Biostatistical Methods in Epidemiology

STEPHEN C. NEWMAN

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

Biostatistical Methods in Epidemiology

Biostatistical Methods in Epidemiology

STEPHEN C. NEWMAN

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This book is printed on acid-free paper. ♾

Library of Congress Cataloging-in-Publication Data:

Newman, Stephen C., 1952–
Biostatistical methods in epidemiology / Stephen C. Newman.
p. cm.—(Wiley series in probability and statistics. Biostatistics section)
Includes bibliographical references and index.
ISBN 0-471-36914-4 (cloth : alk. paper)
1. Epidemiology—Statistical methods. 2. Cohort analysis. I. Title. II. Series.
RA652.2.M3 N49 2001
614.4′07′27—dc21 2001028222

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Sandra

Contents

Preface

The aim of this book is to provide an overview of statistical methods that are important in the analysis of epidemiologic data, the emphasis being on nonregression techniques. The book is intended as a classroom text for students enrolled in an epidemiology or biostatistics program, and as a reference for established researchers. The choice and organization of material is based on my experience teaching biostatistics to epidemiology graduate students at the University of Alberta. In that setting I emphasize the importance of exploring data using nonregression methods prior to undertaking a more elaborate regression analysis. It is my conviction that most of what there is to learn from epidemiologic data can usually be uncovered using nonregression techniques.

I assume that readers have a background in introductory statistics, at least to the stage of simple linear regression. Except for the Appendices, the level of mathematics used in the book is restricted to basic algebra, although admittedly some of the formulas are rather complicated expressions. The concept of confounding, which is central to epidemiology, is discussed at length early in the book. To the extent permitted by the scope of the book, derivations of formulas are provided and relationships among statistical methods are identified. In particular, the correspondence between odds ratio methods based on the binomial model, and hazard ratio methods based on the Poisson model are emphasized (Breslow and Day, 1980, 1987). Historically, odds ratio methods were developed primarily for the analysis of case-control data. Students often find the case-control design unintuitive, and this can adversely affect their understanding of the odds ratio methods. Here, I adopt the somewhat unconventional approach of introducing odds ratio methods in the setting of closed cohort studies. Later in the book, it is shown how these same techniques can be adapted to the case-control design, as well as to the analysis of censored survival data. One of the attractive features of statistics is that different theoretical approaches often lead to nearly identical numerical results. I have attempted to demonstrate this phenomenon empirically by analyzing the same data sets using a variety of statistical techniques.

I wish to express my indebtedness to Allan Donner, Sander Greenland, John Hsieh, David Streiner, and Stephen Walter, who generously provided comments on a draft manuscript. I am especially grateful to Sander Greenland for his advice on the topic of confounding, and to John Hsieh who introduced me to life table theory when I was a student. The reviewers did not have the opportunity to read the final manuscript and so I alone am responsible for whatever shortcomings there may be in the book. I also wish to acknowledge the professionalism and commitment demonstrated by Steve Quigley and Lisa Van Horn of John Wiley & Sons. I am most interested in receiving your comments, which can be sent by e-mail using a link at the website www.stephennewman.com.

Prior to entering medicine and then epidemiology, I was deeply interested in a particularly elegant branch of theoretical mathematics called Galois theory. While studying the historical roots of the topic, I encountered a monograph having a preface that begins with the sentence "I wrote this book for myself." (Hadlock, 1978). After this remarkable admission, the author goes on to explain that he wanted to construct his own path through Galois theory, approaching the subject as an enquirer rather than an expert. Not being formally trained as a mathematical statistician, I embarked upon the writing of this book with a similar sense of discovery. The learning process was sometimes arduous, but it was always deeply rewarding. Even though I wrote this book partly "for myself," it is my hope that others will find it useful.

STEPHEN C. NEWMAN

Edmonton, Alberta, Canada
May 2001

CHAPTER 1

Introduction

In this chapter some background material from the theory of probability and statistics is presented that will be useful throughout the book. Such fundamental concepts as probability function, random variable, mean, and variance are defined, and several of the distributions that are important in the analysis of epidemiologic data are described. The Central Limit Theorem and normal approximations are discussed, and the maximum likelihood and weighted least squares methods of parameter estimation are outlined. The chapter concludes with a discussion of different types of random sampling. The presentation of material in this chapter is informal, the aim being to give an overview of some key ideas rather than provide a rigorous mathematical treatment. Readers interested in more complete expositions of the theoretical aspects of probability and statistics are referred to Cox and Hinkley (1974), Silvey (1975), Casella and Berger (1990), and Hogg and Craig (1994). References for the theory of probability and statistics in a health-related context are Armitage and Berry (1994), Rosner (1995), and Lachin (2000). For the theory of sampling, the reader is referred to Kish (1965) and Cochran (1977).

1.1 PROBABILITY

1.1.1 Probability Functions and Random Variables

Probability theory is concerned with mathematical models that describe phenomena having an element of uncertainty. Problems amenable to the methods of probability theory range from the elementary, such as the chance of randomly selecting an ace from a well-shuffled deck of cards, to the exceedingly complex, such as predicting the weather. Epidemiologic studies typically involve the collection, analysis, and interpretation of health-related data where uncertainty plays a role. For example, consider a survey in which blood sugar is measured in a random sample of the population. The aims of the survey might be to estimate the average blood sugar in the population and to estimate the proportion of the population with diabetes (elevated blood sugar). Uncertainty arises because there is no guarantee that the resulting estimates will equal the true population values (unless the entire population is enrolled in the survey).

Associated with each probability model is a random variable, which we denote by a capital letter such as X. We can think of X as representing a potential data point for a proposed study. Once the study has been conducted, we have actual data points that will be referred to as realizations (outcomes) of X. An arbitrary realization of X will be denoted by a small letter such as x. In what follows we assume that realizations are in the form of numbers so that, in the above survey, diabetes status would have to be coded numerically—for example, 1 for present and 0 for absent. The set of all possible realizations of X will be referred to as the sample space of X. For blood sugar the sample space is the set of all nonnegative numbers, and for diabetes status (with the above coding scheme) the sample space is $\{0, 1\}$. In this book we assume that all sample spaces are either continuous, as in the case of blood sugar, or discrete, as in the case of diabetes status. We say that X is continuous or discrete in accordance with the sample space of the probability model.

There are several mathematically equivalent ways of characterizing a probability model. In the discrete case, interest is mainly in the probability mass function, denoted by $P(X = x)$, whereas in the continuous case the focus is usually on the probability density function, denoted by $f(x)$. There are important differences between the probability mass function and the probability density function, but for present purposes it is sufficient to view them simply as formulas that can be used to calculate probabilities. In order to simplify the exposition we use the term probability function to refer to both these constructs, allowing the context to make the distinction clear. Examples of probability functions are given in Section 1.1.2. The notation $P(X = x)$ has the potential to be confusing because both X and x are “variables.” We read $P(X = x)$ as the probability that the discrete random variable X has the realization x. For simplicity it is often convenient to ignore the distinction between X and x. In particular, we will frequently use x in formulas where, strictly speaking, X should be used instead.

The correspondence between a random variable and its associated probability function is an important concept in probability theory, but it needs to be emphasized that it is the probability function which is the more fundamental notion. In a sense, the random variable represents little more than a convenient notation for referring to the probability function. However, random variable notation is extremely powerful, making it possible to express in a succinct manner probability statements that would be cumbersome otherwise. A further advantage is that it may be possible to specify a random variable of interest even when the corresponding probability function is too difficult to describe explicitly. In what follows we will use several expressions synonymously when describing random variables. For example, when referring to the random variable associated with a binomial probability function we will variously say that the random variable “has a binomial distribution,” “is binomially distributed,” or simply “is binomial.”

We now outline a few of the key definitions and results from introductory probability theory. For simplicity we focus on discrete random variables, keeping in mind that equivalent statements can be made for the continuous case. One of the defining properties of a probability function is the identity

$$\sum_x P(X = x) = 1 \tag{1.1}$$

where here, and in what follows, the summation is over all elements in the sample space of X. Next we define two fundamental quantities that will be referred to repeatedly throughout the book. The mean of X, sometimes called the expected value of X, is defined to be

$$E(X) = \sum_x x\,P(X = x) \tag{1.2}$$

and the variance of X is defined to be

$$\text{var}(X) = \sum_x [x - E(X)]^2 P(X = x). \tag{1.3}$$

It is important to note that when the mean and variance exist, they are constants, not random variables. In most applications the mean and variance are unknown and must be estimated from study data. In what follows, whenever we refer to the mean or variance of a random variable it is being assumed that these quantities exist—that is, are finite constants.

Example 1.1 Consider the probability function given in Table 1.1. Evidently (1.1) is satisfied. The sample space of X is {0, 1, 2}, and the mean and variance of X are

$$E(X) = (0 \times .20) + (1 \times .50) + (2 \times .30) = 1.1$$

and

$$\text{var}(X) = [(0 - 1.1)^2.20] + [(1 - 1.1)^2.50] + [(2 - 1.1)^2.30] = .49.$$

Transformations can be used to derive new random variables from an existing random variable. Again we emphasize that what is meant by such a statement is that we can derive new probability functions from an existing probability function. When the probability function at hand has a known formula it is possible, in theory, to write down an explicit formula for the transformed probability function. In practice, this

TABLE 1.1 Probability Function of X

x	$P(X = x)$
0	.20
1	.50
2	.30

TABLE 1.2 Probability Function of Y

y	$P(Y = y)$
5	.20
7	.50
9	.30

may lead to a very complicated expression, which is one of the reasons for relying on random variable notation.

Example 1.2 With X as in Example 1.1, consider the random variable $Y = 2X + 5$. The sample space of Y is obtained by applying the transformation to the sample space of X, which gives $\{5, 7, 9\}$. The values of $P(Y = x)$ are derived as follows: $P(Y = 7) = P(2X + 5 = 7) = P(X = 1) = .50$. The probability function of Y is given in Table 1.2.

The mean and variance of Y are

$$E(Y) = (5 \times .20) + (7 \times .50) + (9 \times .30) = 7.2$$

and

$$\text{var}(Y) = [(5 - 7.2)^2 .20] + [(7 - 7.2)^2 .50] + [(9 - 7.2)^2 .30] = 1.96.$$

Comparing Examples 1.1 and 1.2 we note that X and Y have the same probability values but different sample spaces.

Consider a random variable which has as its only outcome the constant β, that is, the sample space is $\{\beta\}$. It is immediate from (1.2) and (1.3) that the mean and variance of the random variable are β and 0, respectively. Identifying the random variable with the constant β, and allowing a slight abuse of notation, we can write $E(\beta) = \beta$ and $\text{var}(\beta) = 0$. Let X be a random variable, let α and β be arbitrary constants, and consider the random variable $\alpha X + \beta$. Using (1.2) and (1.3) it can be shown that

$$E(\alpha X + \beta) = \alpha E(X) + \beta \tag{1.4}$$

and

$$\text{var}(\alpha X + \beta) = \alpha^2 \,\text{var}(X). \tag{1.5}$$

Applying these results to Examples 1.1 and 1.2 we find, as before, that $E(Y) = 2(1.1) + 5 = 7.2$ and $\text{var}(Y) = 4(.49) = 1.96$.

Example 1.3 Let X be an arbitrary random variable with mean μ and variance σ^2, where $\sigma > 0$, and consider the random variable $(X - \mu)/\sigma$. With $\alpha = 1/\sigma$ and

$\beta = -\mu/\sigma$ in (1.4) and (1.5), it follows that

$$E\left(\frac{X-\mu}{\sigma}\right) = 0$$

and

$$\operatorname{var}\left(\frac{X-\mu}{\sigma}\right) = 1.$$

In many applications it is necessary to consider several related random variables. For example, in a health survey we might be interested in age, weight, and blood pressure. A probability function characterizing two or more random variables simultaneously is referred to as their joint probability function. For simplicity we discuss the case of two discrete random variables, X and Y. The joint probability function of the pair of random variables (X, Y) is denoted by $P(X = x, Y = y)$. For the present discussion we assume that the sample space of the joint probability function is the set of pairs $\{(x, y)\}$, where x is in the sample space of X and y is in the sample space of Y. Analogous to (1.1), the identity

$$\sum_x \sum_y P(X = x, Y = y) = 1 \tag{1.6}$$

must be satisfied. In the joint distribution of X and Y, the two random variables are considered as a unit. In order to isolate the distribution of X, we "sum over" Y to obtain what is referred to as the marginal probability function of X,

$$P(X = x) = \sum_y P(X = x, Y = y).$$

Similarly, the marginal probability function of Y is

$$P(Y = y) = \sum_x P(X = x, Y = y).$$

From a joint probability function we are to able obtain marginal probability functions, but the process does not necessarily work in reverse. We say that X and Y are independent random variables if $P(X = x, Y = y) = P(X = x)\, P(Y = y)$, that is, if the joint probability function is the product of the marginal probability functions. Other than the case of independence, it is not generally possible to reconstruct a joint probability function in this way.

Example 1.4 Table 1.3 is an example of a joint probability function and its associated marginal probability functions. For example, $P(X = 1, Y = 3) = .30$. The marginal probability function of X is obtained by summing over Y, for example,

$$P(X = 1) = P(X = 1, Y = 1) + P(X = 1, Y = 2) + P(X = 1, Y = 3) = .50.$$

TABLE 1.3 Joint Probability Function of X and Y

	$P(X=x, Y=y)$			
	y			
x	1	2	3	$P(X=x)$
0	.02	.06	.12	.20
1	.05	.15	.30	.50
2	.03	.09	.18	.30
$P(Y=y)$	.10	.30	.60	1

It is readily verified that X and Y are independent, for example, $P(X = 1, Y = 2) = .15 = P(X = 1)\,P(Y = 2)$.

Now consider Table 1.4, where the marginal probability functions of X and Y are the same as in Table 1.3 but where, as is easily verified, X and Y are not independent.

We now present generalizations of (1.4) and (1.5). Let $X_1, X_2, \ldots, X_n$ be arbitrary random variables, let $\alpha_1, \alpha_2, \ldots, \alpha_n, \beta$ be arbitrary constants, and consider the random variable $\sum_{i=1}^{n} \alpha_i X_i + \beta$. It can be shown that

$$E\left(\sum_{i=1}^{n} \alpha_i X_i + \beta\right) = \sum_{i=1}^{n} \alpha_i E(X_i) + \beta \tag{1.7}$$

and, if the X_i are independent, that

$$\operatorname{var}\left(\sum_{i=1}^{n} \alpha_i X_i + \beta\right) = \sum_{i=1}^{n} \alpha_i^2 \operatorname{var}(X_i). \tag{1.8}$$

In the case of two independent random variables X_1 and X_2,

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

$$E(X_1 - X_2) = E(X_1) - E(X_2)$$

TABLE 1.4 Joint Probability Function of X and Y

	$P(X=x, Y=y)$			
	y			
x	1	2	3	$P(X=x)$
0	.01	.05	.14	.20
1	.06	.18	.26	.50
2	.03	.07	.20	.30
$P(Y=y)$	.10	.30	.60	1

and

$$\mathrm{var}(X_1 + X_2) = \mathrm{var}(X_1 - X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2). \tag{1.9}$$

If $X_1, X_2, \ldots, X_n$ are independent and all have the same distribution, we say the X_i are a sample from that distribution and that the sample size is n. Unless stated otherwise, it will be assumed that all samples are simple random samples (Section 1.3). With the distribution left unspecified, denote the mean and variance of X_i by μ and σ^2, respectively. The sample mean is defined to be

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Setting $\alpha_i = 1/n$ and $\beta = 0$ in (1.7) and (1.8), we have

$$E(\overline{X}) = \mu \tag{1.10}$$

and

$$\mathrm{var}(\overline{X}) = \frac{\sigma^2}{n}. \tag{1.11}$$

1.1.2 Some Probability Functions

We now consider some of the key probability functions that will be of importance in this book.

Normal (Gaussian)

For reasons that will become clear after we have discussed the Central Limit Theorem, the most important distribution is undoubtedly the normal distribution. The normal probability function is

$$f(z|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[\frac{-(z-\mu)^2}{2\sigma^2}\right]$$

where the sample space is all numbers and exp stands for exponentiation to the base e. We denote the corresponding normal random variable by Z. A normal distribution is completely characterized by the parameters μ and $\sigma > 0$. It can be shown that the mean and variance of Z are μ and σ^2, respectively.

When $\mu = 0$ and $\sigma = 1$ we say that Z has the standard normal distribution. For $0 < \gamma < 1$, let z_γ denote that point which cuts off the upper γ-tail probability of the standard normal distribution; that is, $P(Z \geq z_\gamma) = \gamma$. For example, $z_{.025} = 1.96$. In some statistics books the notation z_γ is used to denote the lower γ-tail. An important property of the normal distribution is that, for arbitrary constants α and $\beta > 0$, $(Z-\alpha)/\beta$ is also normally distributed. In particular this is true for $(Z-\mu)/\sigma$ which, in view of Example 1.3, is therefore standard normal. This explains why statistics books only need to provide values of z_γ for the standard normal distribution rather than a series of tables for different values of μ and σ.

Another important property of the normal distribution is that it is additive. Let $Z_1, Z_2, \ldots, Z_n$ be independent normal random variables and suppose that Z_i has mean μ_i and variance σ_i^2 $(i = 1, 2, \ldots, n)$. Then the random variable $\sum_{i=1}^{n} Z_i$ is also normally distributed and, from (1.7) and (1.8), it has mean $\sum_{i=1}^{n} \mu_i$ and variance $\sum_{i=1}^{n} \sigma_i^2$.

Chi-Square

The formula for the chi-square probability function is complicated and will not be presented here. The sample space of the distribution is all nonnegative numbers. A chi-square distribution is characterized completely by a single positive integer r, which is referred to as the degrees of freedom. For brevity we write $\chi^2_{(r)}$ to indicate that a random variable has a chi-square distribution with r degrees of freedom. The mean and variance of the chi-square distribution with r degrees of freedom are r and $2r$, respectively.

The importance of the chi-square distribution stems from its connection with the normal distribution. Specifically, if Z is standard normal, then Z^2, the transformation of Z obtained by squaring, is $\chi^2_{(1)}$. More generally, if Z is normal with mean μ and variance σ^2 then, as remarked above, $(Z - \mu)/\sigma$ is standard normal and so $[(Z - \mu)/\sigma]^2 = (Z - \mu)^2/\sigma^2$ is $\chi^2_{(1)}$. In practice, most chi-square distributions with 1 degree of freedom originate as the square of a standard normal distribution. This explains why the usual notation for a chi-square random variable is X^2, or sometimes χ^2.

Like the normal distribution, the chi-square distribution has an additive property. Let $X_1^2, X_2^2, \ldots, X_n^2$ be independent chi-square random variables and suppose that X_i^2 has r_i degrees of freedom $(i = 1, 2, \ldots, n)$. Then $\sum_{i=1}^{n} X_i^2$ is chi-square with $\sum_{i=1}^{n} r_i$ degrees of freedom. As a special case of this result, let $Z_1, Z_2, \ldots, Z_n$ be independent normal random variables, where Z_i has mean μ_i and variance σ_i^2 $(i = 1, 2, \ldots, n)$. Then $(Z_i - \mu_i)^2/\sigma_i^2$ is $\chi^2_{(1)}$ for all i, and so

$$X^2 = \sum_{i=1}^{n} \frac{(Z_i - \mu_i)^2}{\sigma_i^2} \tag{1.12}$$

is $\chi^2_{(n)}$.

Binomial

The binomial probability function is

$$P(A = a|\pi) = \binom{r}{a} \pi^a (1 - \pi)^{r-a}$$

where the sample space is the (finite) set of integers $\{0, 1, 2, \ldots, r\}$. A binomial distribution is completely characterized by the parameters π and r which, for convenience, we usually write as (π, r). Recall that, for $0 \leq a \leq r$, the binomial coefficient is defined to be

$$\binom{r}{a} = \frac{r!}{a!\,(r-a)!}$$

where $r! = r\,(r-1)\cdots 2 \cdot 1$. We adopt the usual convention that $0! = 1$. The binomial coefficient $\binom{r}{a}$ equals the number of ways of choosing a items out of r without regard to order of selection. For example, the number of possible bridge hands is $\binom{52}{13} = 6.35 \times 10^{11}$. It can be shown that

$$\sum_{a=0}^{r} \binom{r}{a} \pi^a (1-\pi)^{r-a} = [\pi + (1-\pi)]^r = 1$$

and so (1.1) is satisfied. The mean and variance of A are πr and $\pi(1-\pi)r$, respectively; that is,

$$E(A) = \sum_{a=0}^{r} a \binom{r}{a} \pi^a (1-\pi)^{r-a} = \pi r$$

and

$$\text{var}(A) = \sum_{a=0}^{r} (a - \pi r)^2 \binom{r}{a} \pi^a (1-\pi)^{r-a} = \pi(1-\pi)r.$$

Like the normal and chi-square distributions, the binomial distribution is additive. Let $A_1, A_2, \ldots, A_n$ be independent binomial random variables and suppose that A_i has parameters $\pi_i = \pi$ and r_i $(i = 1, 2, \ldots, n)$. Then $\sum_{i=1}^{n} A_i$ is binomial with parameters π and $\sum_{i=1}^{n} r_i$. A similar result does not hold when the π_i are not all equal.

The binomial distribution is important in epidemiology because many epidemiologic studies are concerned with counted (discrete) outcomes. For instance, the binomial distribution can be used to analyze data from a study in which a group of r individuals is followed over a defined period of time and the number of outcomes of interest, denoted by a, is counted. In this context the outcome of interest could be, for example, recovery from an illness, survival to the end of follow-up, or death from some cause. For the binomial distribution to be applicable, two conditions need to be satisfied: The probability of an outcome must be the same for each subject, and subjects must behave independently; that is, the outcome for each subject must be unrelated to the outcome for any other subject. In an epidemiologic study the first condition is unlikely to be satisfied across the entire group of subjects. In this case, one strategy is to form subgroups of subjects having similar characteristics so that, to a greater or lesser extent, there is uniformity of risk within each subgroup. Then the binomial distribution can be applied to each subgroup separately. As an example where the second condition would not be satisfied, consider a study of influenza in a classroom of students. Since influenza is contagious, the risk of illness in one student is not independent of the risk in others. In studies of noninfectious diseases, such as cancer, stroke, and so on, the independence assumption is usually satisfied.

Poisson

The Poisson probability function is

$$P(D = d|\nu) = \frac{e^{-\nu}\nu^d}{d!} \tag{1.13}$$

where the sample space is the (infinite) set of nonnegative integers $\{0, 1, 2, \ldots\}$. A Poisson distribution is completely characterized by the parameter ν, which is equal to both the mean and variance of the distribution, that is,

$$E(D) = \sum_{d=0}^{\infty} d\left(\frac{e^{-\nu}\nu^d}{d!}\right) = \nu$$

and

$$\text{var}(D) = \sum_{d=0}^{\infty} (d - \nu)^2\left(\frac{e^{-\nu}\nu^d}{d!}\right) = \nu.$$

Similar to the other distributions considered above, the Poisson distribution has an additive property. Let $D_1, D_2, \ldots, D_n$ be independent Poisson random variables, where D_i has the parameter ν_i $(i = 1, 2, \ldots, n)$. Then $\sum_{i=1}^{n} D_i$ is Poisson with parameter $\sum_{i=1}^{n} \nu_i$.

Like the binomial distribution, the Poisson distribution can be used to analyze data from a study in which a group of individuals is followed over a defined period of time and the number of outcomes of interest, denoted by d, is counted. In epidemiologic studies where the Poisson distribution is applicable, it is not the number of subjects that is important but rather the collective observation time experienced by the group as a whole. For the Poisson distribution to be valid, the probability that an outcome will occur at any time point must be “small.” Expressed another way, the outcome must be a “rare” event.

As might be guessed from the above remarks, there is a connection between the binomial and Poisson distributions. In fact the Poisson distribution can be derived as a limiting case of the binomial distribution. Let D be Poisson with mean ν, and let $A_1, A_2, \ldots, A_i, \ldots$ be an infinite sequence of binomial random variables, where A_i has parameters (π_i, r_i). Suppose that the sequence satisfies the following conditions: $\pi_i r_i = \nu$ for all i, and the limiting value of π_i equals 0. Under these circumstances the sequence of binomial random variables “converges” to D; that is, as i gets larger the distribution of A_i gets closer to that of D. This theoretical result explains why the Poisson distribution is often used to model rare events. It also suggests that the Poisson distribution with parameter ν can be used to approximate the binomial distribution with parameters (π, r), provided $\nu = \pi r$ and π is “small.”

TABLE 1.5 Binomial and Poisson Probability Functions (%)

	Binomial			
	$\pi = .2$	$\pi = .1$	$\pi = .01$	Poisson
x	$r = 10$	$r = 20$	$r = 200$	$\nu = 2$
0	10.74	12.16	13.40	13.53
1	26.84	27.02	27.07	27.07
2	30.20	28.52	27.20	27.07
3	20.13	19.01	18.14	18.04
4	8.81	8.98	9.02	9.02
5	2.64	3.19	3.57	3.61
6	.55	.89	1.17	1.20
7	.08	.20	.33	.34
8	.01	.04	.08	.09
9	< .01	.01	.02	.02
10	< .01	< .01	< .01	< .01
⋮	—	⋮	⋮	⋮

Example 1.5 Table 1.5 gives three binomial distributions with parameters $(.2, 10)$, $(.1, 20)$, and $(.01, 200)$, so that in each case the mean is 2. Also shown is the Poisson distribution with a mean of 2. The sample spaces have been truncated at 10. As can be seen, as π becomes smaller the Poisson distribution provides a progressively better approximation to the binomial distribution.

1.1.3 Central Limit Theorem and Normal Approximations

Let $X_1, X_2, \ldots, X_n$ be a sample from an arbitrary distribution and denote the common mean and variance by μ and σ^2. It was shown in (1.10) and (1.11) that $\overline{X}$ has mean $E(\overline{X}) = \mu$ and variance $\text{var}(\overline{X}) = \sigma^2/n$. So, from Example 1.3, the random variable $\sqrt{n}(\overline{X}-\mu)/\sigma$ has mean 0 and variance 1. If the X_i are normal then, from the properties of the normal distribution, $\sqrt{n}(\overline{X} - \mu)/\sigma$ is standard normal. The Central Limit Theorem is a remarkable result from probability theory which states that, even when the X_i are not normal, $\sqrt{n}(\overline{X}-\mu)/\sigma$ is "approximately" standard normal, provided n is sufficiently "large." We note that the X_i are not required to be continuous random variables. Probability statements such as this, which become more accurate as n increases, are said to hold asymptotically. Accordingly, the Central Limit Theorem states that $\sqrt{n}(\overline{X} - \mu)/\sigma$ is asymptotically standard normal.

Let A be binomial with parameters (π, n) and let $A_1, A_2, \ldots, A_n$ be a sample from the binomial distribution with parameters $(\pi, 1)$. Similarly, let D be Poisson with parameter ν, where we assume that $\nu = n$, an integer, and let $D_1, D_2, \ldots, D_n$ be a sample from the Poisson distribution with parameter 1. From the additive properties of binomial and Poisson distributions, A has the same distribution as $\sum_{i=1}^{n} A_i$, and D has the same distribution as $\sum_{i=1}^{n} D_i$. It follows from the Central Limit Theorem that, provided n is large, A and D will be asymptotically normal. We illustrate this phenomenon below with a series of graphs.

Let $D_1, D_2, \ldots, D_n$ be independent Poisson random variables, where D_i has the parameter ν_i $(i = 1, 2, \ldots, n)$. From the arguments leading to (1.12) and the Central Limit Theorem, it follows that

$$X^2 = \sum_{i=1}^{n} \frac{(D_i - \nu_i)^2}{\nu_i} \tag{1.14}$$

is approximately $\chi^2_{(n)}$. More generally, let $X_1, X_2, \ldots, X_n$ be independent random variables where X_i has mean μ_i and variance σ_i^2 $(i = 1, 2, \ldots, n)$. If each X_i is approximately normal then

$$X^2 = \sum_{i=1}^{n} \frac{(X_i - \mu_i)^2}{\sigma_i^2} \tag{1.15}$$

is approximately $\chi^2_{(n)}$.

Example 1.6 Table 1.6(a) gives the exact and approximate values of the lower and upper tail probabilities of the binomial distribution with parameters (.3, 10). In statistics the term “exact” means that an actual probability function is being used to perform calculations, as opposed to a normal approximation. The mean and variance of the binomial distribution are .3(10) = 3 and .3(.7)(10) = 2.1. The approximate values were calculated using the following approach. The normal approximation to $P(A \leq 2 \,|\, .3)$, for example, equals the area under the standard normal curve to the left of $[(2+.5)-3]/\sqrt{2.1}$, and the normal approximation to $P(A \geq 2 \,|\, .3)$ equals the area under the standard normal curve to the right of $[(2-.5)-3]/\sqrt{2.1}$. The continuity correction factors ±.5 have been included because the normal distribution, which is continuous, is being used to approximate a binomial distribution, which is discrete (Breslow and Day, 1980, §4.3). As can be seen from Table 1.6(a), the exact and approximate values show quite good agreement. Table 1.6(b) gives the results for the

TABLE 1.6(a) Exact and Approximate Tail Probabilities (%) for the Binomial Distribution with Parameters (.3,10)

	$P(A \leq a \,|\, .3)$		$P(A \geq a \,|\, .3)$	
a	Exact	Approximate	Exact	Approximate
2	38.28	36.50	85.07	84.97
4	84.97	84.97	35.04	36.50
6	98.94	99.21	4.73	4.22
8	99.99	99.99	.16	.10

TABLE 1.6(b) Exact and Approximate Tail Probabilities (%) for the Binomial Distribution with Parameters (.3,100)

	$P(A \leq a \mid .3)$		$P(A \geq a \mid .3)$	
a	Exact	Approximate	Exact	Approximate
20	1.65	1.91	99.11	98.90
25	16.31	16.31	88.64	88.50
30	54.91	54.34	53.77	54.34
35	88.39	88.50	16.29	16.31
40	98.75	98.90	2.10	1.91

binomial distribution with parameters (.3,100), which shows even better agreement due to the larger sample size.

Arguments were presented above which show that binomial and Poisson distributions are approximately normal when the sample size is large. The obvious question is, How large is "large"? We approach this matter empirically and present a sample size criterion that is useful in practice. The following remarks refer to Figures 1.1(a)–1.8(a), which show graphs of selected binomial and Poisson distributions. The points in the sample space have been plotted on the horizontal axis, with the corresponding probabilities plotted on the vertical axis. Magnitudes have not been indicated on the axes since, for the moment, we are concerned only with the shapes of distributions. The horizontal axes are labeled with the term "count," which stands for the number of binomial or Poisson outcomes. Distributions with the symmetric, bell-shaped appearance of the normal distribution have a satisfactory normal approximation.

The binomial and Poisson distributions have sample spaces consisting of consecutive integers, and so the distance between neighboring points is always 1. Consequently the graphs could have been presented in the form of histograms (bar charts). Instead they are shown as step functions so as to facilitate later comparisons with the remaining graphs in the same figures. Since the base of each step has a length of 1, the area of the rectangle corresponding to that step equals the probability associated with that point in the sample space. Consequently, summing across the entire sample space, the area under each step function equals 1, as required by (1.1). Some of the distributions considered here have tails with little associated probability (area). This is obviously true for the Poisson distributions, where the sample space is infinite and extreme tail probabilities are small. The graphs have been truncated at the extremes of the distributions corresponding to tail probabilities of 1%.

The binomial parameters used to create Figures 1.1(a)–1.5(a) are (.3,10), (.5,10), (.03,100), (.05,100), and (.1,100), respectively, and so the means are 3, 5, and 10. The Poisson parameters used to create Figures 1.6(a)–1.8(a) are 3, 5, and 10, which are also the means of the distributions. As can be seen, for both the binomial and Poisson distributions, a rough guideline is that the normal approximation should be satisfactory provided the mean of the distribution is greater than or equal to 5.

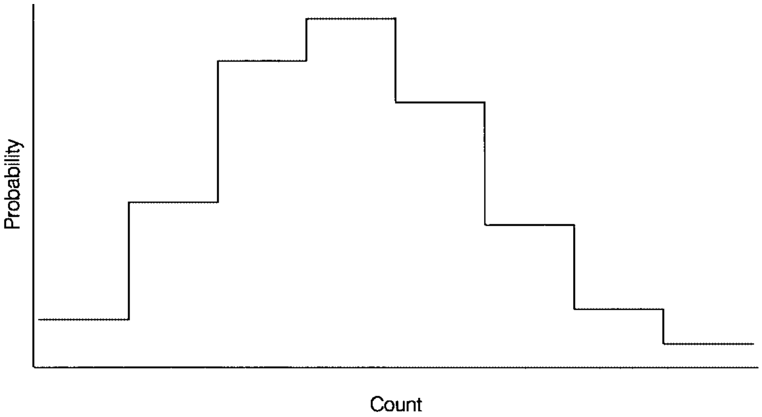

FIGURE 1.1(a) Binomial distribution with parameters (.3, 10)

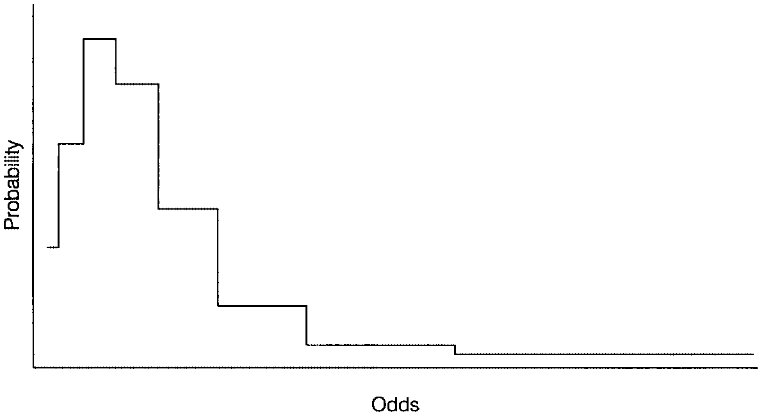

FIGURE 1.1(b) Odds transformation of binomial distribution with parameters (.3, 10)

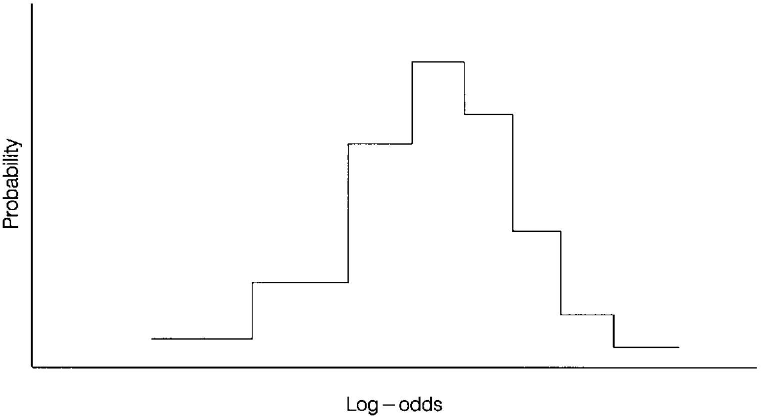

FIGURE 1.1(c) Log-odds transformation of binomial distribution with parameters (.3, 10)

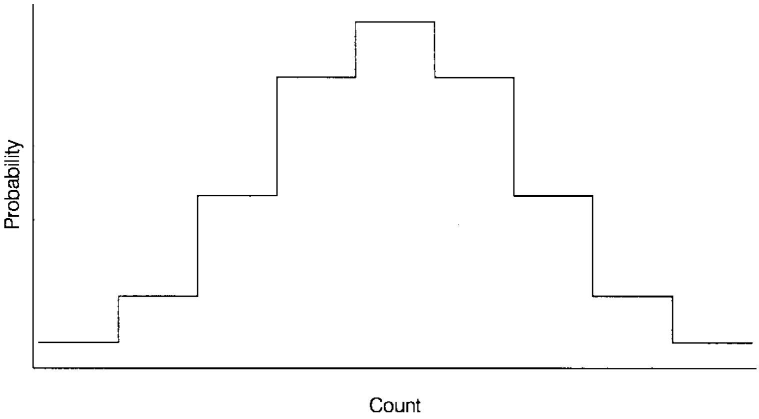

FIGURE 1.2(a) Binomial distribution with parameters (.5, 10)

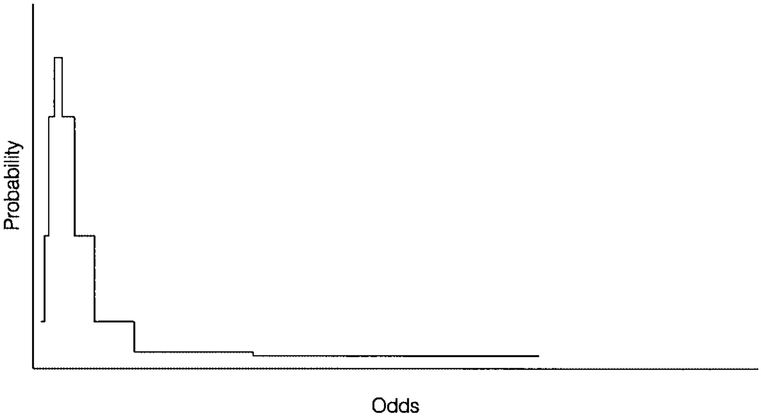

FIGURE 1.2(b) Odds transformation of binomial distribution with parameters (.5, 10)

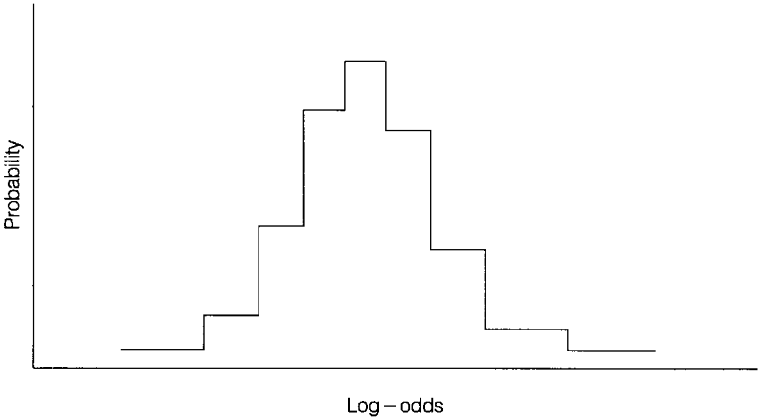

FIGURE 1.2(c) Log-odds transformation of binomial distribution with parameters (.5, 10)

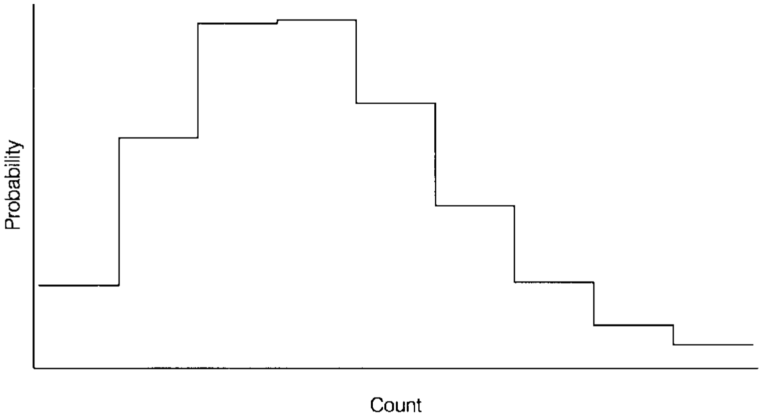

FIGURE 1.3(a) Binomial distribution with parameters (.03, 100)

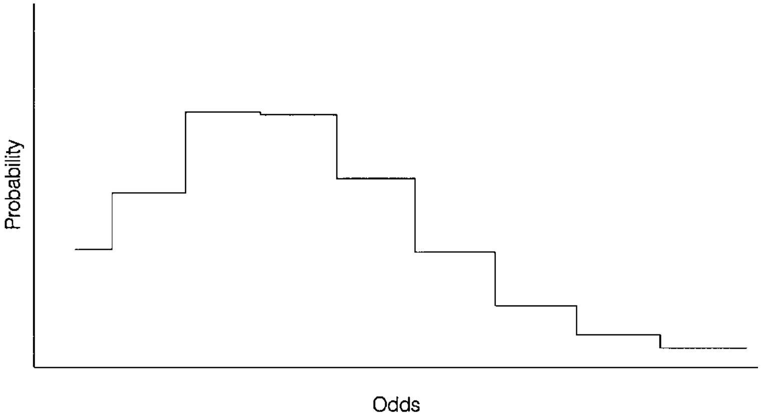

FIGURE 1.3(b) Odds transformation of binomial distribution with parameters (.03, 100)

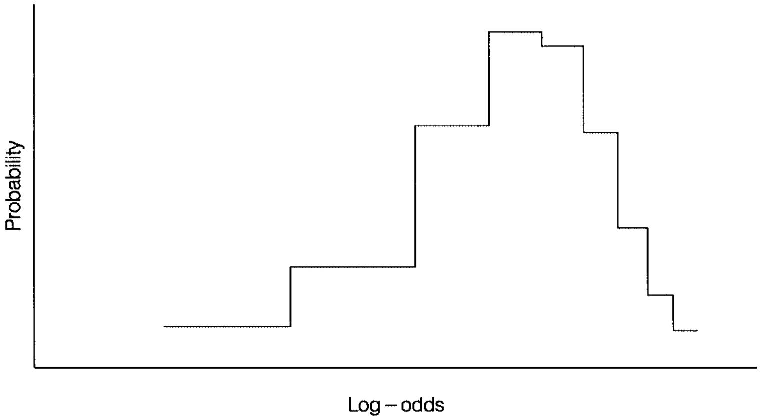

FIGURE 1.3(c) Log-odds transformation of binomial distribution with parameters (.03, 100)