statistic. If $n = 9$, compute the mean and the variance of this W_g.

11.20. If $c_i = 2^i$, $i = 1, 2, \ldots, n$, the corresponding W_g is called the *binary* statistic. Find the mean and the variance of this W_g. Is Liapounov's condition satisfied?

11.21. In the definition of Wilcoxon's statistic, let W_1 be the sum of the ranks of those items of the sample which are positive and let W_2 be the sum of the ranks of those items which are negative. Then $W = W_1 - W_2$. (a) Show that $W = 2W_1 - n(n+1)/2$ and $W = n(n+1)/2 - 2W_2$. (b) Compute the mean and the variance of each of W_1 and W_2.

11.5 The Equality of Two Distributions

In Sections 11.3 and 11.4, some tests of hypotheses about one distribution were investigated. In this section, as in the next section, various tests of the equality of two independent distributions are studied. By the equality of two distributions, we mean that the two distribution functions, say F and G, have $F(z) = G(z)$ for all values of z.

The first test that we discuss is a natural extension of the chi-square test. Let X and Y be stochastically independent variables with distribution functions $F(x)$ and $G(y)$, respectively. We wish to test the hypothesis that $F(z) = G(z)$, for all z. Let us partition the real line into k mutually disjoint sets $A_1, A_2, \ldots, A_k$. Define

$$p_{i1} = \Pr(X \in A_i), \qquad i = 1, 2, \ldots, k,$$

and

$$p_{i2} = \Pr(Y \in A_i), \qquad i = 1, 2, \ldots, k.$$

If $F(z) = G(z)$, for all z, then $p_{i1} = p_{i2}$, $i = 1, 2, \ldots, k$. Accordingly, the hypothesis that $F(z) = G(z)$, for all z, is replaced by the less restrictive hypothesis

$$H_0: p_{i1} = p_{i2}, \qquad i = 1, 2, \ldots, k.$$

But this is exactly the problem of testing the equality of two independent multinomial distributions which was considered in Example 3, p. 312, and the reader is referred to that page for the details.

Some statisticians prefer a procedure which eliminates some of the subjectivity of selecting the partitions. For a fixed positive integer k, proceed as follows. Consider a random sample of size m from the distribution of X and a random sample of size n from the independent distribution of Y. Let the experimental values be denoted by $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$. Then combine the two samples into one sample of size $m + n$ and order the $m + n$ values (not their absolute values) in ascending order of magnitude. These ordered items are then partitioned into k parts in such a way that each part has the same number of items. (If the sample sizes are such that this is impossible, a partition with approximately the same number of items in each group suffices.) In effect, then, the partition $A_1, A_2, \ldots, A_k$ is determined by the experimental values themselves. This does not alter the fact that the statistic, discussed in Example 3, p. 312, has a limiting distribution

which is $\chi^2(k-1)$. Accordingly, the procedures used in that example may be used here.

Among the tests of this type there is one which is frequently used. It is essentially a test of the equality of the medians of two independent distributions. To simplify the discussion, we assume that $m+n$, the size of the combined sample, is an even number, say $m+n=2h$, where h is a positive integer. We take $k=2$ and the combined sample of size $m+n=2h$, which has been ordered, is separated into two parts, a "lower half" and an "upper half," each containing $h=(m+n)/2$ of the experimental values of X and Y. The statistic, suggested by Example 3, p. 312, could be used because it has, when H_0 is true, a limiting distribution which is $\chi^2(1)$. However, it is more interesting to find the exact distribution of another statistic which enables us to test the hypothesis H_0 against the alternative H_1: $F(z) \geq G(z)$ or against the alternative H_1: $F(z) \leq G(z)$ as opposed to merely $F(z) \neq G(z)$. [Here, and in the sequel, alternatives like $F(z) \geq G(z)$ and $F(z) \leq G(z)$ and $F(z) \neq G(z)$ mean that strict inequality holds on some set of positive probability measure.] This other statistic is V, which is the number of observed values of X that are in the lower half of the combined sample. If the observed value of V is quite large, one might suspect that the median of the distribution of X is smaller than that of the distribution of Y. Thus the critical region of this test of the hypothesis H_0: $F(z)=G(z)$, for all z, against H_1: $F(z) \geq G(z)$ is of the form $V \geq c$. Because our combined sample is of even size, there is no unique median of the sample. However, one can arbitrarily insert a number between the hth and $(h+1)$st ordered items and call it the median of the sample. On this account, a test of the sort just described is called a *median* test. Incidentally, if the alternative hypothesis is H_1: $F(z) \leq G(z)$, the critical region is of the form $V \leq c$.

The distribution of V is quite easy to find if the distribution functions $F(x)$ and $G(y)$ are of the continuous type and if $F(z)=G(z)$, for all z. We shall now show that V has a hypergeometric p.d.f. Let $m+n=2h$, h a positive integer. To compute $\Pr(V=v)$, we need the probability that exactly v of $X_1, X_2, \ldots, X_m$ are in the lower half of the ordered combined sample. Under our assumptions, the probability is zero that any two of the $2h$ random variables are equal. The smallest h of the $m+n=2h$ items can be selected in any one of $\binom{2h}{h}$ ways. Each of these ways has the same probability. Of these $\binom{2h}{h}$ ways, we need to count the number of those in which exactly v of the m values of

X (and hence $h - v$ of the n values of Y) appear in the lower h items. But this is $\binom{m}{v}\binom{n}{h-v}$. Thus the p.d.f. of V is the hypergeometric p.d.f.

$$k(v) = \Pr(V = v) = \frac{\binom{m}{v}\binom{n}{h-v}}{\binom{m+n}{h}}, \qquad v = 0, 1, 2, \ldots, m,$$

$$= 0 \text{ elsewhere},$$

where $m + n = 2h$.

The student may be momentarily puzzled by the meaning of $\binom{n}{h-v}$ for $v = 0, 1, 2, \ldots, m$. For example, let $m = 17$, $n = 3$, so that $h = 10$. Then we have $\binom{3}{10-v}$, $v = 0, 1, \ldots, 17$. However, we take $\binom{n}{h-v}$ to be zero if $h - v$ is negative or if $h - v > n$.

If $m + n$ is an odd number, say $m + n = 2h + 1$, it is left to the student to show that the p.d.f. $k(v)$ gives the probability that exactly v of the m values of X are among the lower h of the combined $2h + 1$ values; that is, exactly v of the m values of X are less than the median of the combined sample.

If the distribution functions $F(x)$ and $G(y)$ are of the continuous type, there is another rather simple test of the hypothesis that $F(z) = G(z)$, for all z. This test is based upon the notion of *runs* of values of X and of values of Y. We shall now explain what we mean by runs. Let us again combine the sample of m values of X and the sample of n values of Y into one collection of $m + n$ ordered items arranged in ascending order of magnitude. With $m = 7$ and $n = 8$ we might find that the 15 ordered items were in the arrangement

$$\underline{x}\ \underline{yyy}\ \underline{xx}\ \underline{y}\ \underline{x}\ \underline{yy}\ \underline{xxx}\ \underline{yy}.$$

Note that in this ordering we have underscored the groups of successive values of the random variable X and those of the random variable Y. If we read from left to right, we would say that we have a *run* of one value of X, followed by a *run* of three values of Y, followed by a *run* of two values of X, and so on. In our example, there is a total of eight runs. Three are runs of length one, three are runs of length two, and two are runs of length three. Note that the total number of runs is always one more than the number of unlike adjacent symbols.

Of what can runs be suggestive? Suppose, with $m = 7$ and $n = 8$ we have the following ordering:

$$\underline{xxxxx}\ \underline{y}\ \underline{xx}\ \underline{yyyyyyy}.$$

To us, this strongly suggests that $F(z) > G(z)$. For if, in fact, $F(z) = G(z)$ for all z, we would anticipate a greater number of runs. And if the first run of five values of X were interchanged with the last run of seven values of Y, this would suggest that $F(z) < G(z)$. But runs can be suggestive of other things. For example, with $m = 7$ and $n = 8$, consider the runs

$$\underline{yyyy}\ \underline{xxxxxxx}\ \underline{yyyy}.$$

This suggests to us that the medians of the distributions of X and Y may very well be about the same, but that the "spread" (measured possibly by the standard deviation) of the distribution of X is considerably less than that of the distribution of Y.

Let the random variable R equal the number of runs in the combined sample, once the combined sample has been ordered. Because our random variables X and Y are of the continuous type, we may assume that no two of these sample items are equal. We wish to find the p.d.f. of R. To find this distribution, when $F(z) = G(z)$, we shall suppose that all arrangements of the m values of X and the n values of Y have equal probabilities. We shall show that

$$\Pr(R = 2k + 1) = \left\{\binom{m-1}{k}\binom{n-1}{k-1} + \binom{m-1}{k-1}\binom{n-1}{k}\right\}\Big/\binom{m+n}{m}$$

(1)

$$\Pr(R = 2k) = 2\binom{m-1}{k-1}\binom{n-1}{k-1}\Big/\binom{m+n}{m}$$

when $2k$ and $2k + 1$ are elements of the space of R.

To prove Formulas (1), note that we can select the m positions for the m values of X from the $m + n$ positions in any one of $\binom{m+n}{m}$ ways. Since each of these choices yields one arrangement, the probability of each arrangement is equal to $1\Big/\binom{m+n}{m}$. The problem is now to determine how many of these arrangements yield $R = r$, where r is an integer in the space of R. First, let $r = 2k + 1$, where k is a positive integer. This means that there must be $k + 1$ runs of the ordered values

of X and k runs of the ordered values of Y or vice versa. Consider first the number of ways of obtaining $k + 1$ runs of the m values of X. We can form $k + 1$ of these runs by inserting k "dividers" into the $m - 1$ spaces between the values of X, with no more than one divider per space. This can be done in any one of $\binom{m-1}{k}$ ways. Similarly, we can construct k runs of the n values of Y by inserting $k - 1$ dividers into the $n - 1$ spaces between the values of Y, with no more than one divider per space. This can be done in any one of $\binom{n-1}{k-1}$ ways. The joint operation can be performed in any one of $\binom{m-1}{k}\binom{n-1}{k-1}$ ways. These two sets of runs can be placed together to form $r = 2k + 1$ runs. But we could also have k runs of the values of X and $k + 1$ runs of the values of Y. An argument similar to the preceding shows that this can be effected in any one of $\binom{m-1}{k-1}\binom{n-1}{k}$ ways. Thus,

$$\Pr(R = 2k + 1) = \frac{\binom{m-1}{k}\binom{n-1}{k-1} + \binom{m-1}{k-1}\binom{n-1}{k}}{\binom{m+n}{m}},$$

which is the first of Formulas (1).

If $r = 2k$, where k is a positive integer, we see that the ordered values of X and the ordered values of Y must each be separated into k runs. These operations can be performed in any one of $\binom{m-1}{k-1}$ and $\binom{n-1}{k-1}$ ways, respectively. These two sets of runs can be placed together to form $r = 2k$ runs. But we may begin with either a run of values of X or a run of values of Y. Accordingly, the probability of $2k$ runs is

$$\Pr(R = 2k) = \frac{2\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{m+n}{m}},$$

which is the second of Formulas (1).

If the critical region of this *run* test of the hypothesis H_0: $F(z) = G(z)$ for all z, is of the form $R \leq c$, it is easy to compute $\alpha = \Pr(R \leq c; H_0)$, provided m and n are small. Although it is not easy to

show, the distribution of R can be approximated, with large sample sizes m and n, by a normal distribution with mean

$$\mu = E(R) = 2\frac{mn}{m+n} + 1$$

and variance

$$\sigma^2 = \frac{(\mu - 1)(\mu - 2)}{m + n - 1}.$$

The *run* test may also be used to test for *randomness*. That is, it can be used as a check to see if it is reasonable to treat $X_1, X_2, \ldots, X_s$ as a random sample of size s from some continuous distribution. To facilitate the discussion, take s to be even. We are given the s values of X to be $x_1, x_2, \ldots, x_s$, which are not ordered by magnitude but by the order in which they were observed. However, there are $s/2$ of these values, each of which is smaller than the remaining $s/2$ values. Thus we have a "lower half" and an "upper half" of these values. In the sequence $x_1, x_2, \ldots, x_s$, replace each value of X which is in the lower half by the letter L and each value in the upper half by the letter U. Then, for example, with $s = 10$, a sequence such as

$$L\ L\ L\ L\ U\ L\ U\ U\ U\ U$$

may suggest a trend toward increasing values of X; that is, these values of X may not reasonably be looked upon as being the items of a random sample. If trend is the only alternative to randomness, we can make a test based upon R and reject the hypothesis of randomness if $R \leq c$. To make this test, we would use the p.d.f. of R, as given on p. 368, with $m = n = s/2$. On the other hand if, with $s = 10$, we find a sequence such as

$$L\ H\ L\ H\ L\ H\ L\ H\ L\ H,$$

our suspicions are aroused that there may be a nonrandom effect which is cyclic even though $R = 10$. Accordingly, to test for a trend or a cyclic effect, we could use a critical region of the form $R \leq c_1$ or $R \geq c_2$.

If the sample size s is odd, the number of sample items in the "upper half" and the number in the "lower half" will differ by one. Then, for example, we could use the p.d.f. of R, given on p. 368, with $m = (s - 1)/2$ and $n = (s + 1)/2$ or vice versa.

EXERCISES

11.22. Let 3.1, 5.6, 4.7, 3.8, 4.2, 3.0, 5.1, 3.9, 4.8 and 5.3, 4.0, 4.9, 6.2, 3.7, 5.0, 6.5, 4.5, 5.5, 5.9, 4.4, 5.8 be observed samples of sizes $m = 9$ and $n = 12$ from two independent distributions. With $k = 3$, use a chi-square test to test, with $\alpha = 0.05$ approximately, the equality of the two distributions.

11.23. In the median test, with $m = 9$ and $n = 7$, find the p.d.f. of the random variable V, the number of values of X in the lower half of the combined sample. In particular, what are the values of the probabilities Pr $(V = 0)$ and Pr $(V = 9)$?

11.24. In the notation of the text, use the median test and the data given in Exercise 11.22 to test, with $\alpha = 0.05$, approximately, the hypothesis of the equality of the two independent distributions against the alternative hypothesis that $F(z) \geq G(z)$. If the exact probabilities are too difficult to determine for $m = 9$ and $n = 12$, approximate these probabilities.

11.25. In the discussion of the run test, let the random variables R_1 and R_2 be, respectively, the number of runs of the values of X and the number of runs of the values of Y. Then $R = R_1 + R_2$. Let the pair (r_1, r_2) of integers be in the space of (R_1, R_2); then $|r_1 - r_2| \leq 1$. Show that the joint p.d.f. of R_1 and R_2 is $2\binom{m-1}{r_1-1}\binom{n-1}{r_2-1}\Big/\binom{m+n}{m}$, if $r_1 = r_2$; that this joint p.d.f. is $\binom{m-1}{r_1-1}\binom{n-1}{r_2-1}\Big/\binom{m+n}{m}$ if $|r_1 - r_2| = 1$; and is zero elsewhere. Show that the marginal p.d.f. of R_1 is $\binom{m-1}{r_1-1}\binom{n+1}{r_1}\Big/\binom{m+n}{m}$, $r_1 = 1, \ldots, m$, and is zero elsewhere. Find $E(R_1)$. In a similar manner, find $E(R_2)$. Compute $E(R) = E(R_1) + E(R_2)$.

11.6 The Mann-Whitney-Wilcoxon Test

We return to the problem of testing the equality of two independent distributions of the continuous type. Let X and Y be stochastically independent random variables of the continuous type. Let $F(x)$ and $G(y)$ denote, respectively, the distribution functions of X and Y and let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ denote independent samples from these distributions. We shall discuss the Mann-Whitney-Wilcoxon test of the hypothesis H_0: $F(z) = G(z)$ for all values of z.

Let us define

$$\begin{aligned} Z_{ij} &= 1, \quad X_i < Y_j, \\ &= 0, \quad X_i > Y_j, \end{aligned}$$

and consider the statistic

$$U = \sum_{j=1}^{n} \sum_{i=1}^{m} Z_{ij}.$$

We note that

$$\sum_{i=1}^{m} Z_{ij}$$

counts the number of values of X which are less than Y_j, $j = 1, 2, \ldots, n$. Thus U is the sum of these n counts. For example, with $m = 4$ and $n = 3$, consider the observations

$$x_2 < y_3 < x_1 < x_4 < y_1 < x_3 < y_2.$$

There are 3 values of x which are less than y_1; there are 4 values of x which are less than y_2; and there is 1 value of x which is less than y_3. Thus the experimental value of U is $u = 3 + 4 + 1 = 8$.

Clearly, the smallest value which U can take is zero, and the largest value is mn. Thus, the space of U is $\{u; u = 0, 1, 2, \ldots, mn\}$. If U is large, the values of Y tend to be larger than the values of X, and this suggests that $F(z) \geq G(z)$ for all z. On the other hand, a small value of U suggests that $F(z) \leq G(z)$ for all z. Thus, if we test the hypothesis H_0: $F(z) = G(z)$ for all z against the alternative hypothesis H_1: $F(z) \geq G(z)$ for all z, the critical region is of the form $U \geq c_1$. If the alternative hypothesis is H_1: $F(z) \leq G(z)$ for all z, the critical region is of the form $U \leq c_2$. To determine the size of a critical region, we need the distribution of U when H_0 is true.

If u belongs to the space of U, let us denote $\Pr(U = u)$ by the symbol $h(u; m, n)$. This notation focuses attention on the sample sizes m and n. To determine the probability $h(u; m, n)$, we first note that we have $m + n$ positions to be filled by m values of X and n values of Y. We can fill m positions with the values of X in any one of $\binom{m+n}{m}$ ways. Once this has been done, the remaining n positions can be filled with the values of Y. When H_0 is true, each of these arrangements has the same probability, namely, $1/\binom{m+n}{m}$. The final right-hand position of an arrangement may be either a value of X or a value of Y. This position can be filled in any one of $m + n$ ways, m of which are favorable to X and n of which are favorable to Y. Accordingly, the probability that an arrangement ends with a value of X is $m/(m + n)$ and the probability that an arrangement terminates with a value of Y is $n/(m + n)$.

Now U can equal u in two mutually exclusive and exhaustive ways: (1) The final right-hand position (the largest of the $m + n$ values) in the arrangement may be a value of X and the remaining $(m - 1)$ values of X and the n values of Y can be arranged so as to have $U = u$. The probability that $U = u$, given an arrangement which terminates with a value of X, is given by $h(u; m - 1, n)$. Or, (2) the largest value in the arrangement can be a value of Y. This value of Y is greater than m values of X. If we are to have $U = u$, the sum of $n - 1$ counts of the m values of X with respect to the remaining $n - 1$ values of Y must be $u - m$. Thus, the probability that $U = u$, given an arrangement which terminates in a value of Y, is given by $h(u - m; m, n - 1)$. Accordingly, the probability that $U = u$ is

$$h(u; m, n) = \left(\frac{m}{m+n}\right)h(u; m-1, n) + \left(\frac{n}{m+n}\right)h(u-m; m, n-1).$$

We impose the following reasonable restrictions upon the function $h(u; m, n)$:

$$\begin{aligned} h(u; 0, n) &= 1, \qquad u = 0, \\ &= 0, \qquad u > 0, n \geq 1, \end{aligned}$$

and

$$\begin{aligned} h(u; m, 0) &= 1, \qquad u = 0, \\ &= 0, \qquad u > 0, m \geq 1, \end{aligned}$$

and

$$h(u; m, n) = 0, \qquad u < 0, m \geq 0, n \geq 0.$$

Then it is easy, for small values m and n, to compute these probabilities. For example, if $m = n = 1$, we have

$$\begin{aligned} h(0; 1, 1) &= \tfrac{1}{2}h(0; 0, 1) + \tfrac{1}{2}h(-1; 1, 0) = \tfrac{1}{2}\cdot 1 + \tfrac{1}{2}\cdot 0 = \tfrac{1}{2}, \\ h(1; 1, 1) &= \tfrac{1}{2}h(1; 0, 1) + \tfrac{1}{2}h(0; 1, 0) = \tfrac{1}{2}\cdot 0 + \tfrac{1}{2}\cdot 1 = \tfrac{1}{2}; \end{aligned}$$

and if $m = 1$, $n = 2$, we have

$$\begin{aligned} h(0; 1, 2) &= \tfrac{1}{3}h(0; 0, 2) + \tfrac{2}{3}h(-1; 1, 1) = \tfrac{1}{3}\cdot 1 + \tfrac{2}{3}\cdot 0 = \tfrac{1}{3}, \\ h(1; 1, 2) &= \tfrac{1}{3}h(1; 0, 2) + \tfrac{2}{3}h(0; 1, 1) = \tfrac{1}{3}\cdot 0 + \tfrac{2}{3}\cdot \tfrac{1}{2} = \tfrac{1}{3}, \\ h(2; 1, 2) &= \tfrac{1}{3}h(2; 0, 2) + \tfrac{2}{3}h(1; 1, 1) = \tfrac{1}{3}\cdot 0 + \tfrac{2}{3}\cdot \tfrac{1}{2} = \tfrac{1}{3}. \end{aligned}$$

In Exercise 11.26, the student is to determine the distribution of U when $m = 2$, $n = 1$; $m = 2$, $n = 2$; $m = 1$, $n = 3$; and $m = 3$, $n = 1$.

For large values of m and n, it is desirable to use an approximate distribution of U. Consider the mean and the variance of U when the hypothesis H_0: $F(z) = G(z)$, for all values of z, is true. Since $U = \sum_{j=1}^{n} \sum_{i=1}^{m} Z_{ij}$, then

$$E(U) = \sum_{i=1}^{m} \sum_{j=1}^{n} E(Z_{ij}).$$

But

$$E(Z_{ij}) = (1) \Pr (X_i < Y_j) + (0) \Pr (X_i > Y_j) = \tfrac{1}{2}$$

because, when H_0 is true, $\Pr (X_i < Y_j) = \Pr (X_i > Y_j) = \frac{1}{2}$. Thus

$$E(U) = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{1}{2}\right) = \frac{mn}{2}.$$

To compute the variance of U, we first find

$$\begin{aligned} E(U^2) &= \sum_{k=1}^{n} \sum_{h=1}^{m} \sum_{j=1}^{n} \sum_{i=1}^{m} E(Z_{ij}Z_{hk}) \\ &= \sum_{j=1}^{n} \sum_{i=1}^{m} E(Z_{ij}^2) + \sum_{\substack{k=1 \\ k \neq j}}^{n} \sum_{j=1}^{n} \sum_{i=1}^{m} E(Z_{ij}Z_{ik}) \\ &\quad + \sum_{j=1}^{n} \sum_{\substack{h=1 \\ h \neq i}}^{m} \sum_{i=1}^{m} E(Z_{ij}Z_{hj}) + \sum_{\substack{k=1 \\ k \neq j}}^{n} \sum_{j=1}^{n} \sum_{\substack{h=1 \\ h \neq i}}^{m} \sum_{i=1}^{m} E(Z_{ij}Z_{hk}). \end{aligned}$$

Note that there are mn terms in the first of these sums, $mn(n-1)$ in the second, $mn(m-1)$ in the third, and $mn(m-1)(n-1)$ in the fourth. When H_0 is true, we know that X_i, X_h, Y_j, and Y_k, $i \neq h$, $j \neq k$, are mutually stochastically independent and have the same distribution of the continuous type. Thus, $\Pr (X_i < Y_j) = \frac{1}{2}$. Moreover, $\Pr (X_i < Y_j, X_i < Y_k) = \frac{1}{3}$ because this is the probability that a designated one of three items is less than each of the other two. Similarly, $\Pr (X_i < Y_j, X_h < Y_j) = \frac{1}{3}$. Finally, $\Pr (X_i < Y_j, X_h < Y_k) = \Pr (X_i < Y_j) \Pr (X_h < Y_k) = \frac{1}{4}$. Hence we have

$$\begin{aligned} E(Z_{ij}^2) &= (1)^2 \Pr (X_i < Y_j) = \tfrac{1}{2}, \\ E(Z_{ij}Z_{ik}) &= (1)(1) \Pr (X_i < Y_j, X_i < Y_k) = \tfrac{1}{3}, \qquad j \neq k, \\ E(Z_{ij}Z_{hj}) &= (1)(1) \Pr (X_i < Y_j, X_h < Y_j) = \tfrac{1}{3}, \qquad i \neq h, \end{aligned}$$

and

$$E(Z_{ij}Z_{hk}) = (1)(1) \Pr (X_i < Y_j, X_h < Y_k) = \tfrac{1}{4}, \qquad i \neq h, j \neq k.$$

Thus,

$$E(U^2) = \frac{mn}{2} + \frac{mn(n-1)}{3} + \frac{mn(m-1)}{3} + \frac{mn(m-1)(n-1)}{4}$$

and

$$\sigma_U^2 = mn\left[\frac{1}{2} + \frac{n-1}{3} + \frac{m-1}{3} + \frac{(m-1)(n-1)}{4} - \frac{mn}{4}\right]$$
$$= \frac{mn(m+n+1)}{12}.$$

Although it is fairly difficult to prove, it is true, when $F(z) = G(z)$ for all z, that

$$\frac{U - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}$$

has, if each of m and n is large, an approximate distribution which is $n(0, 1)$. This fact enables us to compute, approximately, various significance levels.

Prior to the introduction of the statistic U in the statistical literature, it had been suggested that a test of H_0: $F(z) = G(z)$, for all z, be based upon the following statistic, say T (not Student's t). Let T be the sum of the ranks of $Y_1, Y_2, \ldots, Y_n$ among the $m + n$ items $X_1, \ldots, X_m, Y_1, \ldots, Y_n$, once this combined sample has been ordered. In Exercise 11.28, the student is asked to show that

$$U = T - \frac{n(n+1)}{2}.$$

This formula provides another method of computing U and it shows that a test of H_0 based on U is equivalent to a test based on T.

Example 1. With the assumptions and the notation of this section, let $m = 10$ and $n = 9$. Let the observed values of X be as given in the first row and the observed values of Y as in the second row of the following display:

4.3, 5.9, 4.9, 3.1, 5.3, 6.4, 6.2, 3.8, 7.5, 5.8,
5.5, 7.9, 6.8, 9.0, 5.6, 6.3, 8.5, 4.6, 7.1.

Since, in the combined sample, the ranks of the values of y are 4, 7, 8, 12, 14, 15, 17, 18, 19, we have the experimental value of T to be equal to $t = 114$. Thus, $u = 114 - 45 = 69$. If $F(z) = G(z)$ for all z, then, approximately,

$$0.05 = \Pr\left[\frac{U - 45}{12.247} \geq 1.645\right] = \Pr[U \geq 65.146].$$

Accordingly, at the 0.05 significance level, we reject the hypothesis H_0: $F(z) = G(z)$, for all z, and accept the alternative hypothesis H_1: $F(z) \geq G(z)$, for all z.

EXERCISES

11.26. Compute the distribution of U in each of the following cases: (a) $m = 2, n = 1$; (b) $m = 2, n = 2$; (c) $m = 1, n = 3$; (d) $m = 3, n = 1$.

11.27. Suppose the hypothesis H_0: $F(z) = G(z)$, for all z, is not true. Let $p = \Pr(X_i < Y_j)$. Show that U/mn is an unbiased statistic for p and that it converges stochastically to p as $m \to \infty$ and $n \to \infty$.

11.28. Show that $U = T - [n(n + 1)]/2$. *Hint.* Let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$ be the order statistics of the random sample $Y_1, Y_2, \ldots, Y_n$. If R_i is the rank of $Y_{(i)}$ in the combined ordered sample, note that $Y_{(i)}$ is greater than $R_i - i$ values of X.

11.29. In Example 1 of this section, assume that the values came from two independent normal distributions with means μ_1 and μ_2, respectively, and with common variance σ^2. Calculate the Student's t which is used to test the hypothesis H_0: $\mu_1 = \mu_2$. If the alternative hypothesis is H_1: $\mu_1 < \mu_2$, do we accept or reject H_0 at the 0.05 significance level?

11.7 Distributions under Alternative Hypotheses

In this section we shall discuss certain problems which are related to a nonparametric test when the hypothesis H_0 is not true.

Let X and Y be stochastically independent random variables of the continuous type with distribution functions $F(x)$ and $G(y)$, respectively, and probability density functions $f(x)$ and $g(y)$. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ denote independent random samples from these distributions. Consider the hypothesis H_0: $F(z) = G(z)$ for all values of z. It has been seen that the test of this hypothesis may be based upon the statistic U, which, when the hypothesis H_0 is true, has a distribution that does not depend upon $F(z) = G(z)$. Or this test can be based upon the statistic $T = U + n(n + 1)/2$, where T is the sum of the ranks of $Y_1, Y_2, \ldots, Y_n$ in the combined sample. To elicit some information about the distribution of T when the alternative hypothesis is true, let us consider the joint distribution of the ranks of these values of Y.

Let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$ be the order statistics of the sample $Y_1, Y_2, \ldots, Y_n$. Order the combined sample, and let R_i be the rank of $Y_{(i)}$, $i = 1, 2, \ldots, n$. Thus, there are $i - 1$ values of Y and $R_i - i$

values of X which are less than $Y_{(i)}$. Moreover, there are $R_i - R_{i-1} - 1$ values of X between $Y_{(i-1)}$ and $Y_{(i)}$. If it is given that $Y_{(1)} = y_1 < Y_{(2)} = y_2 < \cdots < Y_{(n)} = y_n$, then the conditional probability

$$(1) \qquad \Pr(R_1 = r_1, R_2 = r_2, \ldots, R_n = r_n | y_1 < y_2 < \cdots < y_n),$$

where $r_1 < r_2 < \cdots < r_n \leq m + n$ are positive integers, can be computed by using the multinomial p.d.f. in the following manner. Define the following sets: $A_1 = \{x; -\infty < x < y_1\}$, $A_i = \{x; y_{i-1} < x < y_i\}$, $i = 2, \ldots, n$, $A_{n+1} = \{x; y_n < x < \infty\}$. The conditional probabilities of these sets are, respectively, $p_1 = F(y_1)$, $p_2 = F(y_2) - F(y_1), \ldots, p_n = F(y_n) - F(y_{n-1}), p_{n+1} = 1 - F(y_n)$. Then the conditional probability of Display (1) is given by

$$\frac{m!\, p_1^{r_1-1} p_2^{r_2-r_1-1} \cdots p_n^{r_n-r_{n-1}-1} p_{n+1}^{m+n-r_n}}{(r_1 - 1)!\,(r_2 - r_1 - 1)! \cdots (r_n - r_{n-1} - 1)!\,(m + n - r_n)!}.$$

To find the unconditional probability $\Pr(R_1 = r_1, R_2 = r_2, \ldots, R_n = r_n)$, which we denote simply by $\Pr(r_1, \ldots, r_n)$, we multiply the conditional probability by the joint p.d.f. of $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$, namely, $n!\, g(y_1)g(y_2)\cdots g(y_n)$, and then integrate on $y_1, y_2, \ldots, y_n$. That is,

$$\Pr(r_1, r_2, \ldots, r_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{y_3} \int_{-\infty}^{y_2} \Pr(r_1, \ldots, r_n | y_1 < \cdots < y_n) n! \times g(y_1)\cdots g(y_n)\, dy_1 \cdots dy_n,$$

where $\Pr(r_1, \ldots, r_n | y_1 < \cdots < y_n)$ denotes the conditional probability in Display (1).

Now that we have the joint distribution of $R_1, R_2, \ldots, R_n$, we can find, theoretically, the distributions of functions of $R_1, R_2, \ldots, R_n$ and, in particular, the distribution of $T = \sum_1^n R_i$. From the latter we can find that of $U = T - n(n + 1)/2$. To point out the extremely tedious computational problems of distribution theory which we encounter, we give an example. In this example we use the assumptions of this section.

Example 1. Suppose that an hypothesis H_0 is not true but that in fact $f(x) = 1$, $0 < x < 1$, zero elsewhere, and $g(y) = 2y$, $0 < y < 1$, zero elsewhere. Let $m = 3$ and $n = 2$. Note that the space of U is the set $\{u; u = 0, 1, \ldots, 6\}$. Consider $\Pr(U = 5)$. This event $U = 5$ occurs when and only when $R_1 = 3$, $R_2 = 5$, since in this section $R_1 < R_2$ are the ranks of

$Y_{(1)} < Y_{(2)}$ in the combined sample and $U = R_1 + R_2 - 3$. Because $F(x) = x$, $0 < x \leq 1$, we have

$$\begin{aligned}\Pr(U = 5) &= \Pr(R_1 = 3, R_2 = 5)\\ &= \int_0^1 \int_0^{y_2} \frac{3!\, y_1^2(y_2 - y_1)}{2!\, 1!} 2!\,(2y_1)(2y_2)\, dy_1\, dy_2\\ &= 24 \int_0^1 \left[\frac{y_2^6}{4} - \frac{y_2^6}{5}\right] dy_2 = \tfrac{6}{35}.\end{aligned}$$

Consider next $\Pr(U = 4)$. The event $U = 4$ occurs if $R_1 = 2$, $R_2 = 5$ or if $R_1 = 3$, $R_2 = 4$. Thus,

$$\Pr(U = 4) = \Pr(R_1 = 2, R_2 = 5) + \Pr(R_1 = 3, R_2 = 4);$$

the computation of each of these probabilities is similar to that of $\Pr(R_1 = 3, R_2 = 5)$. This procedure may be continued until we have computed $\Pr(U = u)$ for each $u \in \{u; u = 0, 1, \ldots, 6\}$.

In the preceding example, the probability density functions and the sample sizes m and n were selected so as to provide relatively simple integrations. The student can discover for himself how tedious, and even difficult, the computations become if the sample sizes are large or if the probability density functions are not of a simple functional form.

EXERCISES

11.30. Let the probability density functions of X and Y be those given in Example 1 of this section. Further let the sample sizes be $m = 5$ and $n = 3$. If $R_1 < R_2 < R_3$ are the ranks of $Y_{(1)} < Y_{(2)} < Y_{(3)}$ in the combined sample, compute $\Pr(R_1 = 2, R_2 = 6, R_3 = 8)$.

11.31. Let $X_1, X_2, \ldots, X_m$ be a random sample of size m from a distribution of the continuous type with distribution function $F(x)$ and p.d.f. $F'(x) = f(x)$. Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a distribution with distribution function $G(y) = [F(y)]^\theta$, $0 < \theta$. If $\theta \neq 1$, this distribution is called a *Lehmann alternative*. With $\theta = 2$, show that

$$\Pr(r_1, r_2, \ldots, r_n) = \frac{2^n r_1(r_2 + 1)(r_3 + 2)\cdots(r_n + n - 1)}{\binom{m+n}{m}(m + n + 1)(m + n + 2)\cdots(m + 2n)}.$$

11.32. To generalize the results of Exercise 11.31, let $G(y) = h[F(y)]$, where $h(z)$ is a differentiable function such that $h(0) = 0$, $h(1) = 1$, and $h'(z) > 0$, $0 < z < 1$. Show that

$$\Pr(r_1, r_2, \ldots, r_n) = \frac{E[h'(V_{r_1})h'(V_{r_2})\cdots h'(V_{r_n})]}{\binom{m+n}{m}},$$

where $V_1 < V_2 < \cdots < V_{m+n}$ are the order statistics of a random sample of size $m + n$ from the uniform distribution over the interval $(0, 1)$.

Chapter 12

Further Normal Distribution Theory

12.1 The Multivariate Normal Distribution

We have studied in some detail normal distributions of one and of two random variables. In this section, we shall investigate a joint distribution of n random variables that will be called a *multivariate normal* distribution. This investigation assumes that the student is familiar with elementary matrix algebra, with real symmetric quadratic forms, and with orthogonal transformations. Henceforth the expression quadratic form means a quadratic form in a prescribed number of variables whose matrix is real and symmetric. All symbols which represent matrices will be set in boldface type.

Let $\mathbf{A}$ denote an $n \times n$ real symmetric matrix which is positive definite. Let $\boldsymbol{\mu}$ denote the $n \times 1$ matrix such that $\boldsymbol{\mu}'$, the transpose of $\boldsymbol{\mu}$, is $\boldsymbol{\mu}' = [\mu_1, \mu_2, \ldots, \mu_n]$, where each μ_i is a real constant. Finally, let $\mathbf{x}$ denote the $n \times 1$ matrix such that $\mathbf{x}' = [x_1, x_2, \ldots, x_n]$. We shall show that if C is an appropriately chosen positive constant, the nonnegative function

$$f(x_1, x_2, \ldots, x_n) = C \exp\left[-\frac{(\mathbf{x} - \boldsymbol{\mu})'\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})}{2}\right],$$

$$-\infty < x_i < \infty, i = 1, 2, \ldots, n,$$

is a joint p.d.f. of n random variables $X_1, X_2, \ldots, X_n$ that are of the continuous type. Thus we need to show that

$$(1) \qquad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n = 1.$$

Let $\mathbf{t}$ denote the $n \times 1$ matrix such that $\mathbf{t}' = [t_1, t_2, \ldots, t_n]$, where $t_1, t_2, \ldots, t_n$ are arbitrary real numbers. We shall evaluate the integral

$$(2) \qquad C \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[\mathbf{t}'\mathbf{x} - \frac{(\mathbf{x} - \boldsymbol{\mu})'\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})}{2}\right] dx_1 \cdots dx_n,$$

and then we shall subsequently set $t_1 = t_2 = \cdots = t_n = 0$, and thus establish Equation (1). First we change the variables of integration in integral (2) from $x_1, x_2, \ldots, x_n$ to $y_1, y_2, \ldots, y_n$ by writing $\mathbf{x} - \boldsymbol{\mu} = \mathbf{y}$, where $\mathbf{y}' = [y_1, y_2, \ldots, y_n]$. The Jacobian of the transformation is one and the n-dimensional x-space is mapped onto an n-dimensional y-space, so that integral (2) may be written as

$$(3) \qquad C \exp(\mathbf{t}'\boldsymbol{\mu}) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(\mathbf{t}'\mathbf{y} - \frac{\mathbf{y}'\mathbf{A}\mathbf{y}}{2}\right) dy_1 \cdots dy_n.$$

Because the real symmetric matrix $\mathbf{A}$ is positive definite, the n characteristic numbers (proper values, latent roots, or eigenvalues) $a_1, a_2, \ldots, a_n$ of $\mathbf{A}$ are positive. There exists an appropriately chosen $n \times n$ real orthogonal matrix $\mathbf{L}(\mathbf{L}' = \mathbf{L}^{-1}$, where $\mathbf{L}^{-1}$ is the inverse of $\mathbf{L})$ such that

$$\mathbf{L}'\mathbf{A}\mathbf{L} = \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_n \end{bmatrix},$$

for a suitable ordering of $a_1, a_2, \ldots, a_n$. We shall sometimes write $\mathbf{L}'\mathbf{A}\mathbf{L} = \operatorname{diag}[a_1, a_2, \ldots, a_n]$. In integral (3), we shall change the variables of integration from $y_1, y_2, \ldots, y_n$ to $z_1, z_2, \ldots, z_n$ by writing $\mathbf{y} = \mathbf{L}\mathbf{z}$, where $\mathbf{z}' = [z_1, z_2, \ldots, z_n]$. The Jacobian of the transformation is the determinant of the orthogonal matrix $\mathbf{L}$. Since $\mathbf{L}'\mathbf{L} = \mathbf{I}_n$, where $\mathbf{I}_n$ is the unit matrix of order n, we have the determinant $|\mathbf{L}'\mathbf{L}| = 1$ and $|\mathbf{L}|^2 = 1$. Thus the absolute value of the Jacobian is one. Moreover, the n-dimensional y-space is mapped onto an n-dimensional z-space. The integral (3) becomes

$$(4) \quad C \exp(\mathbf{t}'\boldsymbol{\mu}) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[\mathbf{t}'\mathbf{L}\mathbf{z} - \frac{\mathbf{z}'(\mathbf{L}'\mathbf{A}\mathbf{L})\mathbf{z}}{2}\right] dz_1 \cdots dz_n.$$

It is computationally convenient to write, momentarily, $\mathbf{t}'\mathbf{L} = \mathbf{w}'$, where $\mathbf{w}' = [w_1, w_2, \ldots, w_n]$. Then

$$\exp[\mathbf{t}'\mathbf{L}\mathbf{z}] = \exp[\mathbf{w}'\mathbf{z}] = \exp\left(\sum_1^n w_i z_i\right).$$

Moreover,

$$\exp\left[-\frac{\mathbf{z}'(\mathbf{L}'\mathbf{A}\mathbf{L})\mathbf{z}}{2}\right] = \exp\left[-\frac{\sum_1^n a_i z_i^2}{2}\right].$$

Then integral (4) may be written as the product of n integrals in the following manner:

$$(5)\quad C\exp(\mathbf{w}'\mathbf{L}'\boldsymbol{\mu})\prod_{i=1}^{n}\left[\int_{-\infty}^{\infty}\exp\left(w_i z_i - \frac{a_i z_i^2}{2}\right)dz_i\right]$$
$$= C\exp(\mathbf{w}'\mathbf{L}'\boldsymbol{\mu})\prod_{i=1}^{n}\left[\sqrt{\frac{2\pi}{a_i}}\int_{-\infty}^{\infty}\frac{\exp\left(w_i z_i - \frac{a_i z_i^2}{2}\right)}{\sqrt{2\pi/a_i}}dz_i\right].$$

The integral that involves z_i can be treated as the moment-generating function, with the more familiar symbol t replaced by w_i, of a distribution which is $n(0, 1/a_i)$. Thus the right-hand member of Equation (5) is equal to

$$(6)\quad C\exp(\mathbf{w}'\mathbf{L}'\boldsymbol{\mu})\prod_{i=1}^{n}\left[\sqrt{\frac{2\pi}{a_i}}\exp\left(\frac{w_i^2}{2a_i}\right)\right]$$
$$= C\exp(\mathbf{w}'\mathbf{L}'\boldsymbol{\mu})\sqrt{\frac{(2\pi)^n}{a_1 a_2\cdots a_n}}\exp\left(\sum_1^n\frac{w_i^2}{2a_i}\right).$$

Now, because $\mathbf{L}^{-1} = \mathbf{L}'$, we have

$$(\mathbf{L}'\mathbf{A}\mathbf{L})^{-1} = \mathbf{L}'\mathbf{A}^{-1}\mathbf{L} = \operatorname{diag}\left[\frac{1}{a_1}, \frac{1}{a_2}, \cdots, \frac{1}{a_n}\right].$$

Thus,

$$\sum_1^n\frac{w_i^2}{a_i} = \mathbf{w}'(\mathbf{L}'\mathbf{A}^{-1}\mathbf{L})\mathbf{w} = (\mathbf{L}\mathbf{w})'\mathbf{A}^{-1}(\mathbf{L}\mathbf{w}) = \mathbf{t}'\mathbf{A}^{-1}\mathbf{t}.$$

Moreover, the determinant $|\mathbf{A}^{-1}|$ of $\mathbf{A}^{-1}$ is

$$|\mathbf{A}^{-1}| = |\mathbf{L}'\mathbf{A}^{-1}\mathbf{L}| = \frac{1}{a_1 a_2\cdots a_n}.$$

Accordingly, the right-hand member of Equation (6), which is equal to integral (2), may be written as

$$(7)\qquad Ce^{\mathbf{t}'\boldsymbol{\mu}}\sqrt{(2\pi)^n|\mathbf{A}^{-1}|}\exp\left(\frac{\mathbf{t}'\mathbf{A}^{-1}\mathbf{t}}{2}\right).$$

If, in this function, we set $t_1 = t_2 = \cdots = t_n = 0$, we have the value of the left-hand member of Equation (1). Thus, we have

$$C\sqrt{(2\pi)^n|\mathbf{A}^{-1}|} = 1.$$

Accordingly, the function

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{|\mathbf{A}^{-1}|}} \exp\left[-\frac{(\mathbf{x} - \boldsymbol{\mu})'\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})}{2}\right],$$

$-\infty < x_i < \infty$, $i = 1, 2, \ldots, n$, is a joint p.d.f. of n random variables $X_1, X_2, \ldots, X_n$ that are of the continuous type. Such a p.d.f. is called a *nonsingular multivariate normal p.d.f.*

We have now proved that $f(x_1, x_2, \ldots, x_n)$ is a p.d.f. However, we have proved more than that. Because $f(x_1, x_2, \ldots, x_n)$ is a p.d.f., integral (2) is the moment-generating function $M(t_1, t_2, \ldots, t_n)$ of this joint distribution of probability. Since integral (2) is equal to function (7), the moment-generating function of the multivariate normal distribution is given by

$$M(t_1, t_2, \ldots, t_n) = \exp\left(\mathbf{t}'\boldsymbol{\mu} + \frac{\mathbf{t}'\mathbf{A}^{-1}\mathbf{t}}{2}\right).$$

Let the elements of the real, symmetric, and positive definite matrix $\mathbf{A}^{-1}$ be denoted by σ_{ij}, $i, j = 1, 2, \ldots, n$. Then

$$M(0, \ldots, 0, t_i, 0, \ldots, 0) = \exp\left(t_i\mu_i + \frac{\sigma_{ii}t_i^2}{2}\right)$$

is the moment-generating function of X_i, $i = 1, 2, \ldots, n$. Thus, X_i is $n(\mu_i, \sigma_{ii})$, $i = 1, 2, \ldots, n$. Moreover, with $i \neq j$, we see that $M(0, \ldots, 0, t_i, 0, \ldots, 0, t_j, 0, \ldots, 0)$, the moment-generating function of X_i and X_j, is equal to

$$\exp\left(t_i\mu_i + t_j\mu_j + \frac{\sigma_{ii}t_i^2 + 2\sigma_{ij}t_it_j + \sigma_{jj}t_j^2}{2}\right).$$

But this is the moment-generating function of a bivariate normal distribution, so that σ_{ij} is the covariance of the random variables X_i and X_j. Thus the matrix $\boldsymbol{\mu}$, where $\boldsymbol{\mu}' = [\mu_1, \mu_2, \ldots, \mu_n]$, is the matrix of the means of the random variables $X_1, \ldots, X_n$. Moreover, the elements on the principal diagonal of $\mathbf{A}^{-1}$ are, respectively, the variances $\sigma_{ii} = \sigma_i^2$, $i = 1, 2, \ldots, n$, and the elements not on the principal diagonal of $\mathbf{A}^{-1}$ are, respectively, the covariances $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$, $i \neq j$, of the random variables $X_1, X_2, \ldots, X_n$. We call the matrix $\mathbf{A}^{-1}$, which is given by

$$\begin{bmatrix} \sigma_{11}, & \sigma_{12}, & \ldots, & \sigma_{1n} \\ \sigma_{12}, & \sigma_{22}, & \ldots, & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{1n}, & \sigma_{2n}, & \ldots, & \sigma_{nn} \end{bmatrix},$$

the *covariance matrix* of the multivariate normal distribution and henceforth we shall denote this matrix by the symbol $\mathbf{V}$. In terms of the positive definite covariance matrix $\mathbf{V}$, the multivariate normal p.d.f. is written

$$\frac{1}{(2\pi)^{n/2}\sqrt{|\mathbf{V}|}} \exp\left[-\frac{(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right], \qquad -\infty < x_i < \infty,$$

$i = 1, 2, \ldots, n$, and the moment-generating function of this distribution is given by

$$\exp\left(\mathbf{t}'\boldsymbol{\mu} + \frac{\mathbf{t}'\mathbf{V}\mathbf{t}}{2}\right)$$

for all real values of $\mathbf{t}$.

Example 1. Let $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution with matrix $\boldsymbol{\mu}$ of means and positive definite covariance matrix $\mathbf{V}$. If we let $\mathbf{X}' = [X_1, X_2, \ldots, X_n]$, then the moment-generating function $M(t_1, t_2, \ldots, t_n)$ of this joint distribution of probability is

$$E(e^{\mathbf{t}'\mathbf{X}}) = \exp\left(\mathbf{t}'\boldsymbol{\mu} + \frac{\mathbf{t}'\mathbf{V}\mathbf{t}}{2}\right). \tag{8}$$

Consider a linear function Y of $X_1, X_2, \ldots, X_n$ which is defined by $Y = \mathbf{c}'\mathbf{X} = \sum_1^n c_i X_i$, where $\mathbf{c}' = [c_1, c_2, \ldots, c_n]$ and the several c_i are real and not all zero. We wish to find the p.d.f. of Y. The moment-generating function $M(t)$ of the distribution of Y is given by

$$M(t) = E(e^{tY}) = E(e^{t\mathbf{c}'\mathbf{X}}).$$

Now the expectation (8) exists for all real values of $\mathbf{t}$. Thus we can replace $\mathbf{t}'$ in expectation (8) by $t\mathbf{c}'$ and obtain

$$M(t) = \exp\left(t\mathbf{c}'\boldsymbol{\mu} + \frac{\mathbf{c}'\mathbf{V}\mathbf{c}t^2}{2}\right).$$

Thus the random variable Y is $n(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\mathbf{V}\mathbf{c})$.

EXERCISES

12.1. Let $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution with positive definite covariance matrix $\mathbf{V}$. Prove that these random variables are mutually stochastically independent if and only if $\mathbf{V}$ is a diagonal matrix.

12.2. Let $n = 2$ and take

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

Determine $|\mathbf{V}|$, $\mathbf{V}^{-1}$, and $(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})$. Compare the bivariate normal p.d.f. with the multivariate normal p.d.f. when $n = 2$.

12.3. Let $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution, where $\boldsymbol{\mu}$ is the matrix of the means and $\mathbf{V}$ is the positive definite covariance matrix. Let $Y = \mathbf{c}'\mathbf{X}$ and $Z = \mathbf{d}'\mathbf{X}$, where $\mathbf{X}' = [X_1, \ldots, X_n]$, $\mathbf{c}' = [c_1, \ldots, c_n]$, and $\mathbf{d}' = [d_1, \ldots, d_n]$ are real matrices. (a) Find $M(t_1, t_2) = E(e^{t_1 Y + t_2 Z})$ to see that Y and Z have a bivariate normal distribution. (b) Prove that Y and Z are stochastically independent if and only if $\mathbf{c}'\mathbf{V}\mathbf{d} = 0$. (c) If $X_1, X_2, \ldots, X_n$ are mutually stochastically independent random variables which have the same variance σ^2, show that the necessary and sufficient condition of part (b) becomes $\mathbf{c}'\mathbf{d} = 0$.

12.4. Let $\mathbf{X}' = [X_1, X_2, \ldots, X_n]$ have the multivariate normal distribution of Exercise 12.3. Consider the p linear functions of $X_1, \ldots, X_n$ defined by $\mathbf{W} = \mathbf{B}\mathbf{X}$, where $\mathbf{W}' = [W_1, \ldots, W_p]$, $p \leq n$, and $\mathbf{B}$ is a $p \times n$ real matrix of rank p. Find $M(v_1, \ldots, v_p) = E(e^{\mathbf{v}'\mathbf{W}})$, where $\mathbf{v}'$ is the real matrix $[v_1, \ldots, v_p]$, to see that $W_1, \ldots, W_p$ have a p-variate normal distribution which has $\mathbf{B}\boldsymbol{\mu}$ for the matrix of the means and $\mathbf{B}\mathbf{V}\mathbf{B}'$ for the covariance matrix.

12.5. Let $\mathbf{X}' = [X_1, X_2, \ldots, X_n]$ have the n-variate normal distribution of Exercise 12.3. Show that $X_1, X_2, \ldots, X_p$, $p < n$, have a p-variate normal distribution. What submatrix of $\mathbf{V}$ is the covariance matrix of $X_1, X_2, \ldots, X_p$? *Hint.* In the moment-generating function $M(t_1, t_2, \ldots, t_n)$ of $X_1, X_2, \ldots, X_n$, let $t_{p+1} = \cdots = t_n = 0$.

12.2 The Distributions of Certain Quadratic Forms

Let X_i, $i = 1, 2, \ldots, n$, denote mutually stochastically independent random variables which are $n(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, n$, respectively. Then $Q = \sum_1^n (X_i - \mu_i)^2/\sigma_i^2$ is $\chi^2(n)$. Now Q is a quadratic form in the $X_i - \mu_i$ and Q is seen to be, apart from the coefficient $-\frac{1}{2}$, the random variable which is defined by the exponent on the number e in the joint p.d.f. of $X_1, X_2, \ldots, X_n$. We shall now show that this result can be generalized.

Let $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution with p.d.f.

$$\frac{1}{(2\pi)^{n/2}\sqrt{|\mathbf{V}|}} \exp\left[-\frac{(\mathbf{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})}{2}\right],$$

where, as usual, the covariance matrix $\mathbf{V}$ is positive definite. We shall show that the random variable Q (a quadratic form in the $X_i - \mu_i$),

which is defined by $(\mathbf{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})$, is $\chi^2(n)$. We have for the moment-generating function $M(t)$ of Q the integral

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}\sqrt{|\mathbf{V}|}} \times \exp\left[t(\mathbf{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu}) - \frac{(\mathbf{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})}{2}\right] dx_1 \cdots dx_n$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}\sqrt{|\mathbf{V}|}} \times \exp\left[-\frac{(\mathbf{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})(1 - 2t)}{2}\right] dx_1 \cdots dx_n.$$

With $\mathbf{V}^{-1}$ positive definite, the integral is seen to exist for all real values of $t < \frac{1}{2}$. Moreover, $(1 - 2t)\mathbf{V}^{-1}$, $t < \frac{1}{2}$, is a positive definite matrix and, since $|(1 - 2t)\mathbf{V}^{-1}| = (1 - 2t)^n|\mathbf{V}^{-1}|$, it follows that

$$\frac{1}{(2\pi)^{n/2}\sqrt{|\mathbf{V}|/(1 - 2t)^n}} \exp\left[-\frac{(\mathbf{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})(1 - 2t)}{2}\right]$$

can be treated as a multivariate normal p.d.f. If we multiply our integrand by $(1 - 2t)^{n/2}$, we have this multivariate p.d.f. Thus the moment-generating function of Q is given by

$$M(t) = \frac{1}{(1 - 2t)^{n/2}}, \qquad t < \tfrac{1}{2},$$

and Q is $(\chi^2 n)$, as we wished to show. This fact is the basis of the chi-square tests which were discussed in Chapter 10.

The remarkable fact that the random variable which is defined by $(\mathbf{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is $\chi^2(n)$ stimulates a number of questions about quadratic forms in normally distributed variables. We would like to treat this problem in complete generality, but limitations of space forbid this, and we find it necessary to restrict ourselves to some special cases.

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution which is $n(0, \sigma^2)$, $\sigma^2 > 0$. Let $\mathbf{X}' = [X_1, X_2, \ldots, X_n]$ and let $\mathbf{A}$ denote an arbitrary $n \times n$ real symmetric matrix. We shall investigate the distribution of the quadratic form $\mathbf{X}'\mathbf{A}\mathbf{X}$. For instance, we know that $\mathbf{X}'\mathbf{I}_n\mathbf{X}/\sigma^2 = \mathbf{X}'\mathbf{X}/\sigma^2 = \sum_1^n X_i^2/\sigma^2$ is $\chi^2(n)$. First we shall find the moment-generating function of $\mathbf{X}'\mathbf{A}\mathbf{X}/\sigma^2$. Then we shall investigate

the conditions which must be imposed upon the real symmetric matrix $\mathbf{A}$ if $\mathbf{X}'\mathbf{A}\mathbf{X}/\sigma^2$ is to have a chi-square distribution. This moment-generating function is given by

$$M(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(\frac{t\mathbf{x}'\mathbf{A}\mathbf{x}}{\sigma^2} - \frac{\mathbf{x}'\mathbf{x}}{2\sigma^2}\right) dx_1 \cdots dx_n$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{\mathbf{x}'(\mathbf{I} - 2t\mathbf{A})\mathbf{x}}{2\sigma^2}\right] dx_1 \cdots dx_n,$$

where $\mathbf{I} = \mathbf{I}_n$. The matrix $\mathbf{I} - 2t\mathbf{A}$ is positive definite if we take $|t|$ sufficiently small, say $|t| < h$, $h > 0$. Moreover, we can treat

$$\frac{1}{(2\pi)^{n/2}\sqrt{|(\mathbf{I} - 2t\mathbf{A})^{-1}\sigma^2|}} \exp\left[-\frac{\mathbf{x}'(\mathbf{I} - 2t\mathbf{A})\mathbf{x}}{2\sigma^2}\right]$$

as a multivariate normal p.d.f. Now $|(\mathbf{I} - 2t\mathbf{A})^{-1}\sigma^2|^{1/2} = \sigma^n/|\mathbf{I} - 2t\mathbf{A}|^{1/2}$. If we multiply our integrand by $|\mathbf{I} - 2t\mathbf{A}|^{1/2}$, we have this multivariate p.d.f. Hence the moment-generating function of $\mathbf{X}'\mathbf{A}\mathbf{X}/\sigma^2$ is given by

$$M(t) = |\mathbf{I} - 2t\mathbf{A}|^{-1/2}, \qquad |t| < h. \tag{1}$$

It proves useful to express this moment-generating function in a different form. To do this, let $a_1, a_2, \ldots, a_n$ denote the characteristic numbers of $\mathbf{A}$ and let $\mathbf{L}$ denote an $n \times n$ orthogonal matrix such that $\mathbf{L}'\mathbf{A}\mathbf{L} = \text{diag}\,[a_1, a_2, \ldots, a_n]$. Thus,

$$\mathbf{L}'(\mathbf{I} - 2t\mathbf{A})\mathbf{L} = \begin{bmatrix} (1 - 2ta_1) & 0 & \cdots & 0 \\ 0 & (1 - 2ta_2) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & (1 - 2ta_n) \end{bmatrix}.$$

Then

$$\prod_{i=1}^{n} (1 - 2ta_i) = |\mathbf{L}'(\mathbf{I} - 2t\mathbf{A})\mathbf{L}| = |\mathbf{I} - 2t\mathbf{A}|.$$

Accordingly we can write $M(t)$, as given in Equation (1), in the form

$$M(t) = \left[\prod_{i=1}^{n} (1 - 2ta_i)\right]^{-1/2}, \qquad |t| < h. \tag{2}$$

Let r, $0 < r \leq n$, denote the rank of the real symmetric matrix $\mathbf{A}$. Then exactly r of the real numbers $a_1, a_2, \ldots, a_n$, say $a_1, \ldots, a_r$, are not zero and exactly $n - r$ of these numbers, say $a_{r+1}, \ldots, a_n$, are zero. Thus we can write the moment-generating function of $\mathbf{X}'\mathbf{A}\mathbf{X}/\sigma^2$ as

$$M(t) = [(1 - 2ta_1)(1 - 2ta_2)\cdots(1 - 2ta_r)]^{-1/2}.$$

Now that we have found, in suitable form, the moment-generating function of our random variable, let us turn to the question of the con-

ditions that must be imposed if $\mathbf{X}'\mathbf{A}\mathbf{X}/\sigma^2$ is to have a chi-square distribution. Assume that $\mathbf{X}'\mathbf{A}\mathbf{X}/\sigma^2$ is $\chi^2(k)$. Then

$$M(t) = [(1 - 2ta_1)(1 - 2ta_2)\cdots(1 - 2ta_r)]^{-1/2} = (1 - 2t)^{-k/2},$$

or, equivalently,

$$(1 - 2ta_1)(1 - 2ta_2)\cdots(1 - 2ta_r) = (1 - 2t)^k, \qquad |t| < h.$$

Because the positive integers r and k are the degrees of these polynomials, and because these polynomials are equal for infinitely many values of t, we have $k = r$, the rank of $\mathbf{A}$. Moreover, the uniqueness of the factorization of a polynomial implies that $a_1 = a_2 = \cdots = a_r = 1$. If each of the nonzero characteristic numbers of a real symmetric matrix is one, the matrix is idempotent, that is, $\mathbf{A}^2 = \mathbf{A}$, and conversely (see Exercise 12.7). Accordingly, if $\mathbf{X}'\mathbf{A}\mathbf{X}/\sigma^2$ has a chi-square distribution, then $\mathbf{A}^2 = \mathbf{A}$ and the random variable is $\chi^2(r)$, where r is the rank of $\mathbf{A}$. Conversely, if $\mathbf{A}$ is of rank r, $0 < r \leq n$, and if $\mathbf{A}^2 = \mathbf{A}$, then $\mathbf{A}$ has exactly r characteristic numbers that are equal to one, and the remaining $n - r$ characteristic numbers are equal to zero. Thus the moment-generating function of $\mathbf{X}'\mathbf{A}\mathbf{X}/\sigma^2$ is given by $(1 - 2t)^{-r/2}$, $t < \frac{1}{2}$, and $\mathbf{X}'\mathbf{A}\mathbf{X}/\sigma^2$ is $\chi^2(r)$. This establishes the following theorem.

Theorem 1. *Let Q denote a random variable which is a quadratic form in the items of a random sample of size n from a distribution which is $n(0, \sigma^2)$. Let $\mathbf{A}$ denote the symmetric matrix of Q and let r, $0 < r \leq n$, denote the rank of $\mathbf{A}$. Then Q/σ^2 is $\chi^2(r)$ if and only if $\mathbf{A}^2 = \mathbf{A}$.*

Remark. If the normal distribution in Theorem 1 is $n(\mu, \sigma^2)$, the condition $\mathbf{A}^2 = \mathbf{A}$ remains a necessary and sufficient condition that Q/σ^2 have a chi-square distribution. In general, however, Q/σ^2 is not $\chi^2(r)$ but, instead, Q/σ^2 has a noncentral chi-square distribution if $\mathbf{A}^2 = \mathbf{A}$. The number of degrees of freedom is r, the rank of $\mathbf{A}$, and the noncentrality parameter is $\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}/\sigma^2$, where $\boldsymbol{\mu}' = [\mu, \mu, \ldots, \mu]$. Since $\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} = \mu^2 \sum_{i,j} a_{ij}$, where $\mathbf{A} = [a_{ij}]$, then, if $\mu \neq 0$, the conditions $\mathbf{A}^2 = \mathbf{A}$ and $\sum_{i,j} a_{ij} = 0$ are necessary and sufficient conditions that Q/σ^2 be central $\chi^2(r)$. Moreover, the theorem may be extended to a quadratic form in random variables which have a multivariate normal distribution with positive definite covariance matrix $\mathbf{V}$; here the necessary and sufficient condition that Q have a chi-square distribution is $\mathbf{A}\mathbf{V}\mathbf{A} = \mathbf{A}$.

EXERCISES

12.6. Let $Q = X_1X_2 - X_3X_4$, where X_1, X_2, X_3, X_4 is a random sample of size 4 from a distribution which is $n(0, \sigma^2)$. Show that Q/σ^2 does not have a chi-square distribution. Find the moment-generating function of Q/σ^2.

12.7. Let $\mathbf{A}$ be a real symmetric matrix. Prove that each of the nonzero characteristic numbers of $\mathbf{A}$ is equal to one if and only if $\mathbf{A}^2 = \mathbf{A}$. *Hint.* Let $\mathbf{L}$ be an orthogonal matrix such that $\mathbf{L'AL} = \text{diag}\,[a_1, a_2, \ldots, a_n]$ and note that $\mathbf{A}$ is idempotent if and only if $\mathbf{L'AL}$ is idempotent.

12.8. The sum of the elements on the principal diagonal of a square matrix $\mathbf{A}$ is called the trace of $\mathbf{A}$ and is denoted by tr $\mathbf{A}$. (a) If $\mathbf{B}$ is $n \times m$ and $\mathbf{C}$ is $m \times n$, prove that tr $(\mathbf{BC}) =$ tr $(\mathbf{CB})$. (b) If $\mathbf{A}$ is a square matrix and if $\mathbf{L}$ is an orthogonal matrix, use the result of part (a) to show that tr $(\mathbf{L'AL}) =$ tr $\mathbf{A}$. (c) If $\mathbf{A}$ is a real symmetric idempotent matrix, use the result of part (b) to prove that the rank of $\mathbf{A}$ is equal to tr $\mathbf{A}$.

12.9. Let $\mathbf{A} = [a_{ij}]$ be a real symmetric matrix. Prove that $\sum_j \sum_i a_{ij}^2$ is equal to the sum of the squares of the characteristic numbers of $\mathbf{A}$. *Hint.* If $\mathbf{L}$ is an orthogonal matrix, show that $\sum_j \sum_i a_{ij}^2 =$ tr $(\mathbf{A}^2) =$ tr $(\mathbf{L'A^2L}) =$ tr $[(\mathbf{L'AL})(\mathbf{L'AL})]$.

12.10. Let $\bar{X}$ and S^2 denote, respectively, the mean and the variance of a random sample of size n from a distribution which is $n(0, \sigma^2)$. (a) If $\mathbf{A}$ denotes the symmetric matrix of $n\bar{X}^2$, show that $\mathbf{A} = (1/n)\mathbf{P}$, where $\mathbf{P}$ is the $n \times n$ matrix, each of whose elements is equal to one. (b) Demonstrate that $\mathbf{A}$ is idempotent and that the tr $\mathbf{A} = 1$. Thus $n\bar{X}^2/\sigma^2$ is $\chi^2(1)$. (c) Show that the symmetric matrix $\mathbf{B}$ of nS^2 is $\mathbf{I} - (1/n)\mathbf{P}$. (d) Demonstrate that $\mathbf{B}$ is idempotent and that tr $\mathbf{B} = n - 1$. Thus nS^2/σ^2 is $\chi^2(n - 1)$, as previously proved otherwise. (e) Show that the product matrix $\mathbf{AB}$ is the zero matrix.

12.3 The Independence of Certain Quadratic Forms

We have previously investigated the stochastic independence of linear functions of normally distributed variables (see Exercise 12.3). In this section we shall prove some theorems about the stochastic independence of quadratic forms. As we remarked on p. 385, we shall confine our attention to normally distributed variables that constitute a random sample of size n from a distribution that is $n(0, \sigma^2)$.

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution which is $n(0, \sigma^2)$. Let $\mathbf{A}$ and $\mathbf{B}$ denote two real symmetric matrices, each of order n. Let $\mathbf{X}' = [X_1, X_2, \ldots, X_n]$ and consider the two quadratic forms $\mathbf{X'AX}$ and $\mathbf{X'BX}$. We wish to show that these quadratic forms are stochastically independent if and only if $\mathbf{AB} = \mathbf{0}$, the zero matrix. We shall first compute the moment-generating function $M(t_1, t_2)$ of the joint distribution of $\mathbf{X'AX}/\sigma^2$ and $\mathbf{X'BX}/\sigma^2$. We have

$$M(t_1, t_2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(\frac{t_1\mathbf{x}'\mathbf{A}\mathbf{x}}{\sigma^2} + \frac{t_2\mathbf{x}'\mathbf{B}\mathbf{x}}{\sigma^2} - \frac{\mathbf{x}'\mathbf{x}}{2\sigma^2}\right) dx_1 \cdots dx_n$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-\frac{\mathbf{x}'(\mathbf{I} - 2t_1\mathbf{A} - 2t_2\mathbf{B})\mathbf{x}}{2\sigma^2}\right) dx_1 \cdots dx_n.$$

The matrix $\mathbf{I} - 2t_1\mathbf{A} - 2t_2\mathbf{B}$ is positive definite if we take $|t_1|$ and $|t_2|$ sufficiently small, say $|t_1| < h_1$, $|t_2| < h_2$, where $h_1, h_2 > 0$. Then, as on p. 386, we have

$$M(t_1, t_2) = |\mathbf{I} - 2t_1\mathbf{A} - 2t_2\mathbf{B}|^{-1/2}, \qquad |t_1| < h_1, \ |t_2| < h_2.$$

Let us assume that $\mathbf{X}'\mathbf{A}\mathbf{X}/\sigma^2$ and $\mathbf{X}'\mathbf{B}\mathbf{X}/\sigma^2$ are stochastically independent (so that likewise are $\mathbf{X}'\mathbf{A}\mathbf{X}$ and $\mathbf{X}'\mathbf{B}\mathbf{X}$) and prove that $\mathbf{A}\mathbf{B} = \mathbf{0}$. Thus we assume that

$$M(t_1, t_2) = M(t_1, 0)M(0, t_2) \tag{1}$$

for all t_1 and t_2 for which $|t_i| < h_i$, $i = 1, 2$. Identity (1) is equivalent to the identity

$$|\mathbf{I} - 2t_1\mathbf{A} - 2t_2\mathbf{B}| = |\mathbf{I} - 2t_1\mathbf{A}|\,|\mathbf{I} - 2t_2\mathbf{B}|, \qquad |t_i| < h_i, \ i = 1, 2. \tag{2}$$

Let $r > 0$ denote the rank of $\mathbf{A}$ and let $a_1, a_2, \ldots, a_r$ denote the r nonzero characteristic numbers of $\mathbf{A}$. There exists an orthogonal matrix $\mathbf{L}$ such that

$$\mathbf{L}'\mathbf{A}\mathbf{L} = \left[\begin{array}{cccc:c} a_1 & 0 & \cdots & 0 & \\ 0 & a_2 & \cdots & 0 & \mathbf{0} \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & \cdots & a_r & \\ \hdashline & \multicolumn{3}{c}{\mathbf{0}} & \mathbf{0} \end{array}\right] = \left[\begin{array}{c:c} \mathbf{C}_{11} & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{0} \end{array}\right] = \mathbf{C}$$

for a suitable ordering of $a_1, a_2, \ldots, a_r$. Then $\mathbf{L}'\mathbf{B}\mathbf{L}$ may be written in the identically partitioned form

$$\mathbf{L}'\mathbf{B}\mathbf{L} = \left[\begin{array}{c:c} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \hdashline \mathbf{D}_{21} & \mathbf{D}_{22} \end{array}\right] = \mathbf{D}.$$

The identity (2) may be written as

$$|\mathbf{L}'|\,|\mathbf{I} - 2t_1\mathbf{A} - 2t_2\mathbf{B}|\,|\mathbf{L}| = |\mathbf{L}'|\,|\mathbf{I} - 2t_1\mathbf{A}|\,|\mathbf{L}|\,|\mathbf{L}'|\,|\mathbf{I} - 2t_2\mathbf{B}|\,|\mathbf{L}|, \tag{2'}$$

or as

$$|\mathbf{I} - 2t_1\mathbf{C} - 2t_2\mathbf{D}| = |\mathbf{I} - 2t_1\mathbf{C}|\,|\mathbf{I} - 2t_2\mathbf{D}|. \tag{3}$$

The coefficient of $(-2t_1)^r$ in the right-hand member of Equation (3) is seen by inspection to be $a_1a_2\cdots a_r|\mathbf{I} - 2t_2\mathbf{D}|$. It is not so easy to find the coefficient of $(-2t_1)^r$ in the left-hand member of Equation (3). Conceive of expanding this determinant in terms of minors of order r formed from the first r columns. One term in this expansion is the product of the minor of order r in the upper left-hand corner, namely, $|\mathbf{I}_r - 2t_1\mathbf{C}_{11} - 2t_2\mathbf{D}_{11}|$, and the minor of order $n - r$ in the lower right-hand corner, namely, $|\mathbf{I}_{n-r} - 2t_2\mathbf{D}_{22}|$. Moreover, this product is the only term in the expansion of the determinant that involves $(-2t_1)^r$. Thus the coefficient of $(-2t_1)^r$ in the left-hand member of Equation (3) is $a_1a_2\cdots a_r|\mathbf{I}_{n-r} - 2t_2\mathbf{D}_{22}|$. If we equate these coefficients of $(-2t_1)^r$, we have, for all t_2, $|t_2| < h_2$,

$$|\mathbf{I} - 2t_2\mathbf{D}| = |\mathbf{I}_{n-r} - 2t_2\mathbf{D}_{22}|. \tag{4}$$

Equation (4) implies that the nonzero characteristic numbers of the matrices $\mathbf{D}$ and $\mathbf{D}_{22}$ are the same (see Exercise 12.17). Recall that the sum of the squares of the characteristic numbers of a symmetric matrix is equal to the sum of the squares of the elements of that matrix (see Exercise 12.9). Thus the sum of the squares of the elements of matrix $\mathbf{D}$ is equal to the sum of the squares of the elements of $\mathbf{D}_{22}$. Since the elements of the matrix $\mathbf{D}$ are real, it follows that each of the elements of $\mathbf{D}_{11}$, $\mathbf{D}_{12}$, and $\mathbf{D}_{21}$ is zero. Accordingly, we can write $\mathbf{D}$ in the form

$$\mathbf{D} = \mathbf{L}'\mathbf{B}\mathbf{L} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{22} \end{bmatrix}.$$

Thus $\mathbf{CD} = \mathbf{L}'\mathbf{A}\mathbf{L}\mathbf{L}'\mathbf{B}\mathbf{L} = \mathbf{0}$ and $\mathbf{L}'\mathbf{A}\mathbf{B}\mathbf{L} = \mathbf{0}$ and $\mathbf{AB} = \mathbf{0}$, as we wished to prove. To complete the proof of the theorem, we assume that $\mathbf{AB} = \mathbf{0}$. We are to show that $\mathbf{X}'\mathbf{AX}/\sigma^2$ and $\mathbf{X}'\mathbf{BX}/\sigma^2$ are stochastically independent. We have, for all real values of t_1 and t_2,

$$(\mathbf{I} - 2t_1\mathbf{A})(\mathbf{I} - 2t_2\mathbf{B}) = \mathbf{I} - 2t_1\mathbf{A} - 2t_2\mathbf{B},$$

since $\mathbf{AB} = \mathbf{0}$. Thus,

$$|\mathbf{I} - 2t_1\mathbf{A} - 2t_2\mathbf{B}| = |\mathbf{I} - 2t_1\mathbf{A}|\,|\mathbf{I} - 2t_2\mathbf{B}|.$$

Since the moment-generating function of the joint distribution of $\mathbf{X}'\mathbf{AX}/\sigma^2$ and $\mathbf{X}'\mathbf{BX}/\sigma^2$ is given by

$$M(t_1, t_2) = |\mathbf{I} - 2t_1\mathbf{A} - 2t_2\mathbf{B}|^{-1/2}, \qquad |t_i| < h_i,\ i = 1, 2,$$

we have

$$M(t_1, t_2) = M(t_1, 0)M(0, t_2),$$

and the proof of the following theorem is complete.

Theorem 2. *Let Q_1 and Q_2 denote random variables which are quadratic forms in the items of a random sample of size n from a distribution which is $n(0, \sigma^2)$. Let* **A** *and* **B** *denote respectively the real symmetric matrices of Q_1 and Q_2. The random variables Q_1 and Q_2 are stochastically independent if and only if* $\mathbf{AB} = \mathbf{0}$.

Remark. Theorem 2 remains valid if the random sample is from a distribution which is $n(\mu, \sigma^2)$, whatever be the real value of μ. Moreover, Theorem 2 may be extended to quadratic forms in random variables that have a joint multivariate normal distribution with a positive definite covariance matrix **V**. The necessary and sufficient condition for the stochastic independence of two such quadratic forms with symmetric matrices **A** and **B** then becomes $\mathbf{AVB} = \mathbf{0}$. In our Theorem 2, we have $\mathbf{V} = \sigma^2\mathbf{I}$, so that $\mathbf{AVB} = \mathbf{A}\sigma^2\mathbf{IB} = \sigma^2\mathbf{AB} = \mathbf{0}$.

We shall next prove the theorem that was used in Chapter 10 (p. 316).

Theorem 3. *Let $Q = Q_1 + \cdots + Q_{k-1} + Q_k$, where $Q, Q_1, \ldots, Q_{k-1}, Q_k$ are $k + 1$ random variables that are quadratic forms in the items of a random sample of size n from a distribution which is $n(0, \sigma^2)$. Let Q/σ^2 be $\chi^2(r)$, let Q_i/σ^2 be $\chi^2(r_i)$, $i = 1, 2, \ldots, k - 1$, and let Q_k be nonnegative. Then the random variables $Q_1, Q_2, \ldots, Q_k$ are mutually stochastically independent and, hence, Q_k/σ^2 is $\chi^2(r_k = r - r_1 - \cdots - r_{k-1})$.*

Proof. Take first the case of $k = 2$ and let the real symmetric matrices of Q, Q_1, and Q_2 be denoted, respectively, by $\mathbf{A}, \mathbf{A}_1, \mathbf{A}_2$. We are given that $Q = Q_1 + Q_2$ or, equivalently, that $\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2$. We are also given that Q/σ^2 is $\chi^2(r)$ and that Q_1/σ^2 is $\chi^2(r_1)$. In accordance with Theorem 1, p. 387, we have $\mathbf{A}^2 = \mathbf{A}$ and $\mathbf{A}_1^2 = \mathbf{A}_1$. Since $Q_2 \geq 0$, each of the matrices $\mathbf{A}, \mathbf{A}_1$, and $\mathbf{A}_2$ is positive semidefinite. Because $\mathbf{A}^2 = \mathbf{A}$, we can find an orthogonal matrix **L** such that

$$\mathbf{L'AL} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

If then we multiply both members of $\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2$ on the left by $\mathbf{L}'$ and on the right by **L**, we have

$$\begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \mathbf{L'A_1L} + \mathbf{L'A_2L}.$$

Now each of $\mathbf{A}_1$ and $\mathbf{A}_2$, and hence each of $\mathbf{L}'\mathbf{A}_1\mathbf{L}$ and $\mathbf{L}'\mathbf{A}_2\mathbf{L}$ is positive semidefinite. Recall that, if a real symmetric matrix is positive semidefinite, each element on the principal diagonal is positive or zero. Moreover, if an element on the principal diagonal is zero, then all elements in that row and all elements in that column are zero. Thus $\mathbf{L}'\mathbf{A}\mathbf{L} = \mathbf{L}'\mathbf{A}_1\mathbf{L} + \mathbf{L}'\mathbf{A}_2\mathbf{L}$ can be written as

$$\left[\begin{array}{c|c} \mathbf{I}_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{G}_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] + \left[\begin{array}{c|c} \mathbf{H}_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]. \tag{5}$$

Since $\mathbf{A}_1^2 = \mathbf{A}_1$, we have

$$(\mathbf{L}'\mathbf{A}_1\mathbf{L})^2 = \mathbf{L}'\mathbf{A}_1\mathbf{L} = \left[\begin{array}{c|c} \mathbf{G}_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right].$$

If we multiply both members of Equation (5) on the left by the matrix $\mathbf{L}'\mathbf{A}_1\mathbf{L}$, we see that

$$\left[\begin{array}{c|c} \mathbf{G}_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{G}_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] + \left[\begin{array}{c|c} \mathbf{G}_r\mathbf{H}_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right],$$

or, equivalently, $\mathbf{L}'\mathbf{A}_1\mathbf{L} = \mathbf{L}'\mathbf{A}_1\mathbf{L} + (\mathbf{L}'\mathbf{A}_1\mathbf{L})(\mathbf{L}'\mathbf{A}_2\mathbf{L})$. Thus, $(\mathbf{L}'\mathbf{A}_1\mathbf{L}) \times (\mathbf{L}'\mathbf{A}_2\mathbf{L}) = \mathbf{0}$ and $\mathbf{A}_1\mathbf{A}_2 = \mathbf{0}$. In accordance with Theorem 2, Q_1 and Q_2 are stochastically independent. This stochastic independence immediately implies that Q_2/σ^2 is $\chi^2(r_2 = r - r_1)$. This completes the proof when $k = 2$. For $k > 2$, the proof may be made by induction. We shall merely indicate how this can be done by using $k = 3$. Take $\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_3$, where $\mathbf{A}^2 = \mathbf{A}$, $\mathbf{A}_1^2 = \mathbf{A}_1$, $\mathbf{A}_2^2 = \mathbf{A}_2$ and $\mathbf{A}_3$ is positive semidefinite. Write $\mathbf{A} = \mathbf{A}_1 + (\mathbf{A}_2 + \mathbf{A}_3) = \mathbf{A}_1 + \mathbf{B}_1$, say. Now $\mathbf{A}^2 = \mathbf{A}$, $\mathbf{A}_1^2 = \mathbf{A}_1$, and $\mathbf{B}_1$ is positive semidefinite. In accordance with the case of $k = 2$, we have $\mathbf{A}_1\mathbf{B}_1 = \mathbf{0}$, so that $\mathbf{B}_1^2 = \mathbf{B}_1$. With $\mathbf{B}_1 = \mathbf{A}_2 + \mathbf{A}_3$, where $\mathbf{B}_1^2 = \mathbf{B}_1$, $\mathbf{A}_2^2 = \mathbf{A}_2$, it follows from the case of $k = 2$ that $\mathbf{A}_2\mathbf{A}_3 = \mathbf{0}$ and $\mathbf{A}_3^2 = \mathbf{A}_3$. If we regroup by writing $\mathbf{A} = \mathbf{A}_2 + (\mathbf{A}_1 + \mathbf{A}_3)$, we obtain $\mathbf{A}_1\mathbf{A}_3 = \mathbf{0}$, and so on.

Remark. In our statement of Theorem 3 we took $X_1, X_2, \ldots, X_n$ to be items of a random sample from a distribution which is $n(0, \sigma^2)$. We did this because our proof of Theorem 2 was restricted to that case. In fact, if $Q', Q_1', \ldots, Q_k'$ are quadratic forms in any normal variables (including multivariate normal variables), if $Q' = Q_1' + \cdots + Q_k'$, if $Q', Q_1', \ldots, Q_{k-1}'$ are central or noncentral chi-square, and if Q_k' is nonnegative, then $Q_1', \ldots, Q_k'$

are mutually stochastically independent and Q_k' is either central or noncentral chi-square.

This section will conclude with a proof of a frequently quoted theorem due to Cochran.

Theorem 4. *Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution which is $n(0, \sigma^2)$. Let the sum of the squares of these items be written in the form*

$$\sum_1^n X_i^2 = Q_1 + Q_2 + \cdots + Q_k,$$

where Q_j is a quadratic form in $X_1, X_2, \ldots, X_n$, with matrix $\mathbf{A}_j$ which has rank r_j, $j = 1, 2, \ldots, k$. The random variables $Q_1, Q_2, \ldots, Q_k$ are mutually stochastically independent and Q_j/σ^2 is $\chi^2(r_j)$, $j = 1, 2, \ldots, k$, if and only if $\sum_1^k r_j = n$.

Proof. First assume the two conditions $\sum_1^k r_j = n$ and $\sum_1^n X_i^2 = \sum_1^k Q_j$ to be satisfied. The latter equation implies that $\mathbf{I} = \mathbf{A}_1 + \mathbf{A}_2 + \cdots + \mathbf{A}_k$. Let $\mathbf{B}_i = \mathbf{I} - \mathbf{A}_i$. That is, $\mathbf{B}_i$ is the sum of the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_k$ exclusive of $\mathbf{A}_i$. Let R_i denote the rank of $\mathbf{B}_i$. Since the rank of the sum of several matrices is less than or equal to the sum of the ranks, we have $R_i \leq \sum_1^k r_j - r_i = n - r_i$. However, $\mathbf{I} = \mathbf{A}_i + \mathbf{B}_i$, so that $n \leq r_i + R_i$ and $n - r_i \leq R_i$. Hence $R_i = n - r_i$. The characteristic numbers of $\mathbf{B}_i$ are the roots of the equation $|\mathbf{B}_i - \lambda\mathbf{I}| = 0$. Since $\mathbf{B}_i = \mathbf{I} - \mathbf{A}_i$, this equation can be written as $|\mathbf{I} - \mathbf{A}_i - \lambda\mathbf{I}| = 0$. Thus, we have $|\mathbf{A}_i - (1 - \lambda)\mathbf{I}| = 0$. But each root of the last equation is one minus a characteristic number of $\mathbf{A}_i$. Since $\mathbf{B}_i$ has exactly $n - R_i = r_i$ characteristic numbers that are zero, then $\mathbf{A}_i$ has exactly r_i characteristic numbers that are equal to one. However, r_i is the rank of $\mathbf{A}_i$. Thus, each of the r_i nonzero characteristic numbers of $\mathbf{A}_i$ is one. That is, $\mathbf{A}_i^2 = \mathbf{A}_i$ and thus Q_i/σ^2 is $\chi^2(r_i)$, $i = 1, 2, \ldots, k$. In accordance with Theorem 3, the random variables $Q_1, Q_2, \ldots, Q_k$ are mutually stochastically independent.

To complete the proof of Theorem 4, take $\sum_1^n X_i^2 = Q_1 + Q_2 + \cdots + Q_k$, let $Q_1, Q_2, \ldots, Q_k$ be mutually stochastically independent, and let Q_j/σ^2 be $\chi^2(r_j)$, $j = 1, 2, \ldots, k$. Then $\sum_1^k Q_j/\sigma^2$ is $\chi^2\left(\sum_1^k r_j\right)$. But $\sum_1^k Q_j/\sigma^2 = \sum_1^n X_i^2/\sigma^2$ is $\chi^2(n)$. Thus, $\sum_1^k r_j = n$ and the proof is complete.

EXERCISES

12.11. Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution which is $n(0, \sigma^2)$. Prove that $\sum_1^n X_i^2$ and every quadratic form, which is nonidentically zero in $X_1, X_2, \ldots, X_n$, are stochastically dependent.

12.12 Let X_1, X_2, X_3, X_4 denote a random sample of size 4 from a distribution which is $n(0, \sigma^2)$. Let $Y = \sum_1^4 a_i X_i$, where a_1, a_2, a_3, and a_4 are real constants. If Y^2 and $Q = X_1X_2 - X_3X_4$ are stochastically independent, determine a_1, a_2, a_3, and a_4.

12.13. Let $\mathbf{A}$ be the real symmetric matrix of a quadratic form Q in the items of a random sample of size n from a distribution which is $n(0, \sigma^2)$. Given that Q and the mean $\bar{X}$ of the sample are stochastically independent. What can be said of the elements of each row (column) of $\mathbf{A}$? *Hint.* Are Q and $\bar{X}^2$ stochastically independent?

12.14. Let $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ be the matrices of $k > 2$ quadratic forms $Q_1, Q_2, \ldots, Q_k$ in the items of a random sample of size n from a distribution which is $n(0, \sigma^2)$. Prove that the pairwise stochastic independence of these forms implies that they are mutually stochastically independent. *Hint.* Show that $\mathbf{A}_i\mathbf{A}_j = \mathbf{0}$, $i \neq j$, permits $E[\exp(t_1Q_1 + t_2Q_2 + \cdots + t_kQ_k)]$ to be written as a product of the moment-generating functions of $Q_1, Q_2, \ldots, Q_k$.

12.15. Let $\mathbf{X}' = [X_1, X_2, \ldots, X_n]$, where $X_1, X_2, \ldots, X_n$ are items of a random sample from a distribution which is $n(0, \sigma^2)$. Let $\mathbf{b}' = [b_1, b_2, \ldots, b_n]$ be a real nonzero matrix, and let $\mathbf{A}$ be a real symmetric matrix of order n. Prove that the linear form $\mathbf{b}'\mathbf{X}$ and the quadratic form $\mathbf{X}'\mathbf{A}\mathbf{X}$ are stochastically independent if and only if $\mathbf{b}'\mathbf{A} = \mathbf{0}$. Use this fact to prove that $\mathbf{b}'\mathbf{X}$ and $\mathbf{X}'\mathbf{A}\mathbf{X}$ are stochastically independent if and only if the two quadratic forms, $(\mathbf{b}'\mathbf{X})^2 = \mathbf{X}'\mathbf{b}\mathbf{b}'\mathbf{X}$ and $\mathbf{X}'\mathbf{A}\mathbf{X}$, are stochastically independent.

12.16. Let Q_1 and Q_2 be two nonnegative quadratic forms in the items of a random sample from a distribution which is $n(0, \sigma^2)$. Show that another quadratic form Q is stochastically independent of $Q_1 + Q_2$ if and only if Q is stochastically independent of each of Q_1 and Q_2. *Hint.* Consider the orthogonal transformation that diagonalizes the matrix of $Q_1 + Q_2$. After this transformation, what are the forms of the matrices of Q, Q_1, and Q_2 if Q and $Q_1 + Q_2$ are stochastically independent?

12.17. Prove that Equation (4) of this section implies that the nonzero characteristic numbers of the matrices $\mathbf{D}$ and $\mathbf{D}_{22}$ are the same. *Hint.* Let $\lambda = 1/(2t_2)$, $t_2 \neq 0$, and show that Equation (4) is equivalent to $|\mathbf{D} - \lambda\mathbf{I}| = (-\lambda)^r|\mathbf{D}_{22} - \lambda\mathbf{I}_{n-r}|$.

Appendix A

References

[1] Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, New York, 1958.

[2] Basu, D., "On Statistics Independent of a Complete Sufficient Statistic," *Sankhyā*, **15**, 377 (1955).

[3] Blackwell, D., and Girshick, M. A., *Theory of Games and Statistical Decisions*, John Wiley & Sons, New York, 1954.

[4] Brunk, H. D., *An Introduction to Mathematical Statistics*, 2nd ed., Blaisdell Publishing Company, Waltham, Mass., 1965.

[5] Carpenter, O., "Note on the Extension of Craig's Theorem to Non-central Variates," *Ann. Math. Stat.*, **21**, 455 (1950).

[6] Cochran, W. G., "The Distribution of Quadratic Forms in a Normal System, with Applications to the Analysis of Covariance," *Proc. Camb. Phil. Soc.*, **30**, 178 (1934).

[7] Craig, A. T., "Bilinear Forms in Normally Correlated Variables," *Ann. Math. Stat.*, **18**, 565 (1947).

[8] Craig, A. T., "Note on the Independence of Certain Quadratic Forms," *Ann. Math. Stat.*, **14**, 195 (1943).

[9] Cramér, H., *Mathematical Methods of Statistics*, Princeton University Press, Princeton, N.J., 1946.

[10] Curtiss, J. H., "A Note on the Theory of Moment Generating Functions," *Ann. Math. Stat.*, **13**, 430 (1942).

[11] Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. I, 2nd ed., John Wiley & Sons, New York, 1957.

[12] Ferguson, T. S., *Mathematical Statistics*, Academic Press, Inc., New York, 1967.

[13] Fisher, R. A., "On the Mathematical Foundation of Theoretical Statistics," *Phil. Trans. Royal Soc. London*, Series A, **222**, 309 (1921).

[14] Fraser, D. A. S., *Nonparametric Methods in Statistics*, John Wiley & Sons, New York, 1957.

[15] Graybill, F. A., *An Introduction to Linear Statistical Models*, Vol. 1, McGraw-Hill, Inc., New York, 1961.

[16] Hogg, R. V., and Craig, A. T., "On the Decomposition of Certain Chi-Square Variables," *Ann. Math. Stat.*, **29**, 608 (1958).

[17] Hogg, R. V., and Craig, A. T., "Sufficient Statistics in Elementary Distribution Theory," *Sankhyā*, **17**, 209 (1956).

[18] Koopman, B. O., "On Distributions Admitting a Sufficient Statistic," *Trans. Amer. Math. Soc.*, **39**, 399 (1936).

[19] Lancaster, H. O., "Traces and Cumulants of Quadratic Forms in Normal Variables," *J. Royal Stat. Soc.*, Series B, **16**, 247 (1954).

[20] Lehmann, E. L., *Testing Statistical Hypotheses*, John Wiley & Sons, New York, 1959.

[21] Lehmann, E. L., and Scheffé, H., "Completeness, Similar Regions, and Unbiased Estimation," *Sankhyā*, **10**, 305 (1950).

[22] Lévy, P., *Théorie de l'addition des variables aléatoires*, Gauthier-Villars, Paris, 1937.

[23] Lindgren, B. W., *Statistical Theory*, 2nd ed., The Macmillan Company, New York, 1968.

[24] Madow, W. G., "Contributions to the Theory of Multivariate Statistical Analysis," *Trans. Amer. Math. Soc.*, **44**, 454 (1938).

[25] Mann, H. B., and Whitney, D. R., "On a Test of whether One of Two Random Variables Is Stochastically Larger than the Other," *Ann. Math. Stat.*, **18**, 50 (1947).

[26] Meyer, P. L., *Introductory Probability and Statistical Applications*, Addison-Wesley Publishing Company, Inc., Reading, Mass., 1965.

[27] Neyman, J., "Su un teorema concernente le cosiddette statistiche sufficienti," *Giornale dell' Istituto degli Attuari*, **6**, 320 (1935).

[28] Neyman, J., and Pearson, E. S., "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Phil. Trans. Royal Soc. London*, Series A, **231**, 289 (1933).

[29] Pearson, K., "On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such that It Can Be Reasonably Supposed to Have Arisen from Random Sampling," *Phil. Mag.*, Series 5, **50**, 157 (1900).

[30] Pitman, E. J. G., "Sufficient Statistics and Intrinsic Accuracy," *Proc. Cam. Phil. Soc.*, **32**, 567 (1936).

[31] Raiffa, H., and Schlaifer, R., *Applied Statistical Decision Theory*, published by Division of Research, Harvard Business School, Harvard University, Boston, 1961.

[32] Rao, C. R., *Linear Statistical Inference and Its Applications*, John Wiley & Sons, Inc., New York, 1965.

[33] Scheffé, H., *The Analysis of Variance*, John Wiley & Sons, New York, 1959.

[34] Wald, A., *Sequential Analysis*, John Wiley & Sons, New York, 1947.
[35] Wilcoxon, F., "Individual Comparisons by Ranking Methods," *Biometrices Bull.*, **1**, 80 (1945).
[36] Wilks, S. S., *Mathematical Statistics*, John Wiley & Sons, New York, 1962.

Appendix B

Tables

TABLE I

The Poisson Distribution

$$\Pr(X \leq x) = \sum_{w=0}^{x} \frac{\mu^w e^{-\mu}}{w!}$$

	$\mu = E(X)$											
x	0.5	1.0	1.5	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
0	0.607	0.368	0.223	0.135	0.050	0.018	0.007	0.002	0.001	0.000	0.000	0.000
1	0.910	0.736	0.558	0.406	0.199	0.092	0.040	0.017	0.007	0.003	0.001	0.000
2	0.986	0.920	0.809	0.677	0.423	0.238	0.125	0.062	0.030	0.014	0.006	0.003
3	0.998	0.981	0.934	0.857	0.647	0.433	0.265	0.151	0.082	0.042	0.021	0.010
4	1.000	0.996	0.981	0.947	0.815	0.629	0.440	0.285	0.173	0.100	0.055	0.029
5		0.999	0.996	0.983	0.916	0.785	0.616	0.446	0.301	0.191	0.116	0.067
6		1.000	0.999	0.995	0.966	0.889	0.762	0.606	0.450	0.313	0.207	0.130
7			1.000	0.999	0.988	0.949	0.867	0.744	0.599	0.453	0.324	0.220
8				1.000	0.996	0.979	0.932	0.847	0.729	0.593	0.456	0.333
9					0.999	0.992	0.968	0.916	0.830	0.717	0.587	0.458
10					1.000	0.997	0.986	0.957	0.901	0.816	0.706	0.583
11						0.999	0.995	0.980	0.947	0.888	0.803	0.697
12						1.000	0.998	0.991	0.973	0.936	0.876	0.792
13							0.999	0.996	0.987	0.966	0.926	0.864
14							1.000	0.999	0.994	0.983	0.959	0.917
15								0.999	0.998	0.992	0.978	0.951
16								1.000	0.999	0.996	0.989	0.973
17									1.000	0.998	0.995	0.986
18										0.999	0.998	0.993
19										1.000	0.999	0.997
20											1.000	0.998
21												0.999
22												1.000

TABLE II

*The Chi-Square Distribution**

$$\Pr(X \leq x) = \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1} e^{-w/2}\, dw$$

	Pr ($X \leq x$)					
r	0.01	0.025	0.050	0.95	0.975	0.99
1	0.000	0.001	0.004	3.84	5.02	6.63
2	0.020	0.051	0.103	5.99	7.38	9.21
3	0.115	0.216	0.352	7.81	9.35	11.3
4	0.297	0.484	0.711	9.49	11.1	13.3
5	0.554	0.831	1.15	11.1	12.8	15.1
6	0.872	1.24	1.64	12.6	14.4	16.8
7	1.24	1.69	2.17	14.1	16.0	18.5
8	1.65	2.18	2.73	15.5	17.5	20.1
9	2.09	2.70	3.33	16.9	19.0	21.7
10	2.56	3.25	3.94	18.3	20.5	23.2
11	3.05	3.82	4.57	19.7	21.9	24.7
12	3.57	4.40	5.23	21.0	23.3	26.2
13	4.11	5.01	5.89	22.4	24.7	27.7
14	4.66	5.63	6.57	23.7	26.1	29.1
15	5.23	6.26	7.26	25.0	27.5	30.6
16	5.81	6.91	7.96	26.3	28.8	32.0
17	6.41	7.56	8.67	27.6	30.2	33.4
18	7.01	8.23	9.39	28.9	31.5	34.8
19	7.63	8.91	10.1	30.1	32.9	36.2
20	8.26	9.59	10.9	31.4	34.2	37.6
21	8.90	10.3	11.6	32.7	35.5	38.9
22	9.54	11.0	12.3	33.9	36.8	40.3
23	10.2	11.7	13.1	35.2	38.1	41.6
24	10.9	12.4	13.8	36.4	39.4	43.0
25	11.5	13.1	14.6	37.7	40.6	44.3
26	12.2	13.8	15.4	38.9	41.9	45.6
27	12.9	14.6	16.2	40.1	43.2	47.0
28	13.6	15.3	16.9	41.3	44.5	48.3
29	14.3	16.0	17.7	42.6	45.7	49.6
30	15.0	16.8	18.5	43.8	47.0	50.9

* This table is abridged and adapted from "Tables of Percentage Points of the Incomplete Beta Function and of the Chi-Square Distribution," *Biometrika*, **32** (1941). It is published here with the kind permission of Professor E. S. Pearson on behalf of the author, Catherine M. Thompson, and of the Biometrika Trustees.

TABLE III

The Normal Distribution

$$\Pr(X \leq x) = N(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw$$

$$[N(-x) = 1 - N(x)]$$

x	$N(x)$	x	$N(x)$	x	$N(x)$
0.00	0.500	1.10	0.864	2.05	0.980
0.05	0.520	1.15	0.875	2.10	0.982
0.10	0.540	1.20	0.885	2.15	0.984
0.15	0.560	1.25	0.894	2.20	0.986
0.20	0.579	1.282	0.900	2.25	0.988
0.25	0.599	1.30	0.903	2.30	0.989
0.30	0.618	1.35	0.911	2.326	0.990
0.35	0.637	1.40	0.919	2.35	0.991
0.40	0.655	1.45	0.926	2.40	0.992
0.45	0.674	1.50	0.933	2.45	0.993
0.50	0.691	1.55	0.939	2.50	0.994
0.55	0.709	1.60	0.945	2.55	0.995
0.60	0.726	1.645	0.950	2.576	0.995
0.65	0.742	1.65	0.951	2.60	0.995
0.70	0.758	1.70	0.955	2.65	0.996
0.75	0.773	1.75	0.960	2.70	0.997
0.80	0.788	1.80	0.964	2.75	0.997
0.85	0.802	1.85	0.968	2.80	0.997
0.90	0.816	1.90	0.971	2.85	0.998
0.95	0.829	1.95	0.974	2.90	0.998
1.00	0.841	1.960	0.975	2.95	0.998
1.05	0.853	2.00	0.977	3.00	0.999

TABLE IV

*The t Distribution**

$$\Pr(T \leq t) = \int_{-\infty}^{t} \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r}\,\Gamma(r/2)(1 + w^2/r)^{(r+1)/2}}\,dw$$

$$[\Pr(T \leq -t) = 1 - \Pr(T \leq t)]$$

r	$\Pr(T \leq t)$ 0.90	0.95	0.975	0.99	0.995
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750

* This table is abridged from Table III of Fisher and Yates; *Statistical Tables for Biological, Agricultural, and Medical Research*, published by Oliver and Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

TABLE V

*The F Distribution**

$$\Pr(F \le f) = \int_0^f \frac{\Gamma[(r_1 + r_2)/2](r_1/r_2)^{r_1/2} w^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)(1 + r_1 w/r_2)^{(r_1+r_2)/2}}\, dw$$

$\Pr(F \le f)$	r_2	r_1 = 1	2	3	4	5	6	7	8	9	10	12	15
0.95	1	161	200	216	225	230	234	237	239	241	242	244	246
0.975		648	800	864	900	922	937	948	957	963	969	977	985
0.99		4052	4999	5403	5625	5764	5859	5928	5982	6023	6056	6106	6157
0.95	2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4
0.975		38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4
0.99		98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4
0.95	3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70
0.975		17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.3
0.99		34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9
0.95	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86
0.975		12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66
0.99		21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2
0.95	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62
0.975		10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43
0.99		16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72

0.95	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94
0.975		8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27
0.99		13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56
0.95	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51
0.975		8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57
0.99		12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31
0.95	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22
0.975		7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10
0.99		11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52
0.95	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01
0.975		7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77
0.99		10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96
0.95	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85
0.975		6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52
0.99		10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56
0.95	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62
0.975		6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18
0.99		9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01
0.95	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40
0.975		6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86
0.99		8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52

* This table is abridged and adapted from "Tables of Percentage Points of the Inverted Beta Distribution," *Biometrika*, **33** (1943). It is published here with the kind permission of Professor E. S. Pearson on behalf of the authors, Maxine Merrington and Catherine M. Thompson, and of the *Biometrika* Trustees.

Appendix C

Answers to Selected Exercises

CHAPTER 1

1.2 (a) $\{x; x = 0, 1, 2, 3, 4\}$; $\{x; x = 2\}$.
(b) $\{x; 0 < x < 3\}$; $\{x; 1 \leq x < 2\}$.

1.3 (a) $\{x; 0 < x < \frac{5}{8}\}$.

1.6 (a) $\{x; 0 < x < 3\}$;
(b) $\{(x, y); 0 < x^2 + y^2 < 4\}$.

1.7 (a) $\{x; x = 2\}$;
(b) null set;
(c) $\{(x, y); x^2 + y^2 = 0\}$.

1.8 $\frac{80}{81}$; 1.

1.9 $\frac{11}{16}$; 0; 1.

1.10 16; 7; 21; 2.

1.12 $\frac{8}{3}$; 0; $\pi/2$.

1.14 10.

1.18 $\frac{1}{4}$; $\frac{1}{13}$; $\frac{1}{52}$; $\frac{4}{13}$.

1.19 $\frac{31}{32}$; $\frac{3}{64}$; $\frac{1}{32}$; $\frac{63}{64}$.

1.24 $\frac{9}{13}$, $\frac{1}{13}$, $\frac{1}{13}$, $\frac{1}{13}$, $\frac{1}{13}$.

1.25 $\int_A (\frac{1}{20})\, dx$, $A \subset \{x; -10 < x < 10\}$.

1.26 (a) $\frac{3}{52}$; (b) $\frac{1}{13}$.

1.28 $\frac{3}{4}$.

1.30 $\frac{5}{8}$; $\frac{7}{8}$; $\frac{3}{8}$.

1.31 $\frac{1}{4}$.

1.32 $e^{-2} - e^{-3}$.

1.33 (a) $\frac{1}{2}$; (b) 1.

1.36 $\frac{15}{64}$; 0; $\frac{1}{2}$; $\frac{1}{2}$.

1.38 (a) 1; (b) $\frac{2}{3}$; (c) 2.

1.44 (a) $0, x < 0$; $1, 0 \leq x$.
(d) $0, x < 0$; $1 - (1 - x)^3$, $0 \leq x < 1$; $1, 1 \leq x$.

1.45 (b) 0; (e) 2; (f) 2.5.

1.46 (a) $\frac{1}{4}$; (b) 0; (c) $\frac{1}{4}$; (d) 0.

1.48 $0, y < 0$; $y^2, 0 \leq y < 1$; $1, 1 \leq y$, $2y, 0 < y < 1$; 0 elsewhere.

1.53 $1/3\sqrt{y}, 0 < y < 1$; $1/6\sqrt{y}$, $1 < y < 4$; 0 elsewhere.

1.54 (a) $\binom{6}{4}\Big/\binom{16}{4}$;
(b) $\binom{10}{4}\Big/\binom{16}{4}$.

1.55 $1 - \binom{990}{5}\Big/\binom{1000}{5}$.

1.57 (b) $1 - \binom{10}{3}\Big/\binom{20}{3}$.
1.60 $-\ln z$, $0 < z < 1$; 0 elsewhere.
1.62 $(x-1)(10-x)(9-x)/420$, $x = 2, 3, \ldots, 8$.
1.64 2; 86.4; −160.8.
1.65 3; 11; 27.
1.66 $\frac{1}{9}$.
1.68 $(\frac{1}{3})(\frac{2}{3}) \neq \frac{1}{4}$.
1.69 (a) $\frac{1}{2}$; (b) $f(1) = \frac{1}{3}$.
1.70 \$7.80.
1.73 $\frac{7}{3}$.
1.74 (a) 1.5, 0.75; (b) 0.5, 0.05; (c) 2; does not exist.
1.75 $e^t/(2-e^t)$, $t < \ln 2$; 2; 2.
1.83 $(1-t)^{-3}$, $t < 1$.
1.86 10; 0; 2; −30.
1.89 $\frac{31}{12}$, $\frac{167}{144}$.
1.94 0.84.

CHAPTER 2

2.3 (a) $\frac{1}{7}$; (b) $\frac{5}{56}$; (c) $\left[\binom{3}{x}\Big/\binom{8}{x}\right][5/(8-x)]$.
2.6 $\frac{1}{3}$.
2.8 $\frac{9}{20}$; $\frac{2}{3}$.
2.9 $\frac{5}{14}$.
2.10 $(3x_1 + 2)/(6x_1 + 3)$; $(6x_1^2 + 6x_1 + 1)/(2)(6x_1 + 3)^2$.
2.12 $3x_2/4$; $3x_2^2/80$.
2.17 (b) $1/e$.
2.18 (a) 1; (b) −1; (c) 0.
2.19 $7/\sqrt{804}$.
2.26 $b_2 = \sigma_1(\rho_{12} - \rho_{13}\rho_{23})/[\sigma_2(1-\rho_{23}^2)]$; $b_3 = \sigma_1(\rho_{13} - \rho_{12}\rho_{23})/[\sigma_3(1-\rho_{23}^2)]$.
2.31 $\frac{7}{8}$.
2.33 $1-(1-y)^{12}$, $0 \leq y < 1$; $12(1-y)^{11}$, $0 < y < 1$.
2.34 $g(y) = [y^3 - (y-1)^3]/6^3$, $y = 1, 2, 3, 4, 5, 6$.
2.35 $\frac{1}{2}$.

CHAPTER 3

3.1 $\frac{40}{81}$.
3.4 $\frac{1471}{5112}$.
3.6 5.
3.9 $\frac{65}{81}$.
3.11 $\frac{5}{72}$.
3.14 $\frac{1}{6}$.
3.15 $\frac{24}{625}$.
3.17 $\frac{11}{6}$; $x_1/2$; $\frac{11}{6}$.
3.18 $\frac{25}{4}$.
3.19 0.09.
3.22 $4^x e^{-4}/x!$, $x = 0, 1, 2, \ldots$.
3.23 0.84.
3.27 (a) $\exp[-2 + e^{t_2}(1 + e^{t_1})]$; (b) $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, $\rho = \sqrt{2}/2$; (c) $y/2$.
3.28 0.05.
3.29 0.831, 12.8.
3.30 0.90.
3.31 $\chi^2(4)$.
3.33 $3e^{-3y}$, $0 < y < \infty$.
3.34 2, 0.95.
3.35 8, $4\sqrt{3}$.

3.36 $\frac{11}{16}$.
3.37 $\chi^2(2)$.
3.39 0.067; 0.685.
3.41 71.3, 189.7.
3.42 $e^{-4}/\sqrt{\pi}$.
3.45 0.774.
3.46 $\sqrt{2/\pi}$; $(\pi - 2)/\pi$.
3.47 0.90.
3.48 0.477.
3.49 0.461.
3.50 $n(0, 1)$.
3.51 0.433.
3.53 $n(0, 2)$.
3.54 (a) 0.264; (b) 0.440; (c) 0.433; (d) 0.642.
3.57 $\rho = \frac{4}{5}$.
3.58 (38.2, 43.4).

CHAPTER 4

4.2 2.4; 0.048.
4.4 0.405.
4.5 $\frac{1}{8}$.
4.7 $(n + 1)/2$; $(n^2 - 1)/12$.
4.8 $a + b\bar{x}$; $b^2 s_x^2$.
4.9 $\chi^2(2)$.
4.11 $\frac{1}{2}$, $0 < y < 1$; $1/2y^2$, $1 < y < \infty$.
4.12 y^{15}, $0 \le y < 1$; $15y^{14}$, $0 < y < 1$.
4.13 $\frac{4}{7}$.
4.14 $\frac{1}{3}$, $y = 3, 5, 7$.
4.16 $(\frac{1}{2})^{\sqrt[3]{y}}$, $y = 1, 8, 27, \ldots$.
4.17

y_1	$g_1(y_1)$
1	$\frac{1}{36}$
2	$\frac{4}{36}$
3	$\frac{6}{36}$
4	$\frac{4}{36}$
6	$\frac{12}{36}$
9	$\frac{9}{36}$

4.21 $\frac{1}{27}$, $0 < y < 27$.
4.26 $\alpha/(\alpha + \beta)$; $\alpha\beta/[(\alpha + \beta + 1)(\alpha + \beta)^2]$.
4.27 (a) 20; (b) 1260; (c) 495.
4.28 $\frac{10}{243}$.
4.31 $(x - 1)^2/4$ or $1 - (x - 1)^2/4$.
4.35 0.05.
4.38 1/4.74, 3.33.
4.42 $(1/\sqrt{2\pi})^3 y_1^2 e^{-y_1^2/2} \sin y_3$, $0 \le y_1 < \infty$, $0 \le y_2 < 2\pi$, $0 \le y_3 \le \pi$.
4.43 $y_2 y_3^2 e^{-y_3}$, $0 < y_1 < 1$, $0 < y_2 < 1$, $0 < y_3 < \infty$.
4.47 $1/(2\sqrt{y})$, $0 < y < 1$.
4.48 $e^{-y_1/2}/(2\pi\sqrt{y_1 - y_2^2})$, $-\sqrt{y_1} < y_2 < \sqrt{y_1}$, $0 < y_1 < \infty$.
4.50 $1 - (1 - e^{-3})^4$.
4.51 $\frac{1}{8}$.
4.55 $\frac{5}{16}$.
4.56 $48 z_1 z_2^3 z_3^5$, $0 < z_1 < 1$, $0 < z_2 < 1$, $0 < z_3 < 1$.
4.57 $\frac{7}{12}$.
4.62 $6uv(u + v)$, $0 < u < v < 1$.
4.64

y	$g(y)$
2	$\frac{1}{36}$
3	$\frac{2}{36}$
4	$\frac{3}{36}$
5	$\frac{4}{36}$
6	$\frac{5}{36}$
7	$\frac{6}{36}$
8	$\frac{5}{36}$
9	$\frac{4}{36}$
10	$\frac{3}{36}$
11	$\frac{2}{36}$
12	$\frac{1}{36}$

4.65 0.24.
4.72 0.323.
4.74 0.818.
4.78 $\sum_1^n a_i b_i = 0$.
4.80 6.41.
4.81 $n = 16$.
4.84 $(n-1)\sigma^2/n$; $2(n-1)\sigma^4/n^2$.
4.85 0.90.
4.87 0.78.
4.88 $\frac{8}{3}$; $\frac{2}{9}$.
4.89 7.
4.91 2.5; 0.25.
4.95 -5; $60 - 12\sqrt{6}$.
4.96 $\sigma_1/\sqrt{\sigma_1^2 + \sigma_2^2}$.
4.99 22.5, $\frac{261}{4}$.
4.100 $r_2 > 4$.
4.102 $\mu_2\sigma_1/\sqrt{\sigma_1^2\sigma_2^2 + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2}$.
4.106 $5/\sqrt{39}$.

CHAPTER 5

5.1 Degenerate at μ.
5.2 Gamma ($\alpha = 1, \beta = 1$).
5.3 Gamma ($\alpha = 1, \beta = 1$).
5.5 Gamma ($\alpha = 2; \beta = 1$).
5.14 0.682.
5.15 (b) 0.815.
5.18 Degenerate at $\mu_2 + (\sigma_2/\sigma_1)(x - \mu_1)$.
5.19 (b) $n(0, 1)$.
5.20 (b) $n(0, 1)$.
5.22 0.954.
5.24 0.840.
5.27 0.08.
5.29 0.267.
5.30 0.682.
5.34 $n(0, 1)$.

CHAPTER 6

6.1 0.318; 2.
6.2 $\frac{13}{16}$.
6.3 $\frac{7}{12}$.
6.5 0.95; $2.86\sigma^2$.
6.6 (77.28, 85.12).
6.7 24 or 25.
6.8 (3.7, 5.7).
6.9 160.
6.15 $(5\bar{x}/6, 5\bar{x}/4)$.
6.16 $(-3.6, 2.0)$.
6.19 135 or 136.
6.21 (0.43, 2.21).
6.23 (2.68, 9.68).
6.24 (0.71, 5.50).
6.30 (b) $\dfrac{(n\beta)^y\Gamma(\alpha + y)}{y!\,(n\beta + 1)^{y+\alpha}\Gamma(\alpha)}$, $y = 0, 1, 2, \ldots$.

CHAPTER 7

7.4 $\frac{1}{3}$, $\frac{2}{3}$.
7.13 $60y_3^2(y_5 - y_3)/\theta^5$; $6y_5/5$; $\theta^2/7$; $\theta^2/35$.
7.14 $(1/\theta^2)e^{-y_1/\theta}$, $0 < y_2 < y_1 < \infty$; $y_1/2$; $\theta^2/2$.
7.23 (a) Y_1/n; (b) Y_1/n.
7.24 $Y_1 - 1/n$.
7.27 $[Q(y_n) - ny_nQ'(y_n)]/nM(y_n)Q^2(y_n)$.
7.28 $Y_1 = \sum_1^n X_i$; $Y_1/4n$; yes.
7.35 $\bar{x}$.
7.36 $Y(n - Y)/(n - 1)$.
7.39 $\left(\dfrac{n-1}{n}\right)^Y\left(1 + \dfrac{Y}{n-1}\right)$.

CHAPTER 8

8.3 θ^2/n; $\theta^2/n(n+2)$.

8.8 (a) $\bar{X}$;
(b) $-n/\ln(X_1X_2\cdots X_n)$;
(c) $\bar{X}$; (d) the median;
(e) the first order statistic;
(f) $\bar{X}$.

8.9 The first order statistic Y_1, $\sum_1^n (X_i - Y_1)/n$.

8.11 $w_1(y)$.

8.12 $b = 0$; does not exist.

8.13 Does not exist.

8.15 $[y\tau^2 + \mu\sigma^2/n]/(\tau^2 + \sigma^2/n)$.

8.16 $\beta(y + \alpha)/(n\beta + 1)$.

CHAPTER 9

9.1 $\frac{1}{4} + \frac{3}{4}\ln\frac{3}{4}$; $\frac{7}{16} + \frac{9}{8}\ln\frac{3}{4}$.

9.2 $\frac{11}{64}$, $(31)3^8/4^9$.

9.5 $n = 19$ or 20.

9.6 $K(\frac{1}{2}) = 0.062$;
$K(\frac{1}{12}) = 0.920$.

9.10 $\sum_1^{10} x_i^2 \geq 18.3$; yes; yes.

9.12 $3\sum_1^{10} x_i^2 + 2\sum_1^{10} x_i \geq c$.

9.13 95 or 96; 76.7.

9.14 38 or 39; 15.

9.15 0.08; 0.875.

9.16 $(1-\theta)^9(1+9\theta)$.

9.17 $1, 0 < \theta \leq \frac{1}{2}$; $1/(16\theta^4)$, $\frac{1}{2} < \theta < 1$; $1 - 15/(16\theta^4)$, $1 \leq \theta$.

9.19 53 or 54, 5.6.

9.22 Reject H_0 if $\bar{x} \geq 77.564$.

9.23 26 or 27;
reject H_0 if $\bar{x} \leq 24$.

9.24 220 or 221;
reject H_0 if $y \geq 17$.

9.26 $\sum_1^n x_i \leq c$ or $\min(x_i) < \theta_1'$; yes.

9.29 $c_0(n) = (14.4) \times (n \ln 1.5 - \ln 9.5)$;
$c_1(n) = (14.4) \times (n \ln 1.5 + \ln 18)$.

9.30 $c_0(n) = [0.05n - \ln 8]/\ln 3.5$;
$c_1(n) = [0.05n + \ln 4.5]/\ln 3.5$.

9.32 (b) $c = 0.18$; 0.64.
(c) $c = 0.5$; 0.16; 0.84.
(d) $c = 0.23$; 0.06; 0.68.

CHAPTER 10

10.1 $t = 3 > 2.262$, reject H_0.

10.2 $|t| = 2.27 > 2.145$, reject H_0.

10.11 $q_3 = \frac{176}{21} > 7.81$, reject H_0.

10.13 $q_3 = \frac{22}{9} < 11.3$, accept H_0.

10.14 $6.4 < 9.49$, accept H_0.

10.16 $\hat{p} = (X_1 + X_2/2)/(X_1 + X_2 + X_3)$.

10.21 $F = 7.6 > 3.89$, reject H_0.

10.24 $r + \theta$, $2r + 4\theta$.

10.25 $r_2(\theta + r_1)/[r_1(r_2 - 2)]$, $r_2 > 2$.

10.28 $F = 18.7 > 5.14$, reject H_0.

10.31 $F = 27.7 > 4.76$, reject H_0.
10.38 $\hat{\beta} = \sum (X_i/nc_i)$,
$\sum [(X_i - \hat{\beta}c_i)^2/nc_i^2]$.
10.41 Reject H_0.
10.42 $\sum X_i$, $\sum Y_i$, $\sum X_i^2$, $\sum Y_i^2$.
10.43 $g(r) = \frac{1}{2}$, $r = -1, 1$.

CHAPTER 11

11.1 $(e^t - 1)/t$, $t \neq 0$; 1, $t = 0$.
11.4 (a) $\frac{15}{16}$; (b) 675/1024;
(c) $(0.8)^4$.
11.6 8.
11.11 0.067.
11.13 Reject H_0.
11.20 0; $4(4^n - 1)/3$; no.
11.30 $\frac{2}{99}$.

CHAPTER 12

12.3 (a) $\exp\{(t_1\mathbf{c} + t_2\mathbf{d})'\boldsymbol{\mu} + [(t_1\mathbf{c} + t_2\mathbf{d})'\mathbf{V} \times (t_1\mathbf{c} + t_2\mathbf{d})]/2\}$.
12.12 $a_i = 0$, $i = 1, 2, 3, 4$.
12.13 $\sum_{j=1}^{n} a_{ij} = 0, i = 1, 2, \ldots, n$.

Index

Analysis of variance, 327, 333
Approximate distribution(s), chi-square, 308
  normal for binomial, 185
  normal for chi-square, 180
  normal for Poisson, 181
  of $\overline{X}$, 182
  Poisson for binomial, 179

Bayes' formula, 61
Bayesian statistics, 207, 261, 293
Bernstein, S., 83
Beta distribution, *see* Distribution
Binary statistic, 364
Binomial distribution, *see* Distribution
Bivariate normal distribution, *see* Distribution
Boole's inequality, 347

Cauchy distribution, *see* Distribution
Central limit theorem, 182
Change of variable, *see* Technique, Transformation
Characterization, 154, 162
Chebyshev's inequality, 55, 89, 177
Chi-square distribution, *see* Distribution
Chi-square test, 308, 357, 365
Cochran's theorem, 393
Comparisons, *see* Multiple comparisons
Completeness, 227, 229, 232, 239, 240, 243
Compounding, 211
Conditional expectation, 65, 67
Conditional probability, 57, 64
Conditional p.d.f., 63, 64, 66, 67, 112
Confidence coefficient, 194
Confidence interval, 193, 197
  for difference of means, 199, 201
  for means, 194, 196
  for p, 197, 202
  for quantiles, 352
  for ratio of variances, 205
  for regression parameters, 338
  for variances, 202
Contingency tables, 313
Contrasts, 346
Convergence
  in probability, 177
  stochastic, 175, 186, 189
  with probability one, 178
Correlation coefficient, 69, 75, 339
Covariance, 69, 170
Covariance matrix, 383
Coverage, 354
Cramér, 179
Critical region, 266, 269, 272
  best, 272, 274, 278
  size, 269
  uniformly most powerful, 281
Curtiss, J. H., 179

Decision function, 258, 290
Degrees of freedom, *see* Distribution, chi-square

Distribution, beta, 134, 139, 352
binomial, 86, 179, 187, 352
bivariate normal, 111, 160
Cauchy, 134
chi-square, 102, 109, 159, 180, 308, 316, 385, 387
conditional, 63, 64, 67, 215
continuous type, 23, 24
degenerate, 53, 173
Dirichlet, 141, 355
discrete type, 22, 24
exponential class, 231, 239
function, 29, 32, 34, 119
gamma, 99
geometric, 90
hypergeometric, 40
limiting, 171, 186, 308, 362
marginal, 61, 66
multinomial, 91, 92, 309, 377
multivariate normal, 164, 308, 379
negative binomial, 90, 211
normal, 104, 107, 158, 182
of coverages, 355
of F, 137, 305, 319
of functions of random variables, 33, 116, 166
of $F(X)$, 349
of linear functions, 158
of noncentral chi-square, 314, 325, 387
of noncentral F, 327, 333, 334
of noncentral T, 304
of nS^2/σ^2, 165
of order statistics, 145
of quadratic forms, 316, 384
of R, 341
of runs, 368
of sample, 119
of T, 135, 211, 303, 338, 341
of $\overline{X}$, 160, 163
Poisson, 94, 179
posterior, 209
prior, 209
trinomial, 91
uniform, 37, 349
Distribution free, 348, 357

Efficiency, 250
Estimation, 190, 212
Bayesian, 207, 261
interval, *see* Confidence interval
maximum likelihood, 254
minimax, 260
point, 213, 248
Events, 2, 15
exhaustive, 14
mutually exclusive, 14
Expectation (expected value), 41, 42, 79, 166
of a product, 79
Exponential class, 231, 239

F distribution, *see* Distribution
Factorization theorem, 219, 221, 238
Family of distributions, 212
Fisher-Neyman criterion, 216, 238
Frequency, 2, 310
relative, 2, 11, 89
Function, characteristic, 52
decision, 258, 261
distribution, 29, 32, 34, 119
exponential probability density, 231, 239
gamma, 99
likelihood, 254
loss, 258, 261
moment-generating, 47, 49, 73, 82, 154
of parameter, 234
of random variables, 33, 42, 116, 166
point, 8
power, 266, 269, 376
probability density, 22, 23, 24, 29
probability distribution, 12
probability set, 11, 16, 32

risk, 258, 261, 291
set, 8

Hypothesis, *see* Statistical hypothesis, Test of a statistical hypothesis

Independence, *see* Stochastic independence
Inequality, Boole, 347
Chebyshev, 54, 89, 177
Rao-Blackwell, 223
Rao-Cramér, 248
Interaction, 334
Interval estimation, *see* Estimation

Jacobian, 127, 128, 139, 143
Joint conditional distribution, 67
Joint distribution function, 61
Joint probability density function, 61

Law of large numbers, 89, 170, 178
Lehmann alternative, 378
Lehmann-Scheffé, 229
Lévy, P., 179
Liapounov, 362
Likelihood function, 254
Limiting distribution, 172, 182, 187, 308, 362
Limiting moment-generating function, 179
Linear functions, covariance, 384
mean, 168
moment-generating function, 161, 383
variance, 168
Loss function, 258, 261

Madow, W. G., 258
Mann-Whitney, 371
Marginal p.d.f., 62, 66, 67
Maximum likelihood, 254
method of, 254
statistic, 254, 256
Mean, 46, 118
conditional, 70, 71, 112
of a sample, 118
of linear function, 168
of X, 168
Median, 28, 152
Midrange, 152
Minimax, criterion, 260
decision function, 260, 290
Mode, 28, 93
Moment-generating function, 47, 49, 73, 80, 82, 179
of binomial distribution, 87
of bivariate normal distribution, 113
of chi-square distribution, 102
of gamma distribution, 101
of multinomial distribution, 92
of multivariate normal distribution, 382
of normal distribution, 105
of Poisson distribution, 96
of $\overline{X}$, 161
Moments, 50, 53, 94
method of, 253
Multinomial distribution, *see* Distribution
Multiple comparisons, 343, 347
Multiplication rule, 59, 60
Multivariate normal distribution, *see* Distribution

Neyman-Pearson theorem, 274
Noncentral chi-square, *see* Distribution
Noncentral F, *see* Distribution
Noncentral parameter, 304, 326, 327
Noncentral T, *see* Distribution
Nonparametric, 348, 357
Normal distribution, *see* Distribution

Normal scores, 364

Order statistics, 145, 351
  distribution, 145
  functions, 152, 351

Parameter, 86
Parameter space, 212
Percentile, 28
Poisson distribution, *see* Distribution
Poisson process, 95, 100
Power, *see* Function, Test of a statistical hypothesis
Probability, 2, 11, 31, 38
  conditional, 63, 64
  induced, 15, 58
  measure, 2, 12, 144
  models, 24, 36, 59
  subjective, 209
Probability density function, 22, 23, 24, 25, 29
  conditional, 63, 64, 67
  exponential class, 231, 239
Probability set function, 11, 16, 32

Quadratic forms, 316
  distribution, 316, 384
  independence, 316, 391
Quantiles, 28, 348
  confidence intervals for, 352

Random, experiment, 1, 11, 38, 193
  interval, 190, 196, 344
  sample, 118, 160
  sampling distribution theory, 119
Random variable, 15, 33
  continuous-type, 23, 24
  discrete-type, 22, 24
  mixture of types, 33
  space of, 15, 17, 18, 25
Range, 152
Rao-Blackwell theorem 223, 225
Rao-Cramér inequality, 248

Regression, 335
Relative frequency, 2, 11, 89
Risk function, 258, 261, 291

Sample, correlation coefficient, 339
  mean of, 118
  median of, 152
  random, 118, 160, 241
  space, 1, 11, 193, 310
  variance, 118
Scheffé, H., 344
Set, 3
  complement of, 6
  element of, 3
  function, 8
  null, 4, 12
  of discrete points, 22
  probability measure, 2, 12
  subset, 4, 12
Sets, algebra, 3
  intersection, 5
  union, 4
Significance level of test, 269
Space, 6
  of random variables, 15, 17, 18, 25, 193
  parameter, 212, 300
  product, 76
  sample, 1, 11, 57, 193
Standard deviation, 47
Statistic, 116
  best, 213, 229
  consistent, 252
  efficient, 250
  maximum likelihood, 254, 256
  minimum mean-square-error, 260
  order, 145, 351
  unbiased, 213, 236
  see also Sufficient statistic(s)
Statistical hypothesis, 265, 268
  alternative, 265
  composite, 268
  simple, 268
  test of, 266

Statistical inference, 193
Stochastic convergence, 175, 186, 189
Stochastic dependence, 76, 77
Stochastic independence, 76, 77, 80, 114, 133, 243, 306, 316, 339, 391
  mutual, 81, 82, 83
  of linear forms, 162
  of quadratic forms, 316, 388
  of $\overline{X}$ and S^2, 163, 245
  pairwise, 83
  test of, 339
Sufficient statistic(s) 213, 216, 222, 225, 232, 243, 256, 283
  function, 222
  joint, 237, 240

T distribution, *see* Distribution
Technique, change of variable, 121, 123, 125, 128, 139
  distribution function, 119
  moment-generating function, 154
Test of a statistical hypothesis, 269, 272
  Bayesian, 293
  best, 272, 274, 278
  chi-square, 308, 357, 365
  likelihood ratio, 297, 300
  median, 366
  minimax, 291
  of equality of distributions, 365
  of equality of means, 301, 321
  of equality of variances, 305
  of randomness, 370
  of stochastic independence, 313
  power of, 266, 269
  run, 367
  sequential probability ratio, 285
  sign, 357
  significance level, 269
  uniformly most powerful, 281
Tolerance interval, 353
Transformation, not one-to-one, 141
  of continuous-type variables, 125, 128, 139
  of discrete-type variables, 121, 123
  one-to-one, 122, 126, 128
Types I and II errors, 271, 208, 292

Uniqueness, of best statistic, 229
  of characteristic function, 52
  of moment-generating function, 48

Variance, analysis, 327, 333
  conditional, 65, 72
  of a distribution, 46
  of a linear function, 168
  of a sample, 118
  of $\overline{X}$, 168
Venn diagram, 5

Wilcoxon, 359, 371