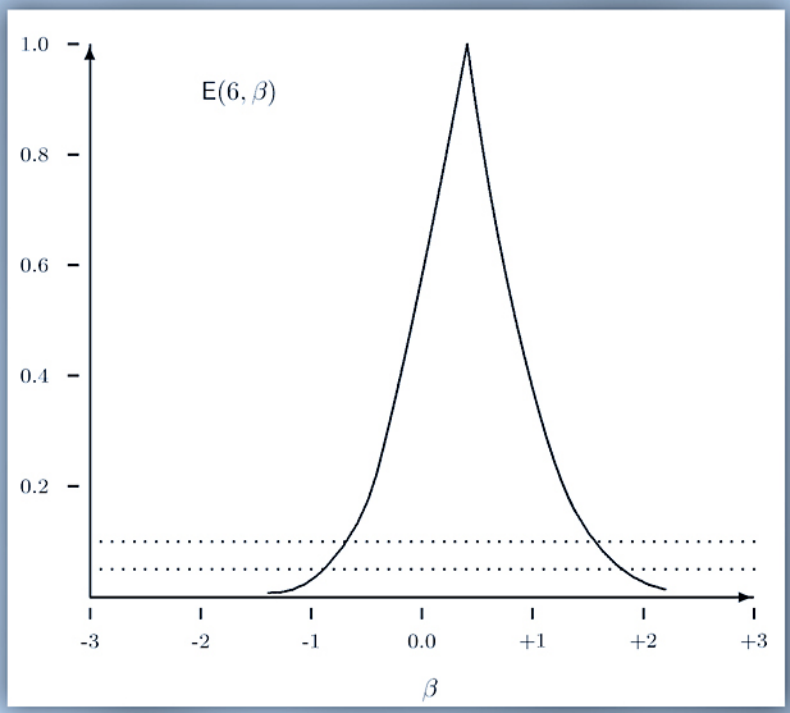
EXACT ANALYSIS OF DISCRETE DATA



Karim F. Hirji

Chapman & Hall/CRC
Taylor & Francis Group

EXACT ANALYSIS OF DISCRETE DATA

EXACT ANALYSIS OF DISCRETE DATA

Karim F. Hirji

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Published in 2006 by
Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-58488-070-8 (Hardcover)
International Standard Book Number-13: 978-1-58488-070-7 (Hardcover)
Library of Congress Card Number 2005053883

Library of Congress Cataloging-in-Publication Data

Hirji, Karim F.
Introduction to the exact analysis of discrete data / Karim F. Hirji.
p. cm.
Includes bibliographical references and index.
ISBN 1-58488-070-8 (alk. paper)
1. Discrete groups. 2. Computer science--Mathematics. I. Title.

QA178.H57 2005
512'.2--dc22 2005053883

informa
Taylor & Francis Group
is the Academic Division of Informa plc.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

For

Emma, Rosa, Farida and Sakarma

Contents

List of figures

List of tables

Abbreviations

ACL	:	actual coverage level
ASL	:	actual significance level
CAM	:	Cochran–Armitage–Mantel
cdf	:	cumulative distribution function
CI	:	confidence interval
cmle	:	conditional maximum likelihood estimate
cmue	:	conditional median unbiased estimate
COR	:	common odds ratio
CS	:	conditional score
CT	:	combined tails
df	:	degrees of freedom
EC	:	exponent check(s)
FH	:	Freeman–Halton
gp	:	generating polynomial
HOR	:	heterogeneous odds ratio
iid	:	independent and identically distributed
iip	:	independent with identical parameters
lhs	:	left hand side
LR	:	likelihood ratio
max	:	maximum
MH	:	Mantel–Haenszel
min	:	minimum
mle	:	maximum likelihood estimate
mue	:	median unbiased estimate
rhs	:	right hand side
PBD	:	polynomial based distribution
RAM	:	random access memory
RCT	:	randomized controlled trial
RBG	:	Robins–Breslow–Greenland
RPM	:	recursive polynomial multiplication
SMR	:	standardized mortality ratio
TBD	:	Tarone–Breslow–Day
TOR	:	trend in odds ratios
TST	:	twice the smaller tail

Foreword

Researchers in biology, medicine, public health, psychology, sociology, law and economics regularly encounter variables that are discrete or categorical in nature. And there is no dearth of books - elementary to advanced - on the methods of analysis and interpretation of such data. These books mainly cover large sample methods. When the sample size is not large, or the data are otherwise sparse, their accuracy is suspect. In that event, methods not based on asymptotic theory, called exact methods, are desirable.

The origins of the exact method for analysis of discrete data lie in the early analysis of binomial and Poisson outcomes, and is related to the growth of nonparametric analysis of continuous data. Its emergence as a distinct branch of statistics, however, dates to the works of Sir Ronald A. Fisher, Frank Yates and James O. Irwin (Fisher 1934, 1935b; Irwin 1935; Yates 1934). All of them worked on the exact analysis of a 2×2 table. Despite the early start, the forms of discrete data for which exact analysis was feasible were, until 1980, rather limited. The analysis was mostly restricted to univariate data, an equiprobable multinomial, or a 2×2 table. For complex models, the exact analysis was formulated theoretically, and creative applications of the Monte–Carlo method were devised. Yet, it was not a practical option since the computational effort rose exponentially with sample size. In fact, more energy was expended in the periodic controversies about exact analysis than on expanding its scope.

A dramatic change occurred during the 1980s. Application of fast Fourier transform based techniques by Marcello Pagano, David Tritchler and their coworkers, on the one hand, and of network algorithms by Cyrus Mehta, Nitin Patel and their coworkers, on the other, were principally behind that turnaround. I had the fortune to be associated with the latter group, and worked on extending exact analysis of logistic models and the multivariate shift algorithm. Today, such efficient algorithms and enhanced computing power have made exact analysis eminently feasible for a vastly enlarged spectrum of discrete data problems. As such, it is often performed when the traditional large sample methods are in question. (The references relevant to this paragraph are noted later in the text.)

Yet, despite the plethora of original research and review papers on the subject, and the existence of computer software, no book with exact analysis of discrete data as its prime focus is currently available. Most books on discrete or categorical data analysis tend to have a small section on exact tests. The Fisher exact test for a 2×2 table is always covered. Recent books contain more material on the topic. The first book with the word "exact" in its title has one half of a chapter on exact tests for discrete data (Weerahandi 1995). Even when they cover more complex problems, the books essentially present exact methods as a set of recipes. None develops them from first principles, covers a broad class of models, gives a variety of worked examples, addresses the conceptual issues and also presents related computational algorithms, all within an integrated yet accessible framework.

This book begins the task of filling the void. My aim has been to present, in a unified but elementary and applications-oriented framework, the distributional theory, statistical methods and computational methods for exact conditional analysis of discrete data. To shape it into

a sturdy and coherent edifice, I have relied on two key ideas, namely, that of a polynomial based distribution and an evidence function. Their roots lie in the pioneering work of Sir R. A. Fisher. His analysis of the odds ratio in a 2×2 table employed the conditional hypergeometric probability. For this purpose, he formulated the tail probability as a ratio of two polynomials in which the numerator was a segment of the denominator. The specific example in Fisher (1935b) was:

$$\mathsf{F}(\psi) \;=\; \frac{1 + 102\psi + 2992\psi^2}{1 + 102\psi + \cdots + 476\psi^{12}}$$

Fisher used this ratio of polynomials to assess a significance level for testing if the odds ratio was equal to one, and determine what we now call 99% and 95% upper confidence bounds for it. A two-sided version of the basic idea, which is relevant to continuous data and large sample analysis as well, was later elaborated in a generalized form by Birnbaum (1961). Allan Birnbaum named this entity a confidence curve; the Fisher polynomial ratio then is but a one-sided confidence curve.

The confidence curve, or in our terminology, the evidence function is, after a long hiatus, steadily showing up in the texts on statistics and epidemiology; a recent case in point is Rothman (2002), where it occupies a prominent position. Over the years, it has been given several different names, based on the aspect emphasized. I have chosen to call it an evidence function, a name that captures the overall spirit of what it stands for. No matter what the name, its primacy lies in that it embodies the three key tools of data analysis, namely, a *p*-value, a point estimate and a confidence interval within a single construct. Thereby, it serves a positive pedagogic purpose and allows a conceptually unified presentation of the results of data analysis.

The importance of the polynomial form for distributions lies in that not only the hypergeometric but other common discrete distributions like the binomial, Poisson, negative binomial, multinomial, product multinomial as well as many of the distributions derived from them are polynomial based distributions. The polynomial formulation is not rare; it is implicit in several classic works on discrete data analysis, e.g., Cox (1970) and Zelen (1971), and in the run of the mill old and new papers like Bennett and Nakamura (1964) and Emerson (1994), to name a few. But, apart from a specialized and rarely cited branch of research, its primacy was not noted, and, until recently, its computational utility was not appreciated.

This book utilizes the research published in the 1990s showing that the polynomial formulation produces an integrated framework for exact inference for discrete data, both in terms of theory and computation. It links the many algorithms in the field and, unlike the other formulations, it is based on a simple idea. In fact, the basic idea underlying it is not too distinct from the high school algebra method of multiplying a set of polynomials in a step by step fashion and selecting some terms from the product.

My research on polynomial multiplication algorithms for exact analysis began in 1988. On my own and with Stein E. Vollset, Isildinha Reis, Man Lai Tang and Timothy Johnson - my doctoral students at UCLA - I wrote a number of papers on exact analysis using such algorithms. The idea was also independently developed by David O. Martin and Harland Austin. Further, the pioneering ideas of Cyrus Mehta and Nitin Patel in a related algorithmic area are well reflected in the polynomial algorithms based literature. Two other papers, Baglivo, Pagano and Spino (1996) and van de Wiel, Di Bucchianico and van der Laan (1999), contain equivalent formulations of the same basic idea. (The other references relevant to this paragraph are noted later in the text.)

Thus far this material has been buried in specialized journals. It is time a wider audience of

students, data analysts, applied researchers and statisticians has access to it. I hope this book will serve that purpose.

****

The first chapter reviews relevant discrete distributions, lays out the notation and defines key concepts. Apart from Chapter 4 and Chapter 7, the chapters which follow, up to and including Chapter 10, develop and illustrate exact conditional methods for various models for discrete data. Of these, Chapters 2, 3 & 5 cover simple one and two variable models. Chapters 6, 8, 9 & 10 deal with several 2×2 tables, and one and several $2 \times K$ tables. Chapters 4, 11 & 12, on the other hand, deal with computational techniques needed for implementing the exact method. Chapters 13 and 14 cover statistical and computational material in an integrated fashion, respectively dealing with multinomial data, and matched and dependent data. Chapter 7 is on assessing the tools of inference, and Chapter 15 deals with conceptual matters, and addresses the use and misuse of exact methods in practice. The relevant large sample methods are presented throughout the text.

Readers and teachers seeking a self-contained but elementary exposure to the field may study Chapters 1, 2, 3, 5, 6, parts of 7, 8, 9 and 10. Chapters 4, 11 and 12, and parts of Chapters 13 and 14 are of special importance to those who also want to explored the computational techniques underlying exact analysis.

This is an introductory work. A basic course in statistics, biostatistics or research methods (at the level say of Rosner (2000) or Daniel (2005)) is all that is needed to access most of its material. It developed from my notes for a course on discrete data analysis for masters and doctoral students in biostatistics at UCLA. I hope that graduate students, teachers and practioners in statistics and biostatistics as well as quantitatively inclined researchers in fields like epidemiology, genetics, sociology, education, psychology and business studies will find it of value, both as a learning and teaching text, and a reference work.

I have avoided a cookbook approach. Instead, the analytic methods are developed in a step by step manner from first principles. Yet, elaborate theory is absent. Each chapter contains relevant worked examples and exercises. Some of them can be taken up as class research projects by students. Relevant material on computer implementation is also included.

This book is not related to any existing statistical software. Those who have access to software such as StatXact or SAS, which cover exact methods, can use them with the book. Many of the illustrative examples can be worked out on a calculator, or with simple programming in common software.

****

Over the years, the field of exact analysis has been fraught with much controversy. In essence, the debates were about the role of conditioning in, and definition of the frame of reference for, data analysis. They as such pertain to all forms of statistical analysis, including the traditional large sample methods. What is called an exact analysis can actually be an unconditional, a partly conditional or a fully conditional analysis. In discrete data analysis, moreover, there are distortions of the form not found in the analysis of continuous data. But these, in different ways no doubt, afflict both approximate and exact methods. Yet mainly because of the context within which the debates were aired, an impression has been created that such controversies are inextricably tied to the exact method. At times, some leading statisticians have not helped the matter by not sufficiently disentangling the separate strands of the arguments. To some, exact

analysis is, as we illustrate in Chapter 15, synonymous with contentious analysis while some others swear by it!

The term "exact" also has had a variety of meanings in the statistical literature. We discuss these issues in the text. But for now, given the historical baggage attached to it, we state that to us the label "exact" surely does not imply that the method in question is the "correct" or "best" one in any absolute sense. It just refers to a method that avoids large sample approximations. In that spirit, we see and present exact methods as an integral part of the spectrum of data analytic techniques available today. After noting their advantages and limitations, we hold that the former often outweigh the latter. Exact analysis is, in our view, not just a valid, but often a better option, for many sparse data problems.

Acknowledgments

This book has benefitted in many ways from the participation of, and valuable suggestions from Dr. Stein E. Vollset, University of Bergen, Norway. He was associated with the initial stages of developing its material, and is a coauthor of Chapter 4. His wisdom, though, prevails in many parts of the work. I am very grateful to him for all his help. Dr. Elliot Landaw, Department of Biomathematics, UCLA and Dr. Roshan Bastani, UCLA School of Public Health, kindly allowed me access to their departmental facilities and resources. Karla Morales helped with the literature search and securing permissions for copyrighted material. My four doctoral students at UCLA (named above) played a key part in developing some of the ideas contained in this work. All of them also have my gratitude.

Dr. Mahmood F. Hameer generously provided data from his research; these have been used for illustrative purposes in several chapters. And Rob Calver, my editor at CRC Press, professionally expedited the various stages of production of this book.

And above all, without the sustained support of Farida, my wife, this book could not have seen the light of the day. Her love and patience are more than any one can ask for. To her, I only say: 'I love you.'

CHAPTER 1

Discrete Distributions

1.1 Introduction

When the variables of a **scientific model** are related through a chance mechanism, we call it a **statistical model.** This book covers statistical models for discrete variables. Its main focus is on the methods for analysis of discrete variable statistical models called exact conditional methods. This chapter prepares the groundwork. This includes introducing a unified framework for representing common discrete probability distributions, defining key concepts and describing the properties of these distributions. The specific aims here are:

- To state the properties of the binomial, Poisson, negative binomial, hypergeometric and related distributions.
- To introduce the multinomial distribution, a distribution that forms the basis of many statistical models for discrete data.
- To present the polynomial formulations for common univariate and multivariate discrete probability distributions.
- To introduce three essential ideas for exact analysis of discrete data, namely, the generating polynomial, conditioning and sufficient statistics.

We do not develop the probability theory for discrete random variables in a rigorous manner. The proofs for the results stated are also not all given. Many introductory probability theory texts contain such material. However, results that are not usually found in standard texts, but which are especially relevant for this book are accorded due elaboration.

1.2 Discrete Random Variables

Variability is an inescapable fact of nature and life. One child gets an earache frequently but another hardly ever suffers from the malady. The symptoms of a common cold may clear up in a few days, or may linger on for weeks. Will this child have an earache in the next six months? How long will a cold persist? Even a trained physician is hard placed to make the prediction since events like these do not have the level of certainty we associate with the statement: "A stone released in midair will fall to the ground." It is fair to regard whether or not a child will get an earache, or the number of days a cold will last as random or chance phenomena.

Randomness does not mean complete unpredictability or chaos. Usually, random phenomena incorporate some systematic components as well. When a large number of cases of the event in question are examined, a pattern often emerges. For example, boys show a higher tendency than girls to develop ear infections. Or, that more than 50% of the cases of the common cold tend to resolve spontaneously within a week or so.

To study processes that are random at a micro-level but which exhibit some degree of regularity

when viewed on an aggregate scale, we use the idea of a **random variable**. A simple example of a random variable obtains from envisioning an event which either occurs or does not occur. Let π denote the probability of occurrence of the event. In that case,

$$0 \leq \pi \leq 1 \tag{1.1}$$

The probability of nonoccurrence of the event then is $1 - \pi$. We define an associated random variable, $\mathcal{Y}$, as follows. Let $\mathcal{Y} = 0$ if the event does not occur (called a failure), and $\mathcal{Y} = 1$ if it does (called a success). Then we write

$$\mathrm{P}[\,\mathcal{Y} = 1\,] \;=\; \pi \quad \text{and} \quad \mathrm{P}[\,\mathcal{Y} = 0\,] \;=\; 1 - \pi \tag{1.2}$$

A compact way of writing this is: For $y \in \{0, 1\}$,

$$\mathrm{P}[\,\mathcal{Y} = y\,] \;=\; \pi^{y}(1-\pi)^{1\,-\,y} \tag{1.3}$$

$\mathcal{Y}$ is called a **binary random variable**, and the random process is referred to as a **Bernoulli trial**. Examples include the outcome of a coin toss, cure or failure in treatment for a disease, the presence or absence of a genetic trait, exposure or nonexposure to a potential occupational carcinogen, and the presence or absence of an ear infection. Equation 1.3 is also a simple statistical model.

A random variable that assumes a finite, or at most, a countable number of values is called a **discrete random variable**. The number of days an episode of the common cold persists is an example of a discrete random variable.

Let the symbol $\mathcal{T}$ designate a discrete random variable. The set of values, realizations, or outcomes of $\mathcal{T}$, denoted by Ω, is the **sample space** or **support set** of $\mathcal{T}$. We assume throughout that Ω is a countable set of real numbers, or vectors with real elements, that is, it can be put into a one-to-one correspondence with a subset of the set of integers.

We let the function $\mathrm{f}(t) = \mathrm{P}[\,\mathcal{T} = t\,]$ represent the probability of an outcome $t \in \Omega$. This **probability function** has to satisfy the following basic properties:

Property 1.1: For any $t \in \Omega$, $\quad \mathrm{P}[\,\mathcal{T} = t\,] > 0$

Property 1.2: If $\Omega_1, \cdots, \Omega_k$ are mutually exclusive subsets of Ω, then

$$\mathrm{P}[\,\mathcal{T} \in \Omega_1 \bigcup \cdots \bigcup \Omega_k\,] \;=\; \sum_{j\,=\,1}^{k} \mathrm{P}[\,\mathcal{T} \in \Omega_j\,] \tag{1.4}$$

for $k = 1, \cdots, \infty$.

Property 1.3:

$$\sum_{t \,\in\, \Omega} \mathrm{P}[\,\mathcal{T} = t\,] \;=\; 1 \tag{1.5}$$

As indicated by Property 1.1, throughout this book we consider events with strictly nonzero probabilities as the relevant subsets of Ω.

****

We consider two scientific investigations with discrete random variables.

Example 1.1: Kiviluoto et al. (1998) report a clinical trial of two surgical procedures, labeled LC and OC, for acute cholecystitis. One outcome of interest was the occurrence of major complications (including death) after surgery. The relevant data from this trial are shown in Table 1.1.

Table 1.1 *Post Surgical Complications*

Major Complications	Surgical Procedure	
	LC	OC
No	32	24
Yes	0	7
Total	32	31

Source: Kiviluoto et al. (1998), Table 3.

The allocation of patients to treatment was done using a random device. Table 1.1 relates two binary random variables, the treatment, and the onset of major complications after treatment. What conclusion can we draw about the complication rates of these treatments?

Example 1.2: Next consider Table 1.2 with data on the status of young children in the pediatrics ward of a hospital in Tanzania. All these children had Kwashiorkor. The main aim of the study was to evaluate the change in the erythrocyte sedimentation rate (ESR) level by type of infection in children with malnutrition. Other children in the study had marasmus and marasmic kwashiorkor, two other forms of malnutrition. The data for these children are given in Table 13.5 and Table 13.9, respectively, and are also summarized at other places in this text.

All the three variables in Table 1.2 are in a discretized form. None is a binary variable. This was a cross sectional type of study; at the time of examination, the age and infection status were known, and the ESR value was the random entity. The latter was converted into a discrete variable using common clinical cut points. After adjusting for the effect, if any, of age, does the ESR profile vary by infection status? - that was a key question of interest.

At this juncture, we emphasize that the above data examples and all the other data examples in this book are given for the sole purpose of illustrating and comparing statistical methods. **They are not used to draw substantive conclusions about the underlying biomedical or other substantive issues**. The latter has to be based on the full data set for the appropriate study.

These examples show us that discrete variables come in various forms. Some are **nominal** variables, also known as **nominal scale**, **qualitative**, or **categorical** variables. Their levels are devoid of any intrinsic order. A case in point is the variable 'type of infection' in Table 1.2 which has three unordered levels. As another example, the primary diagnosis of a hospitalized patient may be classified as cardiac, lung, liver, renal or other disease. For record keeping, we may attach numeric labels such as 1, 2, 3, 4 or 5 to these disease states. But they are arbitrary labels. When infection status is analyzed at two levels 'infection' or 'no infection,' it becomes a binary variable.

Table 1.2 *ESR and Pulmonary Infection by Age in Kwashiorkor*

	ESR Level			
Age $\leq$ 12 Months				
	≤ 10	11–25	26–99	100+
No Infection	12	7	3	0
Pulmonary Tuberculosis	0	0	2	1
Other Pneumonia	4	0	0	0
12 Months $<$ Age $\leq$ 24 Months				
	≤ 10	11–25	26–99	100+
No Infection	11	1	0	0
Pulmonary Tuberculosis	0	0	1	1
Other Pneumonia	3	0	0	0
Age $>$ 24 Months				
	≤ 10	11–25	26–99	100+
No Infection	3	1	1	0
Pulmonary Tuberculosis	1	0	0	1
Other Pneumonia	0	0	1	0

Source: Hameer (1990); Used with permission.

A discrete variable whose levels have a built-in order is called an **ordinal** variable. The variables 'age group' and 'ESR level' in Table 1.2 are of this type. To take another case, the side effects of a medicinal drug may be depicted as: none, not life threatening and life threatening. The second level is more serious than the first, and the third is more serious than the second. Despite the ordering, no quantitative magnitude is attached to any level. Some ordinal variables come with a numeric score attached to each level. The number of side effects from a drug is a case in point. We call an ordered discrete variable which has a numeric score as a **scored** variable. The ESR categories in Table 1.2 (≤ 10, $11-25$, $26-90$ and $100+$) are generally viewed as normal, elevated, highly elevated and very highly elevated levels. They may respectively be assigned **natural** scores 0, 1, 2 and 3 in an analysis. Or, the scores may be the midpoints of the associated numeric range.

For the purpose of data analysis, the same underlying entity is at times deemed a nominal variable and at other times, an ordinal variable. The age of a subject, measured in years or months, as in Table 1.2, is often treated as a discrete variable. Then 'age group' may be used as a nominal variable or as a discrete variable with a numeric score.

Categorizing or scoring a variable may imply a loss of information. If not done with care, it has

the potential to mislead. At times, it may simplify data analysis and interpretation, and at times, it yields valuable insights into nonlinear relationships. Good science requires that discretization or categorization be done at the stage of study design and not after the data are collected and summarized. Further, it should not be done in a way that deviates drastically from the underlying substantive meaning of the variable in question.

Table 1.1 and Table 1.2 are examples of **sparse data**. One or more of the cell counts in both are zero or small. Also, there is an imbalance in the counts at the levels of some variables. Such data occur even with large sample sizes. The methods given in this text are particularly suitable for sparse data.

1.3 Probability Distributions

The **cumulative distribution function (cdf)** of the random variable $\mathcal{T}$ is defined as

$$\mathsf{F}(t) = \mathrm{P}[\,\mathcal{T} \leq t\,] \tag{1.6}$$

The **mean** (or **expectation**) and the **variance** of $\mathcal{T}$ respectively are

$$\mu \;=\; \mathbf{E}[\,\mathcal{T}\,] \quad = \quad \sum_{t\,\in\,\Omega} t\mathrm{P}[\,\mathcal{T} = t\,] \tag{1.7}$$

$$\sigma^2 \;=\; \mathbf{E}[\,(\mathcal{T}-\mu)^2\,] \quad = \quad \sum_{t\,\in\,\Omega} (t-\mu)^2\mathrm{P}[\,\mathcal{T} = t\,] \tag{1.8}$$

The mean is a measure of location and the variance indicates the extent of variability around the mean. The **median** is another measure of location. Conceptually, it is the value of the middle item of a population when all the items are arranged in an increasing order. In a population of discretized values, such a middle item may not be uniquely identifiable. For a discrete random variable taking two or more values, the median is defined as follows:

- Suppose there exists a $t_l \in \Omega$ such that $\mathrm{P}[\,\mathcal{T} \leq t_l\,] = 0.5$. Then $t_r \in \Omega$ is the value such that $\mathrm{P}[\,\mathcal{T} \geq t_r\,] = 0.5$. In this case, the median is not unique, and may be taken as t_l or t_r. In practice, we let $t_m = (t_l + t_r)/2$.
- Suppose there does not exist a $t \in \Omega$ such that $\mathrm{P}[\,\mathcal{T} \leq t\,] = 0.5$. Then the median is the unique value $t_m \in \Omega$ such that $\mathrm{P}[\,\mathcal{T} < t_m\,] < 0.5$ and $\mathrm{P}[\,\mathcal{T} \geq t_m\,] > 0.5$.

Other definitions also exist. The median, like the mean, is not necessarily a member of Ω. It is the preferred measure of location in distributions that are not symmetric. The mean has a key additive property. Suppose $\mathcal{T}_k$ is a random variable with mean μ_k, $k = 1, \cdots, K$. Then, for constants a_k, $k = 1, \cdots, K$,

$$\mathbf{E}[\,\sum_{k\,=\,1}^{K} a_k\mathcal{T}_k\,] \;=\; \sum_{k\,=\,1}^{K} a_k\mu_k \tag{1.9}$$

Unless otherwise specified, when we refer to a random variable from here on, we will mean a discrete random variable.

****

The **conditional probability** of the event Ω_1 given the event Ω_2 is defined by

$$\mathrm{P}[\,\mathcal{T} \in \Omega_1 \mid \mathcal{T} \in \Omega_2\,] \;=\; \frac{\mathrm{P}[\,\mathcal{T} \in \Omega_1 \bigcap \Omega_2\,]}{\mathrm{P}[\,\mathcal{T} \in \Omega_2\,]} \tag{1.10}$$

provided $\mathrm{P}[\,\mathcal{T} \in \Omega_2\,] > 0$. Any two events Ω_1 and Ω_2 are said to be **independent events** if the probability of any of them is not affected by whether the other has occurred or not. Formally, this means that

$$\mathrm{P}[\,\mathcal{T} \in \Omega_1 \mid \mathcal{T} \in \Omega_2\,] \;=\; \mathrm{P}[\,\mathcal{T} \in \Omega_1\,] \tag{1.11}$$

If Ω_1 and Ω_2 are independent events then it follows that

$$\mathrm{P}[\,\mathcal{T} \in \Omega_1 \bigcap \mathcal{T} \in \Omega_2\,] \;=\; \mathrm{P}[\,\mathcal{T} \in \Omega_1\,]\mathrm{P}[\,\mathcal{T} \in \Omega_2\,] \tag{1.12}$$

In particular, if the random variables $\mathcal{T}_1$ and $\mathcal{T}_2$ are independent, then

$$\mathrm{P}[\,\mathcal{T}_1 = t_1,\, \mathcal{T}_2 = t_2\,] \;=\; \mathrm{P}[\,\mathcal{T}_1 = t_1\,]\mathrm{P}[\,\mathcal{T}_2 = t_2\,] \tag{1.13}$$

for all t_1 and t_2.

If $\mathcal{T}_1$ and $\mathcal{T}_2$ are independent random variables with respective finite variances σ_1^2 and σ_2^2, and a_1, a_2 are some constants, then

$$\mathrm{var}[\,a_1\mathcal{T}_1 + a_2\mathcal{T}_2\,] \;=\; a_1^2\,\sigma_1^2 \,+\, a_2^2\,\sigma_2^2 \tag{1.14}$$

where $\mathrm{var}[\mathcal{T}]$ denotes the variance of $\mathcal{T}$.

1.4 Polynomial Based Distributions

A **polynomial based distribution**, denoted **PBD**, is defined as a probability distribution constructed from a polynomial. For example, consider the polynomial

$$f(\phi) \;=\; 2 \,+\, 7\phi^2 \,+\, 3\phi^5$$

with $\phi \geq 0$. Using the exponents of ϕ, we construct the sample space $\{0,2,5\}$. Then we define the random variable $\mathcal{T}$ on this sample space as

$$\begin{aligned}
\mathrm{P}[\,\mathcal{T} = 0;\, \phi\,] &= \frac{2}{f(\phi)} \\
\mathrm{P}[\,\mathcal{T} = 2;\, \phi\,] &= \frac{7\phi^2}{f(\phi)} \\
\mathrm{P}[\,\mathcal{T} = 5;\, \phi\,] &= \frac{3\phi^5}{f(\phi)}
\end{aligned}$$

This satisfies all the required properties for a probability distribution. In general, a polynomial based distribution is defined as follows. Let Ω be a countable set and, for each $u \in \Omega$, let $c(u) > 0$ be an associated coefficient. With a parameter $\phi \geq 0$, we then construct the polynomial

$$f(\phi) \;=\; \sum_{u\,\in\,\Omega} c(u)\phi^u \tag{1.15}$$

From this, we set up a discrete probability distribution for a random variable $\mathcal{T}$ as follows. For any $t \in \Omega$,

$$\mathrm{P}[\,\mathcal{T} = t;\,\phi\,] \;=\; \frac{c(t)\phi^t}{f(\phi)} \tag{1.16}$$

The generic polynomial $f(\phi)$ completely specifies this distribution. The set of its exponents constitutes the sample space, and for any point t in this space, the term with this exponent divided by the polynomial constitutes its probability.

Distribution (1.16) is expressed in an **exponential form** using the parameter $\beta = \ln(\phi)$ and the function

$$h(\beta) \;=\; f(\exp(\beta)) \;=\; \sum_{u\,\in\,\Omega} c(u)\exp(\beta u) \tag{1.17}$$

where $\exp(\beta) = e^\beta$. Then

$$\mathrm{P}[\,\mathcal{T} = t;\,\beta\,] \;=\; \frac{c(t)\exp(\beta t)}{h(\beta)} \tag{1.18}$$

In the statistical literature, distribution (1.16) is called a **Power Series Distribution**. (1.18) is the **exponential form** of the power series distribution. In this book, we prefer to use the less technical sounding and more affable name, namely, Polynomial Based Distribution.

The polynomial $f(\phi)$ is called the **coefficient generating function**, the **series function**, or simply the **generating function** of the PBD. We shall refer to it as the **generating polynomial**, or **gp** of the PBD.

In probability theory, the term generating function refers to functions like the **probability generating function**, or the **moment generating function**, or the **characteristic function** (Gordon 1997). Such functions facilitate the study of properties of probability distributions. For example, the mean and variance are at times easier to determine by using such a function.

In the case of a PBD, the generating polynomial $f(\phi)$ is an equivalent replacement for the more elaborate generating functions. This is one of the many advantages of using the polynomial form. In fact, $f(\phi)$ is an unnormalized version of the probability generating function and also completely specifies it (Exercise 1.33). As we proceed in this text, it will become clear that for exact analysis of many discrete data models, it is simpler and more natural to deal with the generating polynomial than with the other generating functions.

For emphasis, we reiterate the basic property that the gp of a PBD completely and uniquely specifies the distribution of $\mathcal{T}$ in the following sense.

`The Generating Polynomial`

- `The set of exponents of the generating polynomial of a PBD specifies the sample space, Ω.`
- `The probability of a point in this space equals the term of this polynomial with that exponent divided by the whole polynomial.`

The gp of a PBD can also be used to compute its mean and variance.

Theorem 1.1: Suppose $\mathcal{T}$ is a PBD variate with gp $f(\phi)$. If the mean, μ, and variance, σ^2, of $\mathcal{T}$ are finite, then

$$\mu = \phi \frac{d}{d\phi}[\ln f(\phi)] \tag{1.19}$$

$$= \frac{f_*(\phi)}{f(\phi)} = \frac{h'(\beta)}{h(\beta)} \tag{1.20}$$

$$\sigma^2 = \mu + \phi^2 \frac{d^2}{d^2\phi}[\ln f(\phi)] \tag{1.21}$$

$$= \frac{f_{**}(\phi)}{f(\phi)} - \left(\frac{f_*(\phi)}{f(\phi)}\right)^2 = \frac{h''(\beta)}{h(\beta)} - \left(\frac{h'(\beta)}{h(\beta)}\right)^2 \tag{1.22}$$

where

$$f_*(\phi) = \sum_{u \in \Omega} uc(u)\phi^u \quad \text{and} \quad f_{**}(\phi) = \sum_{u \in \Omega} u^2 c(u)\phi^u \tag{1.23}$$

further where $h'(\beta)$ and $h''(\beta)$ are the first two derivatives with respect to β of $h(\beta)$.

Proof: Consider the first portion rhs of the first equation.

$$\frac{d}{d\phi}[\ln f(\phi)] = \sum_{u \in \Omega} uc(u)\phi^{u-1}/f(\phi)$$

Multiplying by ϕ, this equals

$$\sum_{u \in \Omega} u\mathrm{P}[\mathcal{T} = u] = \mu$$

The other relations are proved similarly. □

In a majority of the applications we study, we need to combine a series of independent PBDs. The distribution of the sum of a set of random variables is called the **convolution** of these variables. It is not always easy to directly specify a convolved distribution. When the variables are independent, using the generating polynomial often provides an easier method.

Theorem 1.2: Let $\mathcal{T}_k$, $k = 1, 2$, be independent PBD variates with gp

$$f_k(\phi) = \sum_{u \in \Omega_k} c_k(u)\phi^u \tag{1.24}$$

Then $\mathcal{T} = \mathcal{T}_1 + \mathcal{T}_2$ is a PBD variate with probability function

$$\mathrm{P}[\mathcal{T} = t;\, \phi] = \frac{c(t)\phi^t}{f(\phi)} \tag{1.25}$$

where

$$c(t) \;=\; \sum_{u} c_1(u)c_2(t-u) \tag{1.26}$$

and

$$f(\phi) \;=\; f_1(\phi)f_2(\phi) \tag{1.27}$$

Further, the sample space of $\mathcal{T}$ is the set of exponents of the polynomial $f(\phi)$.

Proof: First note that

$$\begin{aligned} \mathrm{P}[\,\mathcal{T}=t;\,\phi\,] &= \sum_{u} \mathrm{P}[\,\mathcal{T}_1=u, \mathcal{T}_2=t-u\,] \\ &= \sum_{u} \mathrm{P}[\,\mathcal{T}_1=u]\mathrm{P}[\,\mathcal{T}_2=t-u\,] \end{aligned}$$

The last step follows by the independence property. Substituting the probability formula, and after some rearrangement, we have

$$\mathrm{P}[\,\mathcal{T}=t;\,\phi\,] \;=\; \frac{\{\sum_u c_1(u)c_2(t-u)\}\,\phi^t}{f_1(\phi)f_2(\phi)}$$

The desired results are a direct consequence. □

Theorem 1.3: Suppose $\mathcal{T}_1, \cdots, \mathcal{T}_K$ are independent PBD variates with gps $f_1(\phi), \cdots, f_K(\phi)$, and sample spaces $\Omega_1, \cdots, \Omega_K$, respectively. Then $\mathcal{T} = \mathcal{T}_1 + \cdots + \mathcal{T}_K$ has a PBD with gp $f(\phi)$ given by

$$f(\phi) \;=\; \prod_{k=1}^{K} f_k(\phi) \tag{1.28}$$

Proof: Follows by induction from Theorem 1.2. □

In some applications we consider, $\Omega_k = \{l_k, l_k+1, \ldots, u_k\}$, where l_k and u_k are integers. In that case $\Omega = \{l, l+1, \ldots, u\}$ where $l = \Sigma_k l_k$ and $u = \Sigma_k u_k$.

The main message thereby is that the distribution of the sum of independent PBD variates with the same parameters is obtained as a PBD from the product of their generating polynomials.

****

The polynomial formulation is particularly useful because:

```
Central Observation

Many of the discrete probability distributions commonly
used in statistical analysis can be expressed in the
form of a polynomial based distribution.
```

We demonstrate this observation below.

1.5 Binomial Distribution

Suppose n random patients from a homogeneous population receive a therapy for an acute disease. The outcome is cured ($\mathcal{Y} = 1$), or not cured ($\mathcal{Y} = 0$). Let $\mathcal{Y}_i$ denote the treatment result for the ith person, $i = 1, \ldots, n$. We make two assumptions: (i) All patients have the same chance, π, of being cured; (ii) Their outcomes are statistically independent of each other.

For this set of n independent and identically distributed (**iid**) Bernoulli trials, $\mathcal{T} = \Sigma_i \mathcal{Y}_i$ is the random total number of cures (or, in general, successes). The sample space of $\mathcal{T}$ is $\Omega = \{0, 1, \cdots, n\}$. Combinatorial arguments show that $\mathcal{T}$ has a **binomial distribution**, $\mathrm{B}(n, \pi)$, given by

$$\mathrm{P}[\,\mathcal{T} = t;\, \pi\,] \;=\; \binom{n}{t} \pi^t (1-\pi)^n - t, \qquad t \in \{0, 1, \ldots, n\} \tag{1.29}$$

where

$$\binom{n}{t} \;=\; \frac{n!}{t!(n-t)!} \tag{1.30}$$

Formulating the binomial parameter in terms of **odds**, i.e., the chance of success relative to the chance of failure, we have:

$$\phi \;=\; \pi/(1-\pi),$$

or, in terms of the logarithm of the odds (in short, **log-odds**):

$$\beta \;=\; \ln\{\pi/(1-\pi)\}$$

These transformations have the following properties:

- $0 \le \pi \le 1$ implies $0 \le \phi \le +\infty$, and $-\infty \le \beta \le +\infty$.
- These are one-to-one monotonic transformations with

$$\pi \;=\; \frac{\phi}{1+\phi} \;=\; \frac{\exp(\beta)}{1+\exp(\beta)} \tag{1.31}$$

- Any inference on ϕ or β can be translated into one for π and vice versa.

With these transformations, the binomial probability becomes

$$\mathrm{P}[\,\mathcal{T} = t;\, \phi\,] \;=\; \frac{c(t)\phi^t}{(1+\phi)^n} \;=\; \frac{c(t)\exp(\beta t)}{\{1+\exp(\beta t)\}^n} \tag{1.32}$$

where $c(t) = n!/\{t!(n-t)!\}$. Hence, the binomial distribution is a polynomial based distribution with gp equal to

$$(1 \,+\, \phi)^n \;=\; \sum_{u=0}^{n} c(u)\phi^u \tag{1.33}$$

Example 1.3: In a study of mammography screening, Elmore et al. (1998) found that about

one in two women tend to get a false positive report over a decade of annual screening for breast cancer. Then, for a random sample of twenty women, we need to compute the chance that at least 5 women will have a false positive report after a decade of mammography screening. First we compute the chance that at most 4 will have such an outcome. This is

$$2^{-20}\left(1+20+\frac{20\times 19}{2}+\frac{20\times 19\times 18}{3\times 2}+\frac{20\times 19\times 18\times 17}{4\times 3\times 2}\right)$$

which equals 0.006. Hence the probability that at least 5 women will have a false positive result in a decade of screening is about $1-0.006=0.994$.

****

For a binomial, we have that

$$\mu \;=\; n\pi \;=\; \frac{n\phi}{(1+\phi)} \quad \text{and} \quad \sigma^2 \;=\; n\pi(1-\pi) \;=\; \frac{n\phi}{(1+\phi)^2} \tag{1.34}$$

Suppose $\mathcal{T}_1$ and $\mathcal{T}_2$ are independent $\mathsf{B}(n,\pi)$ and $\mathsf{B}(m,\pi)$ variables. Then the gp of $\mathcal{T}_1+\mathcal{T}_2$ is

$$(1+\phi)^n(1+\phi)^m \;=\; (1+\phi)^{(n+m)}$$

Thus $\mathcal{T}_1+\mathcal{T}_2$ is a $\mathsf{B}(n+m,\pi)$ variable.

****

For large n, the probability distribution of $\mathcal{T}$ is approximated by the **normal distribution**. The approximation holds well when π is not far from 0.5. Let $\Phi(x)$ denote the cumulative distribution function of the standard normal. From asymptotic theory, we know that

$$\mathrm{P}[\,\mathcal{T}\le t\,] \;\approx\; \Phi\left(\frac{t-n\pi}{\sqrt{n\pi(1-\pi)}}\right) \tag{1.35}$$

A better approximation is provided with the use of the **continuity correction**. This gives

$$\mathrm{P}[\,\mathcal{T}\le t\,] \;\approx\; \Phi\left(\frac{t+0.5-n\pi}{\sqrt{n\pi(1-\pi)}}\right) \tag{1.36}$$

We apply the normal approximation (without the continuity correction) to the breast cancer screening example given above. In this case, $n=20$, $\pi=0.5$ and $t=5$. Then

$$z \;=\; \left(\frac{t-n\pi}{\sqrt{n\pi(1-\pi)}}\right) \;=\; -\sqrt{5}$$

From the standard normal distribution, this gives the desired probability as 0.988. This is close to the exact binomial probability, 0.994.

The assumption of a uniform cure rate (or identical trials) used to derive the binomial probability may not hold in practice. Alternatively, we may view the trials as a series of independent but

nonidentically distributed Bernoulli trials. Suppose the chance of success in the ith trial is π_i $(i = 1, \cdots, n)$, and π_i is modeled as a function of some features, called covariates, of the subject, or the study. For example, in a study where π_i stands for the chance the ith child will contract an ear infection, it is written as $\pi(x_i)$, where x_i indicates the gender of the ith child. Considerations of such additional variables or covariates often lead to the data being presented in the form of two or more 2×2 tables. Several chapters in this book deal with data in this format.

In some situations it may not be appropriate to assume that events are statistically independent. Modeling the total number of successes in such trials needs to account for the **extra binomial variation** or the interdependence between outcomes. There are many ways of doing it. Chapter 14 examines binary response models that relax the assumption of independence.

1.6 Poisson Distribution

The **Poisson distribution** is used for rare events, for example, for the number of accidents per day at a section of a roadway, or the number of new cases of a disease with a low incidence rate. Empirical studies indicate that it often models such processes with a reasonable accuracy.

Let Ω equal the set of nonnegative integers, $\{0, 1, 2, \ldots\}$. Then, for $t \in \Omega$, the Poisson probability is

$$\mathrm{P}[\mathcal{T} = t;\, \lambda] \;=\; \lambda^t e^{-\lambda}/t! \quad \text{where} \quad \lambda > 0 \tag{1.37}$$

This can also be written as

$$\mathrm{P}[\mathcal{T} = t;\, \lambda] \;=\; \frac{c(t)\lambda^t}{f(\lambda)} \tag{1.38}$$

where

$$c(t) \;= 1/t! \quad \text{and} \quad f(\lambda) \;=\; e^{\lambda} \;=\; \sum_{u=0}^{\infty} c(u)\lambda^u \tag{1.39}$$

Hence the Poisson is a PBD with $\Omega = \{0, 1, 2, \ldots, \}$, and $\phi = \lambda$.

****

Example 1.4: In a population based study of selected areas in Wisconsin, Nordstrom et al. (1998) estimated that the cases of newly diagnosed probable or definite carpal tunnel syndrome occur at the rate of 3.46 cases per 1000 person years. Assume that a random cohort of 100 subjects from the population without the condition is followed up for three years. We model this as a Poisson distribution with mean

$$\lambda \;=\; 3 \times 100 \times 0.00346 \;=\; 1.038$$

The chance that there at most two cases of carpal tunnel syndrome will occur in this sample during the study period is

$$e^{-1.038}\left(1+1.038+\frac{1.038^2}{2}\right) \;=\; 0.913$$

****

For the Poisson variate $\mathcal{T}$,

$$\mu \;=\; \sigma^2 \;=\; \lambda \tag{1.40}$$

If $\mathcal{T}$ represents the number of events occurring within a unit period of time, then λ is the average rate of occurrence per unit time.

An important property of Poisson variables is additivity. That is, if $\mathcal{T}_i$ is Poisson distributed with parameter $\lambda_i, i = 1, \cdots, n$, and if these random variables are mutually independent, then $\Sigma\mathcal{T}_i$ is a Poisson variate with parameter $\Sigma\lambda_i$. This follows from the product of the n gps:

$$\prod_{i=1}^{n} \exp(\lambda_i) \;=\; \exp\left(\sum_{i=1}^{n} \lambda_i\right)$$

and recognizing that this is a gp of Poisson variable with mean $\Sigma\lambda_i$.

The binomial distribution arises from the Poisson as follows. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be independent Poisson variates with parameters λ_1 and λ_2. Consider the conditional distribution of $\mathcal{T}_1$ when their sum is fixed:

$$\mathrm{P}[\,\mathcal{T}_1 = t \mid \mathcal{T}_1 + \mathcal{T}_2 = s\,] \;=\; \frac{\mathrm{P}[\,\mathcal{T}_1 = t,\, \mathcal{T}_1 + \mathcal{T}_2 = s\,]}{\mathrm{P}[\,\mathcal{T}_1 + \mathcal{T}_2 = s\,]}$$

which is equal to

$$\frac{\mathrm{P}[\,\mathcal{T}_1 = t]\mathrm{P}[\,\mathcal{T}_2 = s - t\,]}{\sum_u \mathrm{P}[\,\mathcal{T}_1 = u]\mathrm{P}[\,\mathcal{T}_2 = s - u\,]}$$

Substituting the Poisson formula and simplifying, we get

$$\mathrm{P}[\,\mathcal{T}_1 = t \mid \mathcal{T}_1 + \mathcal{T}_2 = s\,] \;=\; \binom{s}{t}\pi^t(1-\pi)^{s-t} \tag{1.41}$$

for $t = 0, 1, \ldots, s$, and where

$$\pi \;=\; \frac{\lambda_1}{\lambda_1 + \lambda_2} \tag{1.42}$$

The conditional distribution of $\mathcal{T}_1$ is thus a $\mathrm{B}(s, \pi)$ distribution.

****

The Poisson distribution also arises as a limiting distribution to the Binomial. Let $\mathcal{T}$ have $\mathrm{B}(n, \pi)$ distribution. Suppose n is large and π close to zero. Then

$$\mathrm{P}[\,\mathcal{T} = t\,] \;\approx\; \lambda^t e^{-\lambda}/t! \tag{1.43}$$

where $\lambda = n\pi$. This property is known as the Poisson approximation to the binomial. The normal approximation to the Poisson is based on the standardized variate

$$z \;=\; \frac{(t-\lambda)}{\sqrt{\lambda}}$$

Normal approximations to the Poisson and binomial distributions may, however, not be adequate even at fairly large sample sizes, as succinctly demonstrated by Jollife (1995).

Poisson distributions are used to model disease occurrence over time or space. The number of cases of leukemia in the vicinity of a nuclear reactor has been modeled in terms of a Poisson distribution. The number of incident cases of HIV infection in a large population during a time period τ is another case. If λ is the rate of new HIV infection per unit time, the number of new cases during the period may be a Poisson variate with mean $\lambda\tau$.

For events distributed over time or space, the Poisson probability is derived under the following assumptions:

- The probability of an event in a small interval is proportional to the size of the interval.
- The probability of two or more events in a small time interval is negligible.
- Events in disjoint intervals are independent of one another.

The Poisson distribution is also used to model the rate λ as a function of covariates. In the case of incident HIV cases, for example, these may be urban or rural residence, illicit drug use, and gender.

1.7 Negative Binomial Distribution

Consider a series of iid Bernoulli trials, $\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_3, \mathcal{Y}_4, \ldots$, with a common success probability π. For a fixed integer $r \geq 1$, let $\mathcal{T}$ denote the number of successes before the rth failure in the series $\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_3, \mathcal{Y}_4, \ldots$. Then for $t = 0, 1, 2, \ldots$,

$$\mathrm{P}[\,\mathcal{T} = t;\, \pi\,] \;=\; \binom{t+r-1}{r-1}\pi^t(1-\pi)^r \tag{1.44}$$

This is called the **negative binomial distribution**. The special case when $r = 1$ is called the **geometric distribution**.

Now let $\phi = \pi$, $f(\phi) = (1-\phi)^{-r}$ and

$$c(t) \;=\; \binom{t+r-1}{r-1}$$

Then we can write

$$\mathrm{P}[\,\mathcal{T} = t;\, \phi\,] \;=\; \frac{c(t)\phi^t}{f(\phi)} \tag{1.45}$$

showing that the negative binomial is also a PBD on the sample space $\Omega = \{0, 1, 2, \ldots\}$. Note that here, $0 < \phi < 1$. In (1.45), we used the property that for ϕ in this interval,

$$(1-\phi)^{-r} \;=\; \sum_{u=0}^{\infty} \binom{u+r-1}{r-1} \phi^u \tag{1.46}$$

****

Example 1.5: Lazarou, Pomeranz and Corey (1998) reported a 6.7% incidence rate of serious or fatal adverse drug reaction among hospitalized patients in the USA. Suppose a hospital institutes a monitoring program for newly admitted patients. Assuming the 6.7% incidence rate, the chance that the first serious or fatal adverse drug reaction will occur in the tenth patient is

$$0.933^9 \times 0.067 \;=\; 0.035$$

****

For the negative binomial variable $\mathcal{T}_r$, we can show that

$$\mu \;=\; \frac{r\pi}{1-\pi} \quad \text{and} \quad \sigma^2 \;=\; \frac{r\pi}{(1-\pi)^2} \tag{1.47}$$

1.8 Hypergeometric Distribution

Of a group of N individuals, assume n have a specific disease, and m do not. If we sample s persons at random without replacement from this group, the chance that t of them have the condition is

$$\mathrm{P}[\,\mathcal{T}=t\,] \;=\; \binom{n}{t}\binom{m}{s-t}\binom{n+m}{s}^{-1} \tag{1.48}$$

Let $l_1 = \max(0, s-m),\ l_2 = \min(n,s)$. The sample space of $\mathcal{T}$ is $\Omega = \{l_1, l_1+1, \ldots, l_2\}$. This is the central **hypergeometric distribution** for which

$$\mu \;=\; \frac{ns}{(n+m)} \tag{1.49}$$

$$\sigma^2 \;=\; \frac{nms(n+m-s)}{(n+m)^2(n+m-1)} \tag{1.50}$$

A more general version of this distribution arises from two binomial distributions. Let $\mathcal{A}$ be $\mathrm{B}(n,\pi_1)$, and let $\mathcal{B}$ be $\mathrm{B}(n,\pi_0)$, with $\mathcal{A}$ independent of $\mathcal{B}$. Consider the distribution of one of the variables if their sum is fixed, that is, $\mathrm{P}[\,\mathcal{A}=t \mid \mathcal{A}+\mathcal{B}=s\,]$. By definition, this equals

$$\frac{\mathrm{P}[\,\mathcal{A}=t]\mathrm{P}[\,\mathcal{B}=s-t\,]}{\mathrm{P}[\,\mathcal{A}+\mathcal{B}=s\,]} \;=\; \frac{\mathrm{P}[\,\mathcal{A}=t]\mathrm{P}[\,\mathcal{B}=s-t\,]}{\sum_u \mathrm{P}[\,\mathcal{A}=u]\mathrm{P}[\,\mathcal{B}=s-u\,]}$$

If we substitute the binomial expressions in the above, and let

$$c(t) \;=\; \binom{n}{t}\binom{m}{s-t} \quad \text{and} \quad \phi \;=\; \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} \tag{1.51}$$

then we get that

$$\mathrm{P}[\,\mathcal{A}=t \mid \mathcal{A}+\mathcal{B}=s\,] \;=\; \frac{c(t)\phi^t}{\sum_{u\,\in\,\Omega}\, c(u)\phi^u} \tag{1.52}$$

where Ω is as defined above. (1.52) is called the noncentral hypergeometric distribution. The central hypergeometric is the special case when $\phi = 1$. The parameter ϕ is the **odds ratio**. We will have the occasion to consider it in depth later.

Now define

$$f(\phi) \;=\; \sum_{u\,\in\,\Omega} c(u)\phi^u \tag{1.53}$$

Then

$$\mathrm{P}[\,\mathcal{A}=t \mid \mathcal{A}+\mathcal{B}=s\,] \;=\; \frac{c(t)\phi^t}{f(\phi)} \tag{1.54}$$

which obviously is a PBD.

****

Example 1.6: Suppose we randomly select 5 girls and 4 boys, all two years of age, from a population. They are monitored for six months for signs of ear disease. Let $\mathcal{A}$ denote the number of boys, and $\mathcal{B}$ denote the number of girls contracting an ear infection. Suppose a total of three children develop the infection. The conditional probability of $\mathcal{A}$, the number of boys with the infection, is obtained by first computing $c(t)$ for each t. The four possible values of t and the associated coefficients appear in Table 1.3

Table 1.3 *Hypergeometric Coefficients*

t	$c(t)$
0	10
1	40
2	30
3	4

With $f(\phi) = 10 + 40\phi + 30\phi^2 + 4\phi^3$, it follows that

$$\begin{aligned}
\mathrm{P}[\,\mathcal{A}=0 \mid \mathcal{A}+\mathcal{B}=3\,] &= 10/f(\phi)\\
\mathrm{P}[\,\mathcal{A}=1 \mid \mathcal{A}+\mathcal{B}=3\,] &= 40\phi/f(\phi)\\
\mathrm{P}[\,\mathcal{A}=2 \mid \mathcal{A}+\mathcal{B}=3\,] &= 30\phi^2/f(\phi)\\
\mathrm{P}[\,\mathcal{A}=3 \mid \mathcal{A}+\mathcal{B}=3\,] &= 4\phi^3/f(\phi)
\end{aligned}$$

****

For future reference, we list in Table 1.4 the gps of three commonly used univariate discrete distributions.

Table 1.4 *Generating Polynomials*

Distribution	ϕ	$f(\phi)$
Binomial	$\pi/(1-\pi)$	$(1+\phi)^n$
Poisson	λ	$\exp(\phi)$
Negative Binomial	π	$(1-\phi)^{-r}$

1.9 A General Representation

Now we define the multivariate polynomial based distribution. Consider a parameter vector, $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_K)$, $\phi_k \geq 0$, and let $\boldsymbol{\mathcal{T}} = (\mathcal{T}_1, \cdots, \mathcal{T}_K)$ be a discrete random vector realizing values $\boldsymbol{t}$ in a countable K-dimensional set Ω. Define

$$\boldsymbol{\phi}^{\boldsymbol{t}} \;=\; \prod_{k\,=\,1}^{K} \phi^{t_k} \tag{1.55}$$

Note $\boldsymbol{\phi}^{\boldsymbol{t}}$ is **not** a vector. Let

$$f(\boldsymbol{\phi}) \;=\; \sum_{\boldsymbol{u}\,\in\,\Omega} c(\boldsymbol{u})\boldsymbol{\phi}^{\boldsymbol{u}} \tag{1.56}$$

be a polynomial in K parameters. $\boldsymbol{\mathcal{T}}$ is said to have a multivariate polynomial based distribution with gp $f(\boldsymbol{\phi})$ if

$$\mathrm{P}[\,\boldsymbol{\mathcal{T}} = \boldsymbol{t}\,] \;=\; \frac{c(\boldsymbol{t})\boldsymbol{\phi}^{\boldsymbol{t}}}{f(\boldsymbol{\phi})} \tag{1.57}$$

where $c(\boldsymbol{t}) > 0$ for all $\boldsymbol{t} \in \Omega$.

To express this distribution in exponential form, we let $\phi_k = \exp(\beta_k)$ with $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_K)$. Then

$$\mathrm{P}[\,\boldsymbol{\mathcal{T}} = \boldsymbol{t}\,] \;=\; \frac{c(\boldsymbol{t})\exp(\boldsymbol{\beta}\boldsymbol{t}')}{h(\boldsymbol{\beta})} \tag{1.58}$$

where

$$h(\boldsymbol{\beta}) \;=\; \sum_{\boldsymbol{u}\,\in\,\Omega} c(\boldsymbol{u})\exp(\boldsymbol{\beta}\boldsymbol{u}') \tag{1.59}$$

The gp $f(\boldsymbol{\phi})$ has properties similar to those of the univariate gp $f(\phi)$. Some of these are listed below:

Property 1.4: $f(\boldsymbol{\phi})$ completely and uniquely specifies the multivariate PBD.

Property 1.5: $f(\boldsymbol{\phi})$ is an equivalent substitute for the probability generating function.

Property 1.6: The moments of a multivariate PBD are derived from the gp in a manner analogous to that for a univariate PBD.

Property 1.7: The sum of independent identically distributed multivariate PBD vectors is a multivariate PBD vector whose gp is a product of the gps of the vectors in the sum.

The proofs of these assertions are straightforward generalizations of respective proofs given for their univariate counterparts and are left to the exercises. Now let us consider an example.

1.10 The Multinomial Distribution

The multivariate distribution frequently applied in discrete data settings is the **multinomial distribution**. Consider a series of trials such that in each trial there are K possible outcome categories. An example of a three category case is: a child may either have no earache, have an ache in a single ear, or have the problem in both ears. Let the chance the kth outcome will materialize be π_k with $0 \leq \pi_k \leq 1$ and $\Sigma\pi_k = 1$. Let $\mathcal{T}_k$ denote the random number of times the kth category occurs in a series of n independent trials. The joint probability for these random variables is

$$\mathrm{P}[\,\mathcal{T}_1 = t_1, \cdots, \mathcal{T}_K = t_K\,] \;=\; \frac{n!}{t_1!t_2!\cdots t_K!}\prod_k \pi_k^{t_k} \tag{1.60}$$

We denote this distribution as $\mathrm{M}(n; \pi_1, \cdots, \pi_K)$. The sample space Ω is the space of vectors $(t_1, \ldots, t_k)$ that satisfy

$$0 \leq t_k \leq n,\; k = 1, \cdots, K \quad \text{and} \quad \Sigma t_k = n$$

With the special case of $K = 3$, we illustrate a multivariate PBD. Suppose we perform n independent trials, and each trial has one of three outcomes labeled A, B and C. Let $(\mathcal{Y}_{1i}, \mathcal{Y}_{2i})$ be the outcome of the ith trial with $\mathcal{Y}_{1i} = 1$ if the ith trial results in "A", and $= 0$ otherwise; and $\mathcal{Y}_{2i} = 1$ if it results in "B", and $= 0$ otherwise. Let $\mathcal{T}_1, \mathcal{T}_2$ and $\mathcal{T}_3$ be the numbers of trials with outcomes A, B and C, respectively. Obviously,

$$\begin{aligned}
\mathcal{T}_1 &= \sum_{i=1}^{n} \mathcal{Y}_{1i} \\
\mathcal{T}_2 &= \sum_{i=1}^{n} \mathcal{Y}_{2i} \\
\mathcal{T}_3 &= n - (\,\mathcal{T}_1 + \mathcal{T}_2\,)
\end{aligned}$$

Further,

$$\mathrm{P}[\,\mathcal{T}_1 = t_1, \mathcal{T}_2 = t_2\,] \;=\; c(t_1, t_2)\pi_1^{t_1}\pi_2^{t_2}\pi_3^{t_3}$$

Table 1.5 *Trinomial Sample Space and Coefficients*

t_1	t_2	t_3	$c(t_1,t_2)$	t_1	t_2	t_3	$c(t_1,t_2)$
0	0	4	1	1	3	0	4
0	1	3	4	2	0	2	6
0	2	2	6	2	1	1	12
0	3	1	4	2	2	0	6
0	4	0	4	3	0	1	4
1	0	3	1	3	1	0	4
1	1	2	12	4	0	0	1
1	2	1	12				

where π_1, π_2 and π_3 are the chances of A, B and C, respectively, in any trial, with $0 \leq \pi_1, \pi_2, \pi_3 \leq 1$, and $\pi_1 + \pi_2 + \pi_3 = 1$; where $t_1 + t_2 + t_3 = n$; and further, where

$$c(t_1, t_2) \;=\; \frac{n!}{t_1! t_2! (n - t_1 - t_2)!} \tag{1.61}$$

Let $\phi_1 = \pi_1/(1 - \pi_1 - \pi_2)$, and $\phi_2 = \pi_2/(1 - \pi_1 - \pi_2)$. With a slight rearrangement, we write this trinomial probability as

$$\mathrm{P}[\,\mathcal{T}_1 = t_1, \mathcal{T}_2 = t_2\,] \;=\; \frac{c(t_1, t_2)\phi_1^{t_1}\phi_2^{t_2}}{f(\phi_1, \phi_2)} \tag{1.62}$$

where

$$f(\phi_1, \phi_2) \;=\; (1 + \phi_1 + \phi_2)^n \tag{1.63}$$

From basic algebra, we know that

$$f(\phi_1, \phi_2) \;=\; \sum_{(u,v)\,\in\,\Omega} c(u, v)\phi_1^{u_1}\phi_2^{u_2} \tag{1.64}$$

where $\Omega = \{(u, v) : 0 \leq u, v \leq n \text{ and } 0 \leq u + v \leq n\}$. Hence, the trinomial distribution is a PBD. The general multinomial distribution is also shown to be a PBD in a similar way.

Example 1.7: Consider a concrete case with $n = 4$ and $K = 3$. A polynomial representation of this trinomial is in Table 1.5.

****

For the multinomial distribution, we can show that

$$\mu_k \;=\; \mathbf{E}[\,\mathcal{T}_k\,] \;=\; n\pi_k \tag{1.65}$$

$$\sigma_k^2 \;=\; \mathrm{var}[\,\mathcal{T}_k\,] \;=\; n\pi_k(1 - \pi_k) \tag{1.66}$$

$$\sigma_{kj} \;=\; \mathbf{E}[\,(\mathcal{T}_k - \mu_k)(\mathcal{T}_j - \mu_j)\,] \;=\; -n\pi_k\pi_j, \quad k \neq j \tag{1.67}$$

The following properties of a multinomial are useful in discrete data analysis.

Property 1.8: Aggregating a set of multinomial random variables will yield a smaller set of random variables which are also multinomial.

For example, suppose $(\mathcal{T}_1, \cdots, \mathcal{T}_5)$ is M$(n; \pi_1, \cdots, \pi_5)$. Define $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ as: $\mathcal{S}_1 = \mathcal{T}_1 + \mathcal{T}_2, \mathcal{S}_2 = \mathcal{T}_3 + \mathcal{T}_4$, and $\mathcal{S}_3 = \mathcal{T}_5$. Then $(\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3)$ is M$(n; \pi_1 + \pi_2, \pi_3 + \pi_4, \pi_5)$.

Property 1.9: The distribution obtained by conditioning on distinct sums of multinomial outcomes is a product of multinomial distributions.

Suppose $(\mathcal{T}_1, \cdots, \mathcal{T}_5)$ is M$(n; \pi_1, \cdots, \pi_5)$. Consider, for example, the probability of $\mathcal{T}_1, \mathcal{T}_3, \mathcal{T}_4$ given $\mathcal{T}_1 + \mathcal{T}_2 = t$ (and thus $\mathcal{T}_3 + \mathcal{T}_4 + \mathcal{T}_5 = n - t$). This distribution is a product of M$(t; \pi_1^*, \pi_2^*)$ and M$(n - t; \pi_1^{**}, \pi_2^{**}, \pi_3^{**})$, where

$$\begin{aligned}
\pi_1^* &= \frac{\pi_1}{\pi_1 + \pi_2} \\
\pi_2^* &= \frac{\pi_2}{\pi_1 + \pi_2} \\
\pi_1^{**} &= \frac{\pi_3}{\pi_3 + \pi_4 + \pi_5} \\
\pi_2^{**} &= \frac{\pi_4}{\pi_3 + \pi_4 + \pi_5} \\
\pi_3^{**} &= \frac{\pi_5}{\pi_3 + \pi_4 + \pi_5}
\end{aligned}$$

Property 1.10: If $\mathcal{T}_k$, $k = 1, \ldots, K$, are independent Poisson(λ_k), then $\mathrm{P}[\, \mathcal{T}_1 = t_1, \cdots, \mathcal{T}_K = t_K \mid \Sigma_k \mathcal{T}_k = s \,]$ is M$(s; \pi_1, \cdots, \pi_K)$ where $\pi_j = \lambda_j / \Sigma_k \lambda_k$.

The multinomial is the subtext for many distributions used to model and analyze discrete data. Properties 1.8, 1.9 and 1.10 are used to produce the conditional distribution or likelihood from which inference on the data is drawn.

The multinomial distribution is also used to model dependent discrete variables. For example, in the case of two Bernoulli variables, consider the occurrence or otherwise of cataract in the two eyes of an individual. These may not be independent events. One formulation of the possible dependence between them is as follows. Consider the variables $\mathcal{Y}_1$ and $\mathcal{Y}_0$ with respective observed value $y_1, y_0 \in \{0, 1\}$. Then let

$$\mathrm{P}[\, \mathcal{Y}_1 = y_1, \mathcal{Y}_0 = y_0 \,] \;=\; \frac{\theta_1^{y_1} \theta_0^{y_0} \theta_{10}^{y_1 y_0}}{1 + \theta_1 + \theta_0 + \theta_1 \theta_0 \theta_{10}} \tag{1.68}$$

where all the θ's are nonnegative parameters. Let $\pi_{ij} = \mathrm{P}[\, \mathcal{Y}_1 = i, \mathcal{Y}_0 = j \,]$. This is a special case of a four outcome multinomial. In this distribution, if $\theta_{10} = 1$, then we get a product of two independent Bernoulli variables with success probabilities $\theta_j / (1 + \theta_j)$, $j = 1, 0$. We use models based on such a distribution in Chapter 14. This formulation can also be extended to more than two dependent binary variables.

1.11 The Negative Trinomial

Consider a sequence of independent trials with three outcomes, labeled $\{1, 2, 3\}$, in each trial. Let their respective probabilities of occurrence be π_1, π_2 and π_3 with $\pi_1 + \pi_2 + \pi_3 = 1$. We

conduct the experiment until the total number of outcomes of type 3 reaches r. Let $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$ respectively be the total number of outcomes of each type. The resulting probability distribution is a **negative trinomial distribution**. Using the arguments like those for the negative binomial, we can show that it has a polynomial form:

$$\mathrm{P}[\,\mathcal{T}_1 = t_1,\, \mathcal{T}_2 = t_2\,] \;=\; \frac{c(t_1,t_2)\phi_1^{t_1}\phi_2^{t_2}}{f(\phi_1,\phi_2)} \tag{1.69}$$

where $\phi_1 = \pi_1$, $\phi_2 = \pi_2$

$$c(t_1,t_2) \;=\; \binom{t_1+t_2+r-1}{t_1 \quad t_2}$$

and

$$f(\phi_1,\phi_2) \;=\; (1-\phi_1-\phi_2)^{-r} \tag{1.70}$$

Note that here $0 < \phi_1, \phi_2 < 1$.

This can be readily generalized to the case with K outcome categories.

****

The other cases of the multivariate PBD, which we will encounter later, include several forms of multivariate extensions of the hypergeometric and conditional distributions that are derived from multivariate discrete probability models.

1.12 Sufficient Statistics

Sufficiency is a key concept in conditional methods for data analysis. We introduce it through a sequence of n iid Bernoulli variables, $\{\mathcal{Y}_1, \cdots, \mathcal{Y}_n\}$. Set

$$\pi = \exp(\beta)/\{1+\exp(\beta)\}$$

Then

$$\mathrm{P}[\,\mathcal{Y}_1 = y_1, \ldots, \mathcal{Y}_n = y_n\,] \;=\; \frac{\exp(\beta\Sigma_i\, y_i)}{\{1+\exp(\beta)\}^n}$$

Let $\mathcal{T} = \Sigma_i \mathcal{Y}_i$, and consider the conditional probability

$$\mathrm{P}[\,\mathcal{Y}_1 = y_1, \ldots, \mathcal{Y}_n = y_n \;\mid\; \mathcal{T} = t\,]$$

$$= \;\frac{\mathrm{P}[\,\mathcal{Y}_1 = y_1, \ldots, \mathcal{Y}_n = y_n\,]}{\mathrm{P}[\,\mathcal{T} = t\,]} \;=\; \frac{t!(n-t)!}{n!}$$

The conditional probability of the n iid Bernoullis given their sum does not contain the parameter β or π. In this sense, the knowledge of $\mathcal{T}$ has served to extricate the parameter from the distribution. Formally, we define:

Sufficient Statistic

A random variable $\mathcal{T}$ is sufficient for the parameter β if the conditional probability of the data given $\mathcal{T}$ does not contain the parameter.

Note $\mathcal{T}$ and β may be vectors. The **factorization theorem**, stated below without proof, often facilitates identification of sufficient statistics.

Theorem 1.4: Let $\mathcal{T}$ be a function of discrete random variable(s) $\mathcal{Y}$ (both may be vectors). $\mathcal{T}$ is sufficient for parameter β if and only if, for any y,

$$\mathrm{P}[\,\mathcal{Y} = y; \beta\,] \;=\; g(t(y), \beta) q(y) \tag{1.71}$$

□

Two important results on sufficiency and PBD need to be stated.

Theorem 1.5: Let $\boldsymbol{\mathcal{T}} = (\mathcal{T}_1, \cdots, \mathcal{T}_K)$ be multivariate PBD. Then $\mathcal{T}_j$ is sufficient for ϕ_j (or for β_j).

Proof: Use the definition or the factorization theorem. □

Theorem 1.6: Let $\boldsymbol{\mathcal{T}} = (\mathcal{T}_1, \cdots, \mathcal{T}_K)$ have a multivariate PBD. The conditional distribution of a set of $\mathcal{T}_i$'s given some other $\mathcal{T}_j$'s is also multivariate PBD.

Proof: Let $\boldsymbol{\mathcal{T}}_1$ and $\boldsymbol{\mathcal{T}}_2$ be vectors of distinct elements of the vector $\boldsymbol{\mathcal{T}}$ with $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$, the corresponding parameter vectors. Write the joint generating polynomial as

$$f(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) \;=\; \sum_{(\boldsymbol{u}_1, \boldsymbol{u}_2) \,\in\, \Omega} c(\boldsymbol{u}_1, \boldsymbol{u}_2) \boldsymbol{\phi}_1^{\boldsymbol{u}_1} \boldsymbol{\phi}_2^{\boldsymbol{u}_2} \tag{1.72}$$

Then $\mathrm{P}[\,\boldsymbol{\mathcal{T}}_1 = \boldsymbol{t}_1 \mid \boldsymbol{\mathcal{T}}_2 = \boldsymbol{t}_2\,]$ is

$$\frac{\mathrm{P}[\,\boldsymbol{\mathcal{T}}_1 = \boldsymbol{t}_1, \boldsymbol{\mathcal{T}}_2 = \boldsymbol{t}_2\,]}{\sum_{\boldsymbol{u} \,\in\, \Omega(.,\boldsymbol{t}_2)} \mathrm{P}[\,\boldsymbol{\mathcal{T}}_1 = \boldsymbol{u}, \boldsymbol{\mathcal{T}}_2 = \boldsymbol{t}_2\,]} \;=\; \frac{c(\boldsymbol{t}_1, \boldsymbol{t}_2) \boldsymbol{\phi}_1^{\boldsymbol{t}_1}}{f(\boldsymbol{\phi}_1, \mathbf{1}; ., \boldsymbol{t}_2)} \tag{1.73}$$

where

$$f(\boldsymbol{\phi}_1, \mathbf{1}; ., \boldsymbol{t}_2) \;=\; \sum_{\boldsymbol{u} \,\in\, \Omega(.,\boldsymbol{t}_2)} c(\boldsymbol{u}, \boldsymbol{t}_2) \boldsymbol{\phi}_1^{\boldsymbol{u}} \tag{1.74}$$

and where $\Omega(., \boldsymbol{t}_2)$ is the set of values $\boldsymbol{\mathcal{T}}_1$ from the vectors $(\boldsymbol{\mathcal{T}}_1, \boldsymbol{\mathcal{T}}_2) \in \Omega$ in which $\boldsymbol{\mathcal{T}}_2 = \boldsymbol{t}_2$.□

This conditional distribution excludes the parameters in $\boldsymbol{\phi}_2$. It is thus of use when the analysis only concerns the parameters in $\boldsymbol{\phi}_1$. For models with sufficient statistics, appropriate conditioning allows us to focus on the parameters of interest. It was observations along such lines that historically gave birth to the field of exact analysis of discrete data. In summary:

- Sufficient statistics in a PBD are readily identifiable.
- The conditional distribution (1.73) is PBD.
- The conditional gp (1.74) has the same properties as the gp of an unconditional PBD.

The conditional PBDs we have seen thus far are: (i) The noncentral hypergeometric of §1.8, (ii) the binomial derived from conditioning on two Poisson distributions in §1.6 and (iii) the multinomial derived by conditioning on the sum of several Poisson variates noted in Property 1.10 of §1.10.

1.13 The Polynomial Form

We have shown many common discrete probability distributions can be expressed in a polynomial form. Using this form for exact (and even large sample) analysis of discrete distributions is additionally suggested for the following reasons. These reasons will become clearer as we proceed through the chapters of this text.

Why Use The Polynomial Form?

- Discrete data analysis often involves polynomial based distributions.
- Discrete distributions for which exact conditional methods have been developed so far mostly are those with the form of a univariate or a multivariate PBD.
- The common underlying distributional form allows us to construct a unified strategy for exact and asymptotic inference.
- The parameterizations that produce the polynomial forms are in accord with the common usage of odds ratios and log-odds ratios in conventional analysis of discrete data.
- The polynomial form promotes a unified development, and simple portrayal, of efficient computational algorithms for exact analysis of discrete data.

Computational algorithms for exact inference are too often described in unnecessarily intricate ways and appear to be very complex. A key aim of this book is to demonstrate that that is not the case.

Before we end this section, we give a general notation for polynomials, first invoked in Theorem 1.6 above, for later use. For example, consider a trivariate polynomial

$$f(\phi_1, \phi_2, \phi_3) \;=\; \sum_{(u_1,u_2,u_3)\,\in\,\Omega} c(u_1, u_2, u_3)\phi_1^{u_1}\phi_2^{u_2}\phi_3^{u_3} \tag{1.75}$$

Suppose we want terms in this polynomial in which the exponents of ϕ_2 are equal to or greater than t_2, and the exponents of ϕ_3 are equal to t_3. We write this subpolynomial as $f(\phi_1, \phi_2, \phi_3;., \geq t_2, t_3)$. That is

$$f(\phi_1, \phi_2, \phi_3;., \geq t_2, t_3) \;=\; \phi_3^{t_3} \sum_{u_1} \sum_{u_2 \geq\, t_2} c(u_1, u_2, t_3)\phi_1^{u_1}\phi_2^{u_2} \tag{1.76}$$

Next suppose we want terms in (1.75) in which the exponents of ϕ_2 are equal to t_2 and those of ϕ_3 are equal to t_3 when $\phi_2 = \phi_3 = 1$. We write this subpolynomial as $f(\phi_1, 1, 1;., t_2, t_3)$. That is

$$f(\phi_1, 1, 1;., t_2, t_3) \;=\; \sum_{u_1} c(u_1, t_2, t_3)\phi_1^{u_1} \tag{1.77}$$

Such a notation is also useful for the sample space Ω. For example, $\Omega(., t_2, \geq t_3)$ represents that segment of Ω when the value along the second dimension is fixed at t_2 and the points along the third dimension are all greater than or equal to t_3.

This notation, which readily generalizes to higher dimensions and other forms of restrictions, is used later to represent conditional distributions derived from multivariate polynomial based distributions as well as their tail portions.

1.14 Relevant Literature

The material in this chapter is but a tiny portion from the vast field of probability and statistics. The references below, among many others, provide the broader background and further elaboration.

The main discrete distributions, the normal distribution and concepts like independence and conditioning are covered in many elementary books. For example, Thomas (1986) covers a wide ground in a readable manner. Gordon (1997) contains a particularly lucid account of discrete probability. Wild and George (2000) is an ideal elementary introduction to probability and statistical inference while Roussas (2003) clearly explains intermediate level material.

Many texts relate common continuous and discrete probability distributions. For instance, the cumulative binomial is linked to the F distribution and the Poisson, to the chisquare distribution. Many other asymptotic approximations to discrete distributions also exist; for example, the arcsine approximation of the binomial to the normal. For a comprehensive overview of discrete distributions, see Johnson and Kotz (1969) and Johnson, Kotz and Kemp (1992).

Books on stochastic processes generally derive the Poisson probability under mild assumptions. A readable work is Goodman (1988). Properties of the multinomial are discussed in Bishop, Fienberg and Holland (1975).

A sizeable theoretical literature on power series distributions, or what we call the PBD, exists. See Johnson and Kotz (1969) and Patil (1986). A multivariate PBD is a member of the discrete version of the multivariate exponential family of distributions (Patil 1985). A proof of a portion of Theorem 1.1 is in Patil (1986). Bivariate PBDs are well covered in Kocherlakota and Kocherlakota (1992). Pe'rez–Abreu (1991) has given a proof of the applicability of the Poisson approximation to power series distributions in general.

Distributions expressed in a polynomial form (or its exponential version) are found in many papers dealing with exact inference on discrete data. Cases in point: The exact distributions in the seminal paper, Zelen (1971), are in a polynomial form. So are several in the comprehensive review, Agresti (1992). The exact distributions for logistic regression models of discrete data in the influential text, Cox and Snell (1989), are in the exponential polynomial form. An explicit linkage between power series distributions and those used in the analysis of discrete data was made in Hirji (1997a).

The history of generating functions in exact inference on discrete data is a long one. See Cox (1958), Cox (1970) and Hirji, Mehta and Patel (1987). Two recent papers showing their use for

discrete data analysis are Baglivo, Pagano and Spino (1996) and Hirji (1997a). van de Wiel, Di Bucchianico and van der Laan (1999) has a broad perspective on the subject and notes other relevant material.

For a rigorous approach to sufficiency, see Lehmann (1986) and Cox and Hinkley (1974). Application of sufficiency to discrete data models is covered in Cox and Snell (1989) and Bishop, Fienberg and Holland (1975). Pratt and Gibbons (1981) has an elementary proof of the factorization theorem.

1.15 Exercises

1.1. Construct a probability distribution based on each of the following polynomials: (i) $f_1(\phi) = 3\phi^3+11\phi^5+7\phi^7$; (ii) $f_2(\phi) = \phi^3+4\phi^4+5\phi^5+6\phi^6+7\phi^7$; and (iii) $f_3(\phi) = \phi^{-1}+3+4\phi+4\phi^2+3\phi^5+\phi^9$. For each distribution compute $\mathrm{P}[\,\mathcal{T}=5;\ \phi=1.5\,]$ and $\mathrm{P}[\,\mathcal{T}<5;\ \phi=1.5\,]$.

1.2. With $f_1(\phi)$ and $f_2(\phi)$ defined above, let $f(\phi) = f_1(\phi)f_2(\phi)$ be the generating polynomial of a discrete PBD random variable $\mathcal{T}$. Compute $\mathrm{P}[\,\mathcal{T}=10;\ \phi=1.5\,]$ and $\mathrm{P}[\,\mathcal{T}\geq 10;\ \phi=1.5\,]$. Also compute the mean, median and variance of this distribution when $\phi = 1.5$.

1.3. A rare disease occurs in region A at an average rate of three cases a month, and in region B, at an average rate of two cases a month. Assume that cases in any one region occur independently of that in the other. Compute the probability that in a given month: (i) a total of at most three cases are reported from the two regions, (ii) at least two cases are reported in each region and (iii) no cases are reported in either region.

1.4. Kessler (1993) estimated that the reporting rate for serious adverse reactions to prescription drugs in the U.S. is about 1%. Using this as the underlying reporting rate, compute the probability that of the 100 serious reactions occurring in a given period at a medical facility, at most 5 will be reported. Use three methods for this: the exact binomial probability, the Poisson approximation to the binomial and the normal approximation to the binomial. Comment on the results.

1.5. For the geometric variate $\mathcal{T}$ with $\pi = 0.2$, what is (i) $\mathrm{P}[\,\mathcal{T}=2\,]$ and (ii) $\mathrm{P}[\,\mathcal{T}>4\,]$?

1.6. Show that the negative binomial variable arises as the sum of r independent geometric variables. Use this to derive its mean and variance.

1.7. Determine the mean, median and variance of the PBDs with generating polynomials

$$f(\phi) \;=\; (1+\phi)^2(1+2\phi+\phi^2)^2$$

and

$$f(\phi) \;=\; (2+3\phi)(3+2\phi)^2$$

1.8. Let $\Omega = \{1, 2, \cdots, n\}$ and let $\mathrm{P}[\,\mathcal{T}=t\,] = 1/n$ for $t \in \Omega$. $\mathcal{T}$ is the discrete uniform random variable. Is $\mathcal{T}$ a PBD variate? Compute the mean, median and variance of $\mathcal{T}$.

1.9. If $\mathcal{T}_1$ and $\mathcal{T}_2$ are discrete uniform variates on $\Omega_1 = \{1, 2, \cdots, n_1\}$ and $\Omega_2 = \{1, 2, \cdots, n_2\}$, what is the distribution of (i) $\mathcal{T}_1 + \mathcal{T}_2$ and (ii) $\min(\mathcal{T}_1, \mathcal{T}_2)$?

1.10. For a Poisson variate $\mathcal{T}$, find the probability that (i) $\mathcal{T}$ is even and (ii) $\mathcal{T}$ is odd.

1.11. Prove the variance formulae in Theorem 1.1.

1.12. Derive the mean and variance of the binomial, Poisson, central hypergeometric and negative binomial distributions in two ways, directly and by using Theorem 1.1.

1.13. What is the median of $\mathrm{B}(n, 0.5)$ when (i) n is even and (ii) n is odd?

1.14. Determine the medians of the Poisson distributions with $\lambda = 0.5, 1.0, 1.5$.

1.15. What is the median of the hypergeometric distribution when (i) $\phi = 1.0$, and (ii) $\phi = 0.5$ and $n = m$.

1.16. Determine the medians of the negative binomial distributions with $r = 5$ and $\pi = 0.25, 0.50, 0.75$.

1.17. How would you determine the median of the geometric distribution?

1.18. The kth factorial moment of $\mathcal{T}$ is

$$\mathbf{E}[\,\mathcal{T}(\mathcal{T}-1)\ldots(\mathcal{T}-k+1)\,]$$

Determine the kth factorial moment of a PBD in terms of its gp, $f(\phi)$. Use this to find the kth factorial moments of the binomial, Poisson and negative binomial distributions.

1.19. (i) If $\mathcal{T}$ is a B(n, π) variate, then prove that

$$\mathbf{E}[(\mathcal{T}-n\pi)^3] \;=\; n\pi(1-\pi)(1-2\pi)$$

(ii) If $\mathcal{T}$ is a Poisson variate with parameter λ, then prove that

$$\mathbf{E}[(\mathcal{T}-\lambda)^3] \;=\; \lambda$$

1.20. Let $(\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3)$ have M$(12; 0.5, 0.25, 0.25)$ distribution. What is the (i) marginal probability of $\mathcal{T}_1$ and (ii) conditional probability of $\mathcal{T}_1$ given $\mathcal{T}_2 = t_2$? Further, compute: (i) $\mathrm{P}[\,\mathcal{T}_1 = 4,\, \mathcal{T}_2 = 3\,]$, (ii) $\mathrm{P}[\,\mathcal{T}_1 \geq 3,\, \mathcal{T}_2 \geq 2\,]$, (iii) $\mathrm{P}[\,\mathcal{T}_1 = 4\,]$, and (iv) $\mathrm{P}[\,\mathcal{T}_1 \leq 4\,]$.

1.21. What is the generating polynomial for a multinomial variate with n trials and K outcome categories? For this distribution, show that

$$\sum_{\boldsymbol{t}\,\in\,\Omega} c(\boldsymbol{t}) \;=\; K^n$$

1.22. For the trinomial distribution of §1.10, show that the conditional distribution of $\mathcal{T}_1$ given $\mathcal{T}_2$ is a PBD.

1.23. Derive the mean, variance and covariance formulae for the multinomial distribution stated in §1.10.

1.24. Consider the model for two dependent Bernoulli variates in §1.10. What is the probability of having cataracts in both eyes given that at least one is affected?

1.25. Suppose $\mathcal{T}_1$ is a B(n, π_1) and $\mathcal{T}_2$ is a B(m, π_2) random variable. If they are independent, determine the distribution of $\mathcal{T}_1 + \mathcal{T}_2$. Is it a PBD?

1.26. Give a formal proof of Theorem 1.3.

1.27. Prove that the sum of iid multivariate PBD vectors is a multivariate PBD vector whose gp is a product of the gps of the vectors in the sum.

1.28. Prove the properties 1.8, 1.9 and 1.10 of the multinomial distribution stated in §1.10, and state them in a general form.

1.29. If $(\mathcal{T}_1, \mathcal{T}_2)$ has a bivariate PBD, then derive the covariance of $\mathcal{T}_1$ and $\mathcal{T}_2$ in terms of its generating function. Extend Theorem 1.1 to a multivariate PBD including the specification of the covariances in terms of the generating polynomial $f(\boldsymbol{\phi})$. Apply these results to the multinomial distribution and the negative trinomial distribution.

1.30. Suppose $\mathcal{T}_j, j = 1, \ldots, K$ is a binomial variate with success probability π and total number of trials equal to n_j. How would you obtain the distribution of $\Sigma_j \mathcal{T}_j^2$? Extend your result to K Poisson, negative binomial and general PBD distributions.

1.31. Give a detailed proof that the negative trinomial of §1.11 is a PBD and clearly specify its sample space. Further: (i) Derive the mean and variance of $\mathcal{T}_1$ and $\mathcal{T}_2$ as well as the covariance between them. (ii) What is the distribution of $\mathcal{T}_1+\mathcal{T}_2$? (iii) What is the marginal distribution of $\mathcal{T}_1$? (iv) Generalize this to the case with K outcome categories.

1.32. Consider an experiment with three outcome categories in each independent trial. It is continued until r outcomes of either type 2 or type 3 occur. Derive the joint distribution of $(\mathcal{T}_1, \mathcal{T}_2)$. Is it a PBD?

1.33. Suppose $\mathcal{T}_1, \cdots, \mathcal{T}_K$ are independent but not necessarily identically distributed PBD variates with gps $f_1(\phi_1), \cdots, f_K(\phi_K)$, and sample spaces $\Omega_1, \cdots, \Omega_K$, respectively. Is the distribution of $\mathcal{T} = \mathcal{T}_1 + \cdots + \mathcal{T}_K$ necessarily a PBD? Apply this to: (i) three independent binomial variates, (ii) three independent Poisson variates and (iii) two independent negative binomial variates.

1.34. For a nonnegative random variable $\mathcal{T}$, and for $|\psi| < 1$, the probability generating function, $G_{\mathcal{T}}(\psi)$, is defined by

$$G_{\mathcal{T}}(\psi) \;=\; \mathbf{E}[\,\psi^{\mathcal{T}}\,] \;=\; \sum_{t} \psi^t \mathrm{P}[\,T = t\,]$$

Suppose $\mathcal{T}$ is a PBD variate with gp $f(\phi)$. Then show that

$$G_{\mathcal{T}}(\psi) \;=\; \frac{f(\psi\phi)}{f(\phi)}$$

Therefore, the probability generating function of a PBD is the ratio of two functions, each of which is obtained from the gp, $f(\phi)$. Extend this to the multivariate PBD.

1.35. Prove the Factorization Theorem for sufficient statistics (Theorem 1.4).

1.36. For a discrete random variable taking nonnegative integer values, $\mathcal{T}$, show that

$$\mathbf{E}[\,\mathcal{T}\,] \;=\; \sum_{j\,=\,0}^{\infty} \mathrm{P}[\,\mathcal{T} > j\,]$$

Use this to find the mean of the geometric distribution.

1.37. Suppose the distribution of $\mathcal{T}$ given $\mathcal{N} = n$ is $\mathrm{B}(n, \pi)$, with the distribution of $\mathcal{N}$ being Poisson with mean λ. What is the unconditional distribution of $\mathcal{T}$? Is it a PBD?

1.38. Suppose the distribution of $\mathcal{T}$ given $\mathcal{N} = n$ is $\mathrm{B}(n, \pi)$, with the distribution of $\mathcal{N}$ being a negative binomial with r, the maximal number of failures allowed, and success probability, π_*. What is the unconditional distribution of $\mathcal{T}$? Is it a PBD?

1.39. Suppose the distribution of $\mathcal{T}$ given $\mathcal{R} = r$ is a negative binomial with r, the maximal number of failures allowed, and success probability, π. Further, the distribution of $\mathcal{R}$ is Poisson with mean λ. What is the unconditional distribution of $\mathcal{T}$? Is it a PBD?

1.40. Consider K independent events, $A_1, \cdots, A_K$. Show that the probability that at least one of them will occur is

$$\mathrm{P}[\,\bigcup_{k\,=1}^{K} A_k\,] \;=\; 1 \;-\; \prod_{k\,=\,1}^{K} (1 \;-\; \mathrm{P}[\,A_k\,])$$

CHAPTER 2

One-Sided Univariate Analysis

2.1 Introduction

This chapter introduces exact methods for analyzing data from polynomial based distributions. The corresponding asymptotic methods are also described. Its specific aims are:

- To formulate one-sided and two-sided hypotheses setups for the parameter of a statistical model.
- To show how the tail area of a probability distribution can be used as a measure of evidence for a one-sided hypothesis.
- To formulate one-sided conventional exact, mid-p exact and asymptotic evidence functions for univariate PBDs.
- To define one-sided p-values and confidence intervals, and relate them to the one-sided evidence function.
- To define the size, significance level and critical region of a statistical test.
- To begin the discussion, to be continued in later chapters, of the linkage between study design and method of data analysis.

While the binomial, Poisson and negative binomial distributions are used to illustrate the ideas of this chapter, they apply broadly to other naturally univariate PBDs and to those univariate PBDs constructed by conditioning from a multivariate PBD. For now, we focus on one-sided methods. Though less frequently applied in practice, they form a useful device for explaining the logic of the tail area approach to statistical inference, and a foundation upon which the more common two-sided methods are formulated.

2.2 One Parameter Inference

The scientific aims of a study often require assessing the value of one or more of the parameters of a statistical model. We seek, for example, to evaluate the diagnostic utility of a neural network algorithm for subjects who have had a myocardial infarction. Its actual but unknown false diagnosis rate among the patients with an infarct is π. We want to check if this rate exceeds π_0, the maximally acceptable rate.

We begin with defining a key idea in study design. Suppose we have a large population of subjects whose characteristics are defined by the random variable $\mathcal{T}$ (it may be a vector). Suppose we sample n subjects from this population. Let $\mathcal{T}_j$ denote the characteristic of the jth subject in this sample. This sample is said to be **a random sample** from the parent population if (i) for any t, $\mathrm{P}[\,\mathcal{T}_j = t\,] = \mathrm{P}[\,\mathcal{T} = t\,]$ and (ii) $\mathcal{T}_j$, $j = 1, \ldots, n$ are mutually independent.

A random sample has the feature that probability statements we make about it apply to the

population from which it was drawn as well. It is then said to protect the **generalizability** or **external validity** of the study.

In particular, for a random sample with binary variables, the expected mean value is the population proportion of the binary characteristic.

$$\mathbf{E}\left[\frac{1}{n}\sum_j \mathcal{T}_j\right] \;=\; \mathbf{E}[\,\mathcal{T}\,] \;=\; \pi$$

Example 2.1: Suppose $\pi_0 = 0.05$, or 5%. In one study, a particular neural network algorithm correctly identified 35 of the 36 patients who actually had a myocardial infarct. The observed error rate thereby was $1/36$ or 2.8% (Baxt 1991; Newman 1995). What do we conclude?

****

The objective of such a study may be stated as a choice between two competing hypotheses regarding the parameter of interest. They are the **null hypothesis**, denoted H_0, and the **alternative hypothesis**, denoted H_1. Several ways of framing such hypotheses exist. In a **one-sided setup,** the null and the alternative hypotheses represent a segment of the real line to the left or right of a specified value. For example, we write:

$$H_0 : \pi \;\leq\; \pi_0 \quad \text{versus} \quad H_1 : \pi \;>\; \pi_0 \tag{2.1}$$

Alternatively, the directions of the one-sided null and alternative hypotheses may be reversed. That is:

$$H_0 : \pi \;\geq\; \pi_0 \quad \text{versus} \quad H_1 : \pi \;<\; \pi_0 \tag{2.2}$$

One-sided hypotheses are also called directional hypotheses. At times, the one-sided hypothesis testing setup (for example, (2.2)) is, not quite accurately, written as testing a simple $H_0 : \pi = \pi_0$ versus $H_1 : \pi < \pi_0$.

****

Example 2.2: The gender of a newborn baby is a random entity. Some environmental factors may, however, alter the male to female ratio in the newborn babies. Irgens et al. (1997) looked at this ratio among the offsprings of workers who had been exposed to low frequency electromagnetic fields (LFEMF) on the job. Let us assume that the proportion of male offsprings in an unexposed but otherwise comparable reference population is known. Call it π_0, and let π be the unknown proportion of males in the LFEMF exposed population. Then we ask:

Does occupational exposure to LFEMF affect the offspring sex ratio? That is, is π different from π_0?

****

Such a query is formalized in a **two-sided hypotheses setup.** Here the null and alternative hypotheses are couched in terms of equivalence or nonequivalence to a given value.