

# 1

## Introduction

These are the fundamental properties, definitions, and general building blocks for everything that follows. Section 1.1 describes discrete distributions in general. The following sections of this chapter provide specific examples of useful and popular models. The reader is referred to (Johnson, Kotz, and Kemp (1992)) as a more comprehensive and definitive resource to this material for univariate distributions. Johnson, Kotz, and Balakrishnan (1997) is the corresponding reference for multivariate discrete distributions.

### 1.1 Discrete Distributions in General

We will describe properties of a discrete valued random variable  $Y$  whose support is taken to be  $0, 1, \dots$  or a finite subset of these nonnegative integers. In general, random variables are denoted by capital letters and their observed values by lower case letters.

The simplest example of a random variable is the *Bernoulli random variable* for which

$$\Pr[ Y = 1 ] = p \quad \text{and} \quad \Pr[ Y = 0 ] = 1 - p \quad (1.1)$$

for probability parameter  $p$  satisfying  $0 \leq p \leq 1$ . The values of  $Y = 1$  and  $Y = 0$  are often referred to as successes and failures, respectively.

More generally, for  $y = 0, 1, \dots$  let

$$\Pr[ Y = y ] = p_y$$

for  $p_y \geq 0$  with  $\sum p_y = 1$ .

The function

$$f(y) = p_y$$

for  $y = 0, 1, \dots$  and zero elsewhere is called the *probability mass function* or the *mass function*.

The function

$$F(y) = \sum_{j=0}^y p_j = \Pr[Y \leq y]$$

is called the *cumulative distribution function* or the *distribution function*.

The value  $\tilde{y}$  for which  $p_{\tilde{y}}$  is a maximum is called the *mode* of the distribution. The mode may not be unique. There may be local modes as well. These ideas are illustrated in the distribution plotted in Fig. 2.3 and discussed in Section 2.2.3.

The *expected value of a function*  $g(Y)$  of the random variable  $Y$  is

$$E[g(Y)] = \sum_y g(y) \Pr[Y = y] = \sum_y g(y)f(y)$$

provided this sum converges.

Expectation is a linear operation so

$$E[g(Y) + h(Y)] = E[g(Y)] + E[h(Y)].$$

The specific example in which  $g(Y)$  is a power of  $Y$  is called a *moment*. The expectation

$$E[Y^j] = \sum_y y^j \Pr[Y = y]$$

for  $j = 1, 2, \dots$  is called the *jth moment* of  $Y$  or sometimes the *jth moment about zero*.

The first moment  $E[Y]$  is called the *mean* or the *expected value*. Let us denote the mean of  $Y$  by  $\mu$ . Moments of  $Y$  about its mean are called *central moments* so that

$$E[(Y - \mu)^j]$$

is called the *jth central moment*.

The first central moment is zero and the second central moment is called the *variance* of  $Y$ . This is denoted by

$$\text{Var}[Y] = E[(Y - \mu)^2]$$

Expanding the square here shows

$$\text{Var}[Y] = E[Y^2] - \mu^2.$$

The square root of the variance is the *standard deviation*.

Higher central moments are usually standardized by the standard deviation. The third central standardized moment

$$E[(Y - \mu)^3] / \{\text{Var}[Y]\}^{3/2}$$

is called the *skewness*.

The *moment generating function* of  $Y$

$$M_Y(t) = E[e^{tY}] = \sum_y e^{ty} \Pr[Y = y]$$

is the expected value of  $e^{tY}$  and may only be defined for values of  $t$  in a neighborhood of zero.

The moment generating function is so called because successive derivatives at zero are equal to the moments for many distributions. Specifically, for all of the distributions we will discuss

$$(d/dt)^j M_Y(t) \Big|_{t=0} = E[Y^j]$$

for every  $j = 1, 2, \dots$ . In all generality, not all the moments exist for all the distributions and the moment generating function may fail to generate the moments even when these exist.

The *characteristic function* of  $Y$

$$\phi_Y(t) = E[e^{itY}] = \sum_y e^{ity} \Pr[Y = y]$$

is the expected value of  $e^{itY}$ , where  $i^2 = -1$ . Unlike the moment generating function, the characteristic function always exists.

The *probability generating function* is

$$G(t) = E[t^Y] = \sum_y t^y \Pr[Y = y]. \tag{1.2}$$

The probability generating function has the property that successive derivatives at  $t = 0$  are equal to the individual probabilities of  $Y$  in the sense that

$$(d/dt)^j G(t) \Big|_{t=0} = j! \Pr[Y = y]$$

for  $j = 0, 1, \dots$

The relation between the probability generating function and the characteristic function is

$$\phi_Y(t) = G_Y(e^{it}).$$

For many discrete distributions, the factorial moments are available in a convenient functional form. For  $k = 1, 2, \dots$ , define the *factorial polynomial*

$$z^{(k)} = z(z-1)\cdots(z-k+1) \tag{1.3}$$

and  $z^{(0)} = 1$ .

The *factorial moments* of  $Y$  are then

$$E[Y^{(k)}] = E[Y(Y-1)\cdots(Y-k+1)].$$

These are also referred to as the *descending factorial moments* by some authors.

Similarly, we have

$$E[Y] = E[Y^{(1)}]$$

and

$$\text{Var}[Y] = E[Y^{(2)}] + E[Y] - \{E[Y]\}^2.$$

The *factorial moment generating function* is

$$G(t+1) = E[(t+1)^Y],$$

where  $G$  is the probability generating function given in (1.2).

The factorial moment generating function has the property

$$\left. \frac{(d/dt)^j G(1+t)}{t=0} \right|_{t=0} = E[Y^{(j)}]$$

for many distributions.

There are, of course, many other useful generating functions available with important relationships between them. Those listed here are limited to those specifically referred to in the volume. The reader is referred to (Johnson, Kotz, and Kemp (1992, Section 1.B)) for a more thorough treatment of the subject.

## 1.2 Multivariate Discrete Distributions

The joint mass function of the random variables  $\{X_1, \dots, X_k\}$  is the function

$$f(x_1, \dots, x_k) = \Pr[X_1 = x_1, \dots, X_k = x_k].$$

The mass function is never negative over the range of  $\{X_1, \dots, X_k\}$  and sums to one:

$$\sum_{x_1} \cdots \sum_{x_k} f(x_1, \dots, x_k) = 1.$$

Marginal distributions are obtained by summing over individual random variables in their joint mass function. So for  $1 \leq j < k$ ,

$$\begin{aligned} f(x_1, \dots, x_j) &= \Pr[X_1 = x_1, \dots, X_j = x_j] \\ &= \sum_{x_{j+1}} \cdots \sum_{x_k} f(x_1, \dots, x_k). \end{aligned}$$

The summations are over the range of the random variables being summed.

Conditional probability mass functions are obtained using the laws of probability so that for  $1 \leq j < k$  we write

$$\begin{aligned} f(x_1, \dots, x_j \mid x_{j+1}, \dots, x_k) &= \Pr[X_1 = x_1, \dots, X_j = x_j \mid X_{j+1} = x_{j+1}, \dots, X_k = x_k] \\ &= \Pr[X_1 = x_1, \dots, X_k = x_k] / \Pr[X_{j+1} = x_{j+1}, \dots, X_k = x_k] \\ &= f(x_1, \dots, x_k) / f(x_{j+1}, \dots, x_k). \end{aligned}$$

Multivariate moments are taken with respect to the joint mass function. Specifically, the expectation of the function  $g(X_1, \dots, X_k)$  is

$$E[g(X_1, \dots, X_k)] = \sum_{x_1} \cdots \sum_{x_k} g(x_1, \dots, x_k) f(x_1, \dots, x_k),$$

provided this summation converges.

Multivariate moments can also be defined. The *covariance* between random variables  $X_1$  and  $X_2$  is

$$\begin{aligned} \text{Cov}[X_1, X_2] &= E[(X_1 - E[X_1])(X_2 - E[X_2])] \\ &= E[X_1 X_2] - E[X_1]E[X_2] \end{aligned}$$

and their *correlation* is

$$\text{Corr}[X_1, X_2] = \text{Cov}[X_1, X_2] / \{\text{Var}[X_1] \text{Var}[X_2]\}^{1/2}.$$

The correlation is always between  $-1$  and  $1$ . The correlation is a measure of the strength of the linear relationship between two random variables. If the random variables are independent then their correlation is zero, but zero correlation does not imply independence. There are also higher-order analogies to the correlation coefficient that extend to more than two random variables at a time. These are introduced and used in Section 7.3.

Another example of multivariate moments are the *joint factorial moments*. Specifically, for nonnegative integers  $r_1, \dots, r_k$  the joint factorial moments of  $X_1, \dots, X_k$  are

$$E[X_1^{(r_1)} \cdots X_k^{(r_k)}] = \sum_{x_1} \cdots \sum_{x_k} \left[ \prod_i x_i^{(r_i)} \right] f(x_1, \dots, x_k).$$

Conditional moments are taken with respect to the conditional distribution. In particular, we have the *iterated expectation*

$$E[X_1] = E\{E[X_1 | X_2]\}, \quad (1.4)$$

where the outer expectation is taken with respect to  $X_2$ .

Similarly,

$$\text{Var}[X_1] = \text{Var}\{E[X_1 | X_2]\} + E\{\text{Var}[X_1 | X_2]\}. \quad (1.5)$$

The sections that follow illustrate these definitions with reference to specific discrete distributions.

### 1.3 Binomial Distribution

The binomial distribution is the sum of  $n$  independent, identically distributed Bernoulli  $p$  random variables. The valid parameter values are  $0 \leq p \leq 1$  and

$n = 1, 2, \dots$ . The parameter  $n$  is often referred to as the *index* or the *sample size* of the binomial distribution.

The probability that the sum of  $n$  independent and identically distributed Bernoulli random variables is equal to  $y$  is

$$\Pr[Y = y] = \binom{n}{y} p^y (1 - p)^{n-y}, \quad (1.6)$$

defined for  $y = 0, 1, \dots, n$  and zero otherwise. Bernoulli random variables are defined in (1.1).

We can reverse the roles of success and failure. Specifically if  $Y$  behaves as binomial with parameters  $n$  and  $p$ , then  $n - Y$  behaves as binomial with parameters  $n$  and  $1 - p$ .

Figure 1.1 illustrates the binomial distribution for  $n = 10$  and parameter  $p = 0.2, 0.5$ , and  $0.8$ . For  $p = 0.5$ , the binomial distribution is symmetric. The distribution has a longer left or right tail, depending on whether  $p > 0.5$  or  $p < 0.5$ , respectively.

The proof that the probabilities in (1.6) sum to one is provided by the expansion of the binomial polynomial

$$[p + (1 - p)]^n = 1.$$

The terms in this expansion are the individual binomial probabilities in (1.6). This relationship has led to the name of the binomial distribution.

The mean of  $Y$  in the binomial distribution (1.6) is  $E[Y] = np$  and the variance is  $\text{Var}[Y] = np(1 - p)$ . The variance is smaller than the mean.

The third central moment

$$E[(Y - np)^3] = np(1 - p)(1 - 2p)$$

is positive for  $0 < p < 1/2$ , negative for  $1/2 < p < 1$ , and zero when  $p = 1/2$ . This symmetry and asymmetry can be seen in Fig. 1.1.

The moment generating function of the binomial distribution is

$$M(t) = E[e^{tY}] = (1 - p + pe^t)^n.$$

The factorial moment generating function is

$$G(t) = E[(1 + t)^Y] = (1 + pt)^n.$$

For  $r = 1, 2, \dots$  the factorial moments of the binomial distribution are

$$E[Y^{(r)}] = n^{(r)} p^r.$$

The sum of two independent binomial random variables with respective parameters  $(n_1, p_1)$  and  $(n_2, p_2)$  behaves as binomial  $(n_1 + n_2, p_1)$  when  $p_1 = p_2$ . If the parameters  $p_1$  and  $p_2$  are not equal, then the sum of the two independent binomial random variables does not behave as binomial.

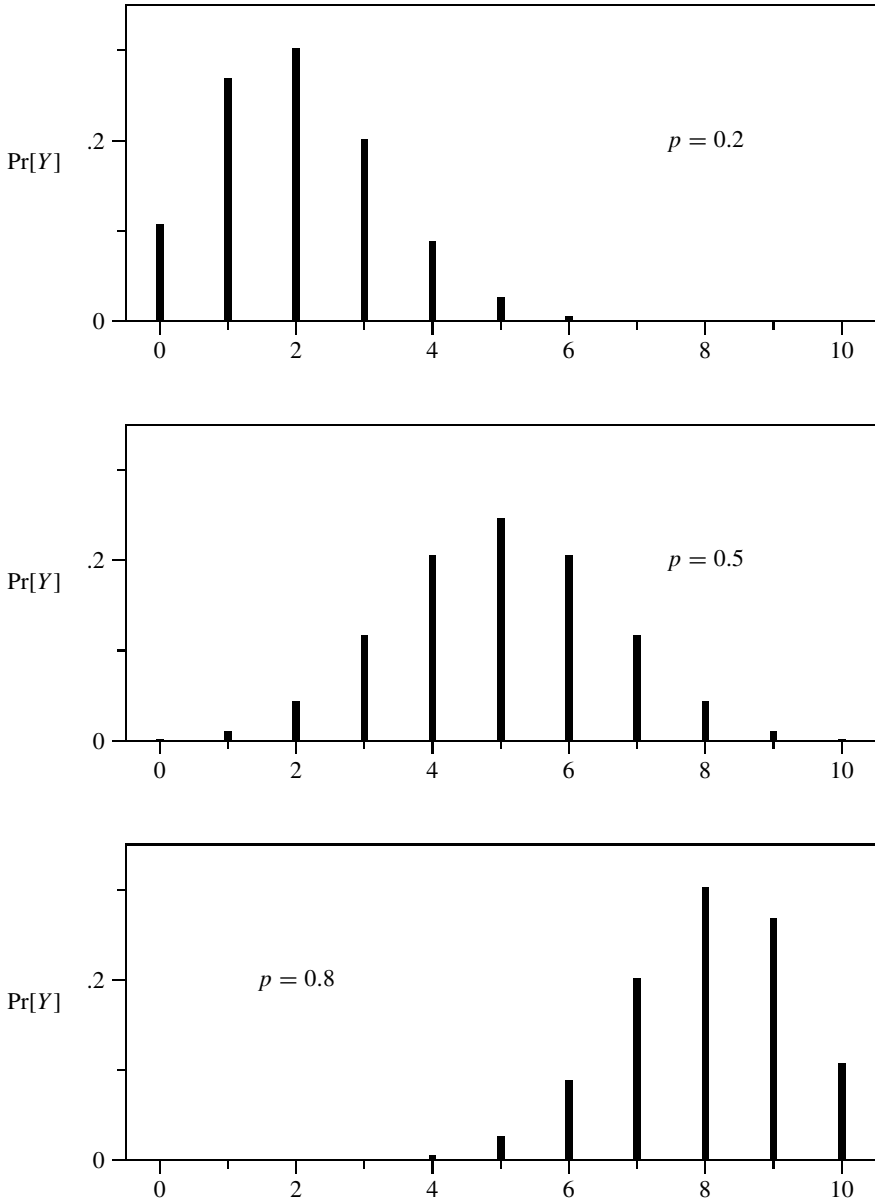


Figure 1.1 The binomial distribution probability mass function is illustrated for  $n = 10$  with parameters  $p = 0.2, 0.5,$  and  $0.8$ .

If  $n$  is large and  $p$  is not close to either 0 or 1 so that the variance is also large, then the binomial can be approximated by the normal distribution. If  $n$  is large and  $p$  is very close to zero, then the binomial can be approximated by the Poisson distribution. Similarly, if  $p$  is very close to 1 then  $n - Y$  will behave approximately as Poisson. The Poisson distribution and these approximations are discussed in Section 1.5.

Chapter 7 describes several generalizations of the binomial distribution when we drop the assumption of independence among the individual Bernoulli indicators.

In another generalization, if the binomial  $p$  parameter varies according to the beta distribution, then the marginal distribution of  $Y$  is beta-binomial. Specifically, suppose that  $Y$  conditional on  $p$  behaves as binomial  $(n, p)$  and that  $p$  behaves as a beta random variable with density function

$$f_{\alpha\beta}(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}$$

for  $0 \leq p \leq 1$  and parameters  $\alpha > 0$  and  $\beta > 0$ .

The marginal mass function of  $Y$  is then

$$\begin{aligned} \Pr[Y = y] &= \int_0^1 \binom{n}{y} p^y (1-p)^{n-y} f_{\alpha\beta}(p) \, dp \\ &= \frac{n! \Gamma(\alpha + \beta) \Gamma(\alpha + y) \Gamma(\beta + n - y)}{y! (n - y)! \Gamma(\alpha) \Gamma(\beta) \Gamma(\alpha + \beta + n)}. \end{aligned} \quad (1.7)$$

This is the mass function of the beta-binomial distribution and is derived in another fashion in Section 7.2.4. More details about this distribution are available in Johnson, Kotz, and Kemp (1992, pp 239–42).

The gamma function  $\Gamma(\cdot)$  is a generalization of the definition of the factorial function extended to nonintegers. The gamma function and its approximations are described in Section 1.8.

## 1.4 The Multinomial Distribution

The multinomial distribution is the generalization of the binomial to more than two mutually exclusive categories. This multivariate distribution is used to describe the joint frequencies of several possible outcomes.

Consider  $n$  mutually independent experiments, each of which results in one of  $k$  ( $k = 2, 3, \dots$ ) possible mutually exclusive and exhaustive outcomes. The probability of each of these outcomes is  $p_1, \dots, p_k$ , where  $p_i \geq 0$  and  $\sum p_i = 1$ . After  $n$  independent replications of this experiment, let  $\{N_1, \dots, N_k\}$  denote the set of joint frequencies of each of the  $k$  outcomes. Each of these frequencies can take on the values  $0, 1, \dots, n$  and satisfy

$$\sum_i^n N_i = n. \quad (1.8)$$

The  $k$  frequencies  $\{N_1, \dots, N_k\}$  represent a  $k - 1$  dimensional random variable as a result of this constraint.

The joint probability mass function of  $\{N_1, \dots, N_k\}$  is

$$\Pr[ N_1 = n_1, \dots, N_k = n_k ] = \frac{n!}{n_1! \dots n_k!} \prod_i^k p_i^{n_i} \tag{1.9}$$

for nonnegative integers  $\{n_i\}$  that satisfy  $\sum n_i = n$ .

All of the probabilities in (1.9) represent the complete expansion of the polynomial

$$(p_1 + p_2 + \dots + p_k)^n.$$

This polynomial is equal to 1 because  $\sum p_i = 1$ , demonstrating that (1.9) is a valid mass function and sums to unity.

The marginal distribution of each  $N_i$  is binomial with parameters  $n$  and  $p_i$ . The conditional distribution of  $N_1$ , given  $N_2 = n_2$  is binomial with parameters  $n - n_2$  and  $p_1/(1 - p_2)$ . The joint marginal distribution of  $\{N_1, N_2, n - N_1 - N_2\}$  is multinomial ( $k = 3$ ), with parameters  $n$  and  $\{p_1, p_2, 1 - p_1 - p_2\}$ .

The joint factorial moments of  $\{N_1, \dots, N_k\}$  are

$$E \left[ \prod_i N_i^{(r_i)} \right] = n^{(r_+)} \prod_i p_i^{r_i}$$

for  $r_i = 0, 1, \dots$  and  $r_+ = \sum r_i$ .

In particular, the covariance of  $N_1$  and  $N_2$  is

$$\text{Cov} [ N_1, N_2 ] = E [ (N_1 - np_1)(N_2 - np_2) ] = -np_1p_2.$$

This covariance is negative, reflecting the constraint (1.8) that the sum of the  $N_i$  is fixed.

The correlation between a pair of frequencies

$$\text{Corr} [ N_1, N_2 ] = - \left[ \frac{p_1}{1 - p_1} \frac{p_2}{1 - p_2} \right]^{1/2}$$

is related to the product of the two odds for the frequencies.

There are two approximations to the multinomial distribution when the sample size  $n$  is large. Suppose  $n$  is large and  $(p_1, p_2)$  become small at a rate such that the limits  $\lambda_1 = np_1$  and  $\lambda_2 = np_2$  are bounded above and away from zero. Then  $(N_1, N_2)$  will jointly behave approximately as independent Poisson random variables with parameters  $(\lambda_1, \lambda_2)$ .

A second approximation occurs when  $n$  is large and the variances of  $(N_1, N_2)$  are both large. Then the joint behavior of  $(N_1, N_2)$  is approximately bivariate normal.

## 1.5 Poisson Distribution

The Poisson distribution is often described as an approximation to the binomial distribution. It is useful in settings where the binomial  $n$  parameter is large and  $p$  is small. These parameters approach limits in such a way that the mean, denoted by  $\lambda = np$ , remains moderate and bounded away from zero. Under these conditions, the binomial distribution can be approximated by the Poisson distribution.

This approximation is based on the sum of a large number of independent and identically distributed Bernoulli random variables. The Poisson distribution is also the approximate behavior of a large number of independent Bernoulli indicators that are not identically distributed but all of whose  $p$  parameters are uniformly small.

The probability mass function for the Poisson distribution with mean parameter  $\lambda > 0$  is

$$\Pr[Y = y] = e^{-\lambda} \lambda^y / y! \quad (1.10)$$

defined for  $y = 0, 1, \dots$  and zero otherwise.

The mean and variance of the Poisson distribution are both equal to  $\lambda$ . The third central moment of the Poisson distribution

$$E[(Y - \lambda)^3] = \lambda$$

is also equal to  $\lambda$ .

Higher Poisson central moments do not continue this pattern. For example, the fourth central moment is

$$E[(Y - \lambda)^4] = 3\lambda^2 + \lambda.$$

For  $r = 1, 2, \dots$ , the factorial moments of the Poisson distribution are

$$E[Y^{(r)}] = \lambda^r \quad (1.11)$$

and the factorial moment generating function is

$$E[(t + 1)^Y] = e^{t\lambda}.$$

The moment generating function is

$$M_Y(t) = E[e^{tY}] = \exp[\lambda(e^t - 1)].$$

When the mean parameter  $\lambda$  is large, then the Poisson distribution can be approximated by the normal distribution. This normal approximate distribution becomes more apparent in Fig. 1.2 as  $\lambda$  grows larger.

The sum of two independent Poisson random variables will also behave as Poisson. Specifically, if  $Y_1$  and  $Y_2$  are independently distributed as Poisson with parameters  $\lambda_1$  and  $\lambda_2$ , respectively, then  $Y_1 + Y_2$  will follow the Poisson distribution with parameter  $\lambda_1 + \lambda_2$ .

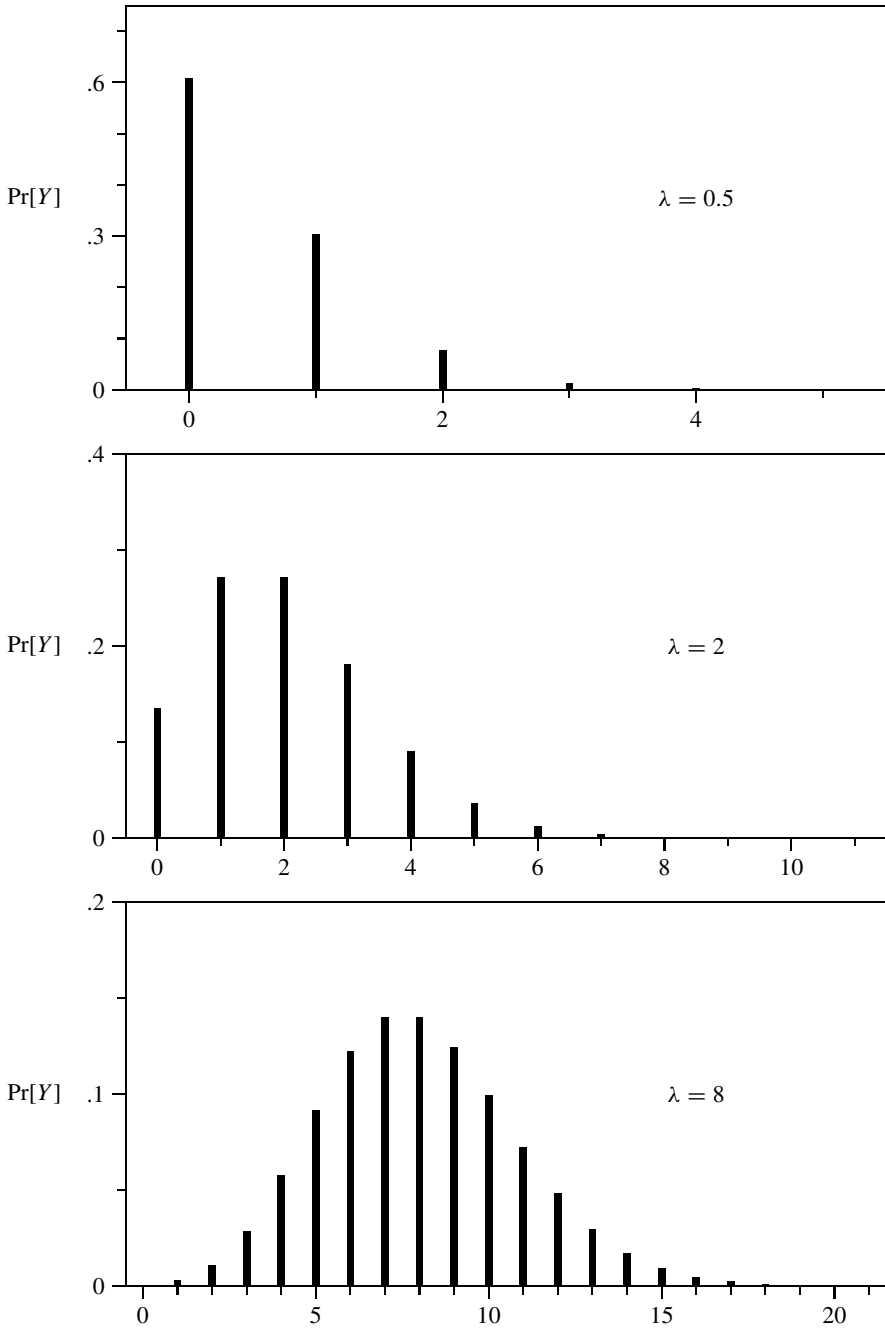


Figure 1.2 The Poisson distribution illustrated for  $\lambda = 0.5, 2, \text{ and } 8$ .

Similarly, the distribution of  $Y_1$  conditional on the value of  $Y_1 + Y_2$  is binomial with parameters

$$n = Y_1 + Y_2$$

and

$$p = \lambda_1 / (\lambda_1 + \lambda_2).$$

Finally, there is a close relationship between the Poisson and the negative binomial distribution that is described in the following section.

As a numerical example of the Poisson distribution, consider the data in Table 1.1 reported by Student (1907) on the distribution of yeast cells using a hemocytometer. WS Gosset was employed by the Guinness brewery and published under the *nom de plume* Student. The hemocytometer is a glass slide with a tiny  $20 \times 20$  grid etched into its surface. It was commonly used to count dilute samples of blood cells, or in this example, yeast cells under the microscope.

The distribution of a sample of 272 yeast cells across the 400 compartments on the glass slide is given in the second column of Table 1.1. Specifically, in 213 of the 400 compartments there were no yeast cells, in 128 compartments there was exactly one yeast cell, and so on. The expected counts and the corresponding  $\chi^2$  values are also given.

The argument we use to justify the Poisson distribution is based on its approximation to the binomial model. In terms of the binomial model there are a large number of  $n = 400$  compartments but there are not so many yeast cells relative to this number. We expect approximately

$$272/400 = 0.680$$

yeast cells per compartment. There is a very small probability that any one compartment is occupied by a given yeast cell. In particular, more than half of the 400 compartments are unoccupied. Kolchin *et al.* (1978, Ch. 2) contains a large

Table 1.1 Observed yeast cell frequencies and fitted Poisson distribution.

Count	Observed frequency	Fitted Poisson frequency	Components of $\chi^2$
0	213	202.342	.561
1	128	137.897	.710
2	37	46.989	2.123
3	18	10.674	} 6.667
4+	4	2.098	
Totals	400	400	10.061

Source: Student (1907)

number of limiting distributions of occupancy problems such as described by this example.

The maximum likelihood estimate  $\hat{\lambda} = 0.6815$  is obtained by numerically maximizing the likelihood function

$$\begin{aligned} \Lambda(\lambda) = & 213 \log \Pr[Y = 0 \mid \lambda] + 128 \log \Pr[Y = 1 \mid \lambda] \\ & + 37 \log \Pr[Y = 2 \mid \lambda] + 18 \log \Pr[Y = 3 \mid \lambda] \\ & + 4 \log \Pr[Y \geq 4 \mid \lambda] \end{aligned} \quad (1.12)$$

This estimate of 0.6815 is slightly larger than  $272/400=0.680$  because of how we treat the category listed as containing four or more yeast cells.

The expected counts using the fitted value of  $\hat{\lambda} = 0.6815$  are given in Table 1.1. The last two categories of three and four or more yeast cells are combined in order to increase the small expected counts in these compartments when calculating the  $\chi^2$  statistic. The  $\chi^2 = 10.06$  with 2 df has a significance level of .007 indicating a poor fit of the Poisson model.

The large values of the components of the  $\chi^2$  statistic show that the lack of fit is due to those compartments with large numbers of yeast cells. This is probably because the yeast cells cling together resulting in compartments with unusually large numbers of cells. The Bernoulli events of cells occupying compartments on the slide are then not independent, so a binomial distribution might be in question. Similarly, the Poisson approximation to this binomial distribution might not be appropriate either. This example is examined again in Section 1.6.

## 1.6 Negative Binomial Distribution

Suppose we continue to sample Bernoulli distributed events with probability parameter  $p$  until we obtain  $c$  successes or 1's. The value of  $c = 1, 2, \dots$  is determined before the sampling begins. The sampling process ends with the observation of the  $c$ -th success. The negative binomial distribution describes the number of failures (0's) observed before the  $c$ th success has been achieved.

The number of failures  $Y$  until the  $c$ th success follows the negative binomial distribution with mass function

$$\Pr[Y = y] = \binom{c + y - 1}{c - 1} p^c (1 - p)^y. \quad (1.13)$$

for  $y = 0, 1, \dots$

The name 'negative binomial' and the proof that the mass function (1.13) sums to one comes from the following expansion. For  $P$  near zero, we have

$$(Q - P)^{-c} = Q^{-c} + \frac{c}{1!} P Q^{-c-1} + \frac{c(c+1)}{2!} P^2 Q^{-c-2} + \dots$$

Then set  $Q = P + 1$  and write

$$Q^{-c} = [(Q - P)/Q]^c = (1 - P/Q)^c$$

to show

$$(Q - P)^{-c} = (1 - P/Q)^c + \frac{c}{1!}(P/Q)(1 - P/Q)^c \\ + \frac{c(c+1)}{2!}(P/Q)^2(1 - P/Q)^c + \dots$$

These terms are the same as those in (1.13), where

$$P = (1 - p)/p.$$

The expected value of the negative binomial random variable with mass function in (1.13) is

$$E[Y] = \mu = c(1 - p)/p \quad (1.14)$$

and the variance is

$$\text{Var}[Y] = c(1 - p)/p^2.$$

The variance of the negative binomial distribution is always larger than its mean.

The probability generating function is

$$G(t) = E[t^Y] = [(1 - p)/(1 - pt)]^c.$$

The factorial moment generating function is

$$G(1 + t) = E[(1 + t)^Y] = [1 - (1 - p)t/p]^{-c}$$

and the factorial moments of the negative binomial distribution are

$$E[Y^{(k)}] = (c + k - 1)^{(k)} [(1 - p)/p]^k$$

for  $k = 0, 1, \dots$

If the  $c$  parameter is large and  $p$  approaches one in such a manner that the mean  $c(1 - p)/p$  approaches a finite, nonzero limit  $\lambda$ , then the negative binomial distribution can be approximated by the Poisson with mean  $\lambda$ .

The negative binomial distribution is plotted in Fig. 1.3 for various values of parameters  $c$  and  $p$ . From top to bottom, this figure displays the negative binomial and Poisson mass functions with means 0.5, 2, and 8. The mass points of these discrete distributions are connected with lines in order to compare the different distributions. This will be the convention in the remainder of the book.

In Fig. 1.3, the values of the  $c$  parameters are set to 1, 2, 4, and 8. Values of the  $p$  parameters vary in order to achieve a constant value of the expected value  $\mu$ . Specifically, from (1.14) we have

$$p = c/(c + \mu) \quad (1.15)$$

for mean  $\mu > 0$  and parameter  $c$ .

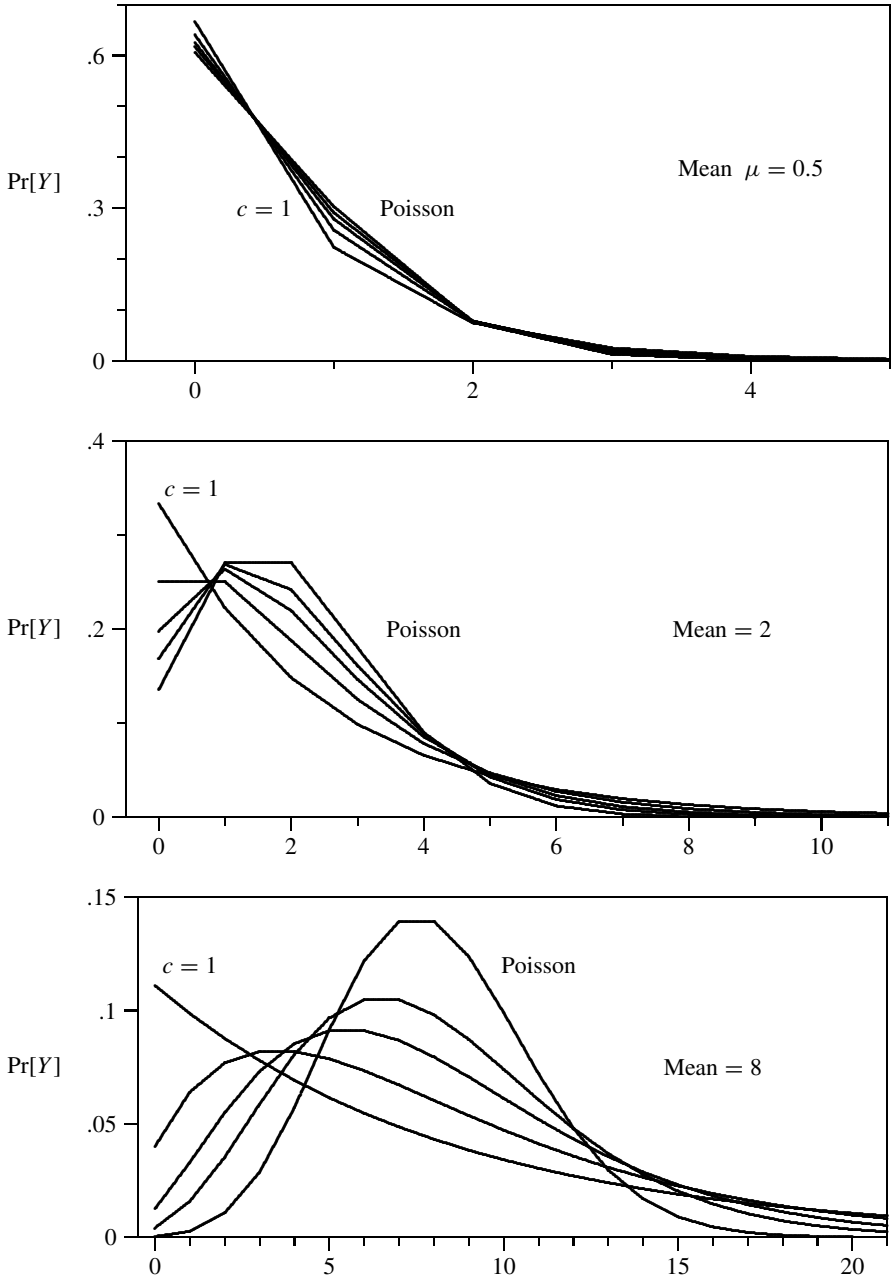


Figure 1.3 Negative binomial and Poisson mass functions with means  $\mu = 0.5, 2,$  and  $8$ . The negative binomial  $c$  parameters in every figure are  $1, 2, 4,$  and  $8$ . The Poisson distribution is the limit for large values of  $c$ .

As the  $c$  parameter gets larger, in every case in Fig. 1.3, the negative binomial distribution looks more like the Poisson distribution with the same mean. When the mean and  $c$  are both large, then the negative binomial distribution can be approximated by the normal distribution, as seen in the bottom of this figure. The normal approximation for the negative binomial distribution is also described in Section 2.3.1.

More generally, the  $c$  parameter need not be an integer. The following derivation of the negative binomial mass function demonstrates its description as a mixture of Poisson distributions. Let  $Y$ , given  $\lambda$ , behave as Poisson with parameter  $\lambda$  and suppose  $\lambda$  behaves as a gamma random variable with density function

$$f_{\Gamma} = (c/\mu)^c \lambda^{c-1} \exp(-\lambda c/\mu) / \Gamma(c)$$

for  $\lambda > 0$ . This gamma distribution has mean parameter  $\mu > 0$  and variance  $\mu^2/c$  for shape parameter  $c > 0$ .

The marginal distribution of  $Y$  has mass function

$$\begin{aligned} \Pr[Y = y] &= \int_0^{\infty} e^{-\lambda} \lambda^y / y! f_{\Gamma}(\lambda) d\lambda \\ &= \frac{\Gamma(c+y)}{y! \Gamma(c)} \left( \frac{c}{\mu+c} \right)^c \left( \frac{\mu}{\mu+c} \right)^y, \end{aligned} \quad (1.16)$$

which is a form of the mass function of the negative binomial distribution.

Expressions (1.13) and (1.16) coincide when  $c$  is an integer and  $p = c/(\mu + c)$ . The derivation of the negative binomial distribution in (1.16) helps explain why it has longer tails than the corresponding Poisson distribution with the same mean, as seen in Fig. 1.3.

Let us use this derivation of the negative binomial distribution to motivate another examination of Student's yeast example given in Table 1.1. The negative binomial distribution is fitted to this data in Table 1.2. The maximum likelihood estimates are obtained by numerically maximizing a likelihood similar to the

Table 1.2 Observed yeast cell frequencies and fitted negative binomial distribution.

Count	Observed frequency	Fitted frequency	Components of $\chi^2$
0	213	214.330	.008
1	128	122.476	.249
2	37	45.004	1.423
3	18	13.477	1.518
4+	4	4.714	.108
Totals	400	400	3.306

function given at (1.12), namely

$$\begin{aligned}\Lambda(c, \mu) &= 213 \log \Pr[Y = 0 \mid c, \mu] + 128 \log \Pr[Y = 1 \mid c, \mu] \\ &\quad + 37 \log \Pr[Y = 2 \mid c, \mu] + 18 \log \Pr[Y = 3 \mid c, \mu] \\ &\quad + 4 \log \Pr[Y \geq 4 \mid c, \mu]\end{aligned}$$

with probability model  $\Pr[\cdot]$  given in (1.16).

These maximum likelihood estimates are  $\hat{c} = 3.496$  and  $\hat{\mu} = .6831$ . Notice that the estimate of the mean parameter  $\hat{\mu}$  is very close in value to the estimate  $\hat{\lambda} = .6815$  for the Poisson parameter that maximizes (1.12). Using the parameterization given in (1.13), we also have

$$\hat{p} = [\hat{c}/(\hat{c} + \hat{\mu})] = .8365.$$

The fitted expected counts and the components of the  $\chi^2$  statistic are given in Table 1.2. The  $\chi^2 = 3.306$  value with 2 df has significance level .19, indicating a very good fit to the data. The negative binomial model has longer tails than the Poisson, so this model is better able to model the large number of compartments on the glass slide containing several yeast cells clinging together. Not all cells exhibit this phenomenon but apparently some do. Similarly, the negative binomial as a mixture of Poisson distributions is able to model some of the dependence among the yeast cells.

The negative binomial distribution is generalized in Chapter 2, in which we describe the number of trials necessary in order to obtain both  $c$  successes and  $c$  failures. The resulting distribution is the larger of two negative binomial distributions: the number of trials necessary until  $c$  successes and the number of trials needed until  $c$  failures are observed.

## 1.7 Hypergeometric Distribution

The binomial, Poisson, and negative binomial distributions all assume that we are sampling from an infinitely large parent population. The hypergeometric distribution is the analogy for sampling from a finite population.

Consider an urn containing  $N$  balls. Of these, suppose  $m$  are of the ‘successful’ color and the remaining  $N - m$  are of the ‘unsuccessful’ type. We reach into the urn and draw out a sample of size  $n$  balls. The hypergeometric distribution describes the number of successful colored balls in our sample of size  $n$ . This sample can also be illustrated as a  $2 \times 2$  table of counts as given in Table 1.3.

The probability mass function of the hypergeometric distribution is

$$\Pr[Y = y] = \binom{m}{y} \binom{N - m}{n - y} / \binom{N}{n}. \quad (1.17)$$

Intuitively,  $y$  out of the  $m$  successful types are sampled and  $n - y$  out of the  $N - m$  unsuccessful types are sampled. The denominator considers all possible ways in which samples of size  $n$  can be drawn out of the total  $N$ .

Table 1.3 The hypergeometric distribution of  $Y$ , given in (1.17) displayed as a  $2 \times 2$  table.

	Types of Items		Totals
	Successful	Unsuccessful	
Items drawn	$Y$	$X = n - Y$	$n$
Items not drawn	$m - Y$	$N - n - m + Y$	$N - n$
Totals	$m$	$N - m$	$N$

The support or range of the hypergeometric distribution in (1.17) is

$$\max(0, m + n - N) \leq y \leq \min(n, m).$$

This range assures that all counts in Table 1.3 are nonnegative.

The derivation of this distribution can be obtained as a conditional distribution of a binomial. Specifically, let  $X$  and  $Y$  denote independent binomial random variables with index parameters  $N - m$  and  $m$  respectively with the same value of their  $p$  parameter.

We want the conditional distribution of  $Y$ , given  $X + Y = n$ . Begin by writing

$$\begin{aligned} \Pr[Y = y \mid X + Y = n] &= \Pr[Y = y \text{ and } X + Y = n] / \Pr[X + Y = n] \\ &= \Pr[Y = y; X = n - y] / \Pr[X + Y = n]. \end{aligned}$$

The distribution of  $X + Y$  in the denominator behaves as binomial with parameters  $N$  and  $p$ . We also use the independence of  $X$  and  $Y$  to show

$$\begin{aligned} \Pr[Y = y \mid X + Y = n] &= \Pr[Y = y] \Pr[X = n - y] / \Pr[X + Y = n] \\ &= \binom{m}{y} p^y (1 - p)^{m-y} \binom{N - m}{n - y} p^{n-y} (1 - p)^{N-n-m+y} \\ &\quad / \binom{N}{n} p^n (1 - p)^{N-n}. \end{aligned} \tag{1.18}$$

All of the terms involving  $p$  and  $1 - p$  cancel in the numerator and the denominator. The binomial coefficients that remain after this cancellation yield the mass function given in (1.17).

The expected value of this distribution is

$$E[Y] = mn/N.$$

This is also the estimate used for the expected counts when we calculate the  $\chi^2$  statistic.

The variance of the hypergeometric distribution is

$$\begin{aligned} \text{Var}[Y] &= mn(N - m)(N - n)/N^2(N - 1) \\ &= E[Y](N - m)(N - n)/N(N - 1). \end{aligned}$$

The variance is always smaller than the mean.

The more general factorial moments of this distribution are

$$E[Y^{(r)}] = m^{(r)}n^{(r)}/N^{(r)}$$

for  $r = 1, 2, \dots, \min(m, n)$  and zero for larger values of  $r$ .

The same method used to prove that the hypergeometric probabilities sum to one is used several times in this book. The easiest way to prove that the probabilities in (1.17) sum to one is to begin by writing the polynomial identity

$$(1 + z)^N = (1 + z)^m (1 + z)^{N-m}. \tag{1.19}$$

This equality is true for all values of  $z$ . The coefficients of each power of  $z$  must agree on both sides of this identity. If we identify the coefficient of  $z^n$  on both sides of this identity, then we have

$$\binom{N}{n} = \sum_y \binom{m}{y} \binom{N-m}{n-y}. \tag{1.20}$$

This is the same equality we need to show that the probabilities in (1.17) sum to one. This equality is known as *Vandermonde's theorem*. The polynomial identity in (1.19) is referred to as the *generating polynomial* of the hypergeometric distribution.

### 1.7.1 Negative hypergeometric distribution

This distribution is the finite sample analogy to the negative binomial distribution described in Section 1.6. The *negative hypergeometric distribution* (Johnson *et al.* 1992, pp. 239–42) is the distribution of the number of unsuccessful draws, one at a time, from an urn with two different colored balls until a specified number of successful draws have been obtained. If  $m$  out of  $N$  balls are of the successful type, then the number of unsuccessful draws  $Y$  observed before  $c$  of the successful types are obtained is

$$\Pr[Y = y] = \binom{c + y - 1}{c - 1} \binom{N - c - y}{m - c} / \binom{N}{m}, \tag{1.21}$$

with parameters satisfying  $1 \leq c \leq m < N$  and range  $y = 0, 1, \dots, N - m$ . The expected value of  $Y$  in (1.21) is  $mc/(N - m - 1)$ .

This distribution is used by Kaigh and Lachenbruch (1982) in applications of resampling for nonparametric quantile estimation. This distribution is a special case of the beta-binomial when all of its parameters are positive integers. The beta-binomial distribution is described in Section 7.2.4.

## 1.7.2 Extended hypergeometric distribution

This distribution is the extension of (1.18) when the binomial  $p$  parameters are not the same for  $X$  and  $Y$ . Let  $X$  and  $Y$  denote independent binomial random variables with parameters  $(N - m, p_1)$  and  $(m, p_2)$  respectively. We want to describe the conditional distribution of  $Y$  given  $X + Y = n$ . This distribution is discussed in detail by Johnson, Kotz, and Kemp (1992, pp. 279–82).

Following the derivation in (1.18), write

$$\Pr[Y = y | X + Y = n] = \Pr[Y = y] \Pr[X = n - y] / \Pr[X + Y = n].$$

This denominator does not have a simple expression because  $X + Y$  does not have a binomial distribution for  $p_1 \neq p_2$ . The denominator  $\Pr[X + Y = n]$  does not depend on  $y$ .

We then have

$$\begin{aligned} \Pr[Y = y | X + Y = n] & \propto \binom{m}{y} p_2^y (1 - p_2)^{m-y} \binom{N - m}{n - y} p_1^{n-y} (1 - p_1)^{N - n - m + y} \\ & \propto \binom{m}{y} \binom{N - m}{n - y} \exp(\lambda y), \end{aligned} \tag{1.22}$$

where  $\lambda$  is the log-odds ratio

$$\lambda = \log[p_2(1 - p_1)/p_1(1 - p_2)].$$

The normalizing constant of proportionality in (1.22) depends on  $N$ ,  $n$ ,  $m$ , and  $\lambda$  but not  $y$ . Similarly, the moments and generating functions of this distribution do not have simple closed form expressions, in general. The special case of (1.22) is the hypergeometric mass function given in (1.17), when  $\lambda = 0$  or equivalently when  $p_1 = p_2$ .

An example of the extended hypergeometric mass function is plotted in Fig. 1.4 for  $m = 12$ ;  $n = 10$  and  $N = 25$ . Values of  $\lambda$  vary from  $-3$  to  $3$  in this figure, shifting the mass from the lower to the upper range of the distribution.

The most common application of this distribution is as a model for the alternative hypothesis in the analysis of  $2 \times 2$  tables. The hypergeometric distribution with mass function given in (1.17) is derived by assuming  $p_1 = p_2$  or independence of population and sampling parameter. This is the basis of the Fisher exact

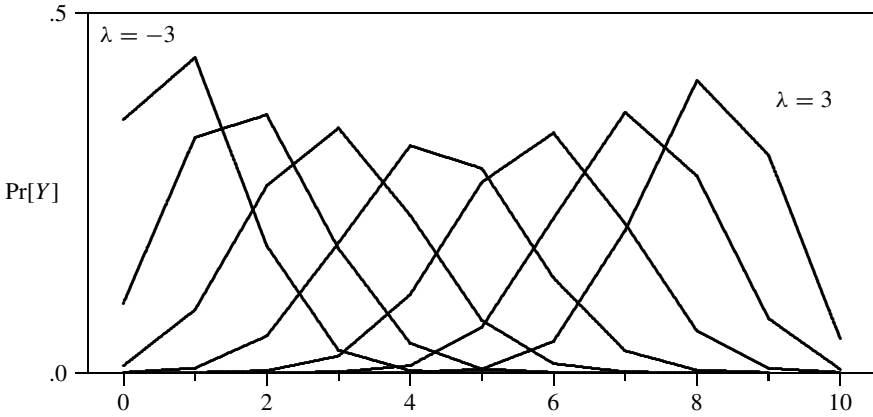


Figure 1.4 Extended hypergeometric mass function for  $m = 12$ ,  $n = 10$ , and  $N = 25$ . Values of  $\lambda$  are from  $-3$  to  $3$  by  $1$ .

test of significance when examining  $2 \times 2$  tables. As an alternative to the independence hypothesis, the extended hypergeometric distribution is a natural choice and provides a single parameter  $\lambda$  that allows expression of an alternative hypothesis.

As an example of the hypergeometric distribution in use, let us consider the data from Innes *et al.* (1969) summarized in Table 1.4. In this experiment, 16 mice were exposed to the fungicide Avadex and an additional 79 mice were kept separately under unexposed conditions. After 85 weeks, all of the 95 mice were sacrificed and their lungs were examined for tumors by pathologists. The frequencies of tumors in the exposed and unexposed mice is summarized in Table 1.4.

The  $\chi^2$  statistic for this table is 5.41 with 1 df and significance level 0.02, indicating that there is some evidence of association between exposure and the tumor outcome. The assumption made using this statistic is that the counts in the table can be approximated using the normal distribution, and the significance level of the  $\chi^2$  statistic is approximated using the corresponding asymptotic distribution.

*Exact methods* for analyzing data of this type require a complete enumeration of all possible outcomes consistent with the margins of the table. The observed

Table 1.4 Incidence of tumors in mice exposed to the fungicide Avadex (Innes *et al.*, 1969).

	Exposed	Control	Totals
Mice with tumors	4	5	9
No tumors	12	74	86
Totals	16	79	95

count of  $Y = 4$  exposed mice with tumors could also have taken on any of the possible values  $0, 1, \dots, 9$  for these sets of marginal totals.

These marginal totals are all considered to be fixed in the *frequentist* statistical analysis of the interaction between exposure and cancer development. The number of mice exposed or not was determined by the experimenters and has no random component. The number of mice developing tumors is an outcome that could be different if the same experiment were repeated today. Nevertheless, we also treat both sets of marginal totals as fixed in the examination of the interaction between exposure and tumor development.

The frequentist approach to the analysis of this data posits that the marginal totals contain no information about the interaction of the row and column categories. The Bayesian approach, on the other hand, might argue that the investigators knew that 16 exposed mice would be sufficient to detect a difference and hence also knew something about the interaction *a priori*.

The hypergeometric distribution in (1.17) is the model of independence of exposure status and the eventual tumor development. Four exposed mice developed tumors. Under this model of independence we expect

$$9 \times 16/95 = 1.516$$

exposed mice to develop tumors.

The observed number (=4) is larger than this expected value. The probability of four exposed mice with tumors is not the significance level. Instead, we need to examine the probability of this outcome plus all of those outcomes that are more extreme than the one observed. The exact probability of observing four or more exposed mice with tumors is

$$\Pr[X \geq 4 \mid N = 95, m = 16, n = 9] = .0411, \quad (1.23)$$

using distribution (1.17), modeling the null hypothesis of independence of fungicide exposure and tumor development.

This *exact significance level* does not rely on making any approximations or using any asymptotic assumptions. While exact methods do have many virtues, it is well recognized that they usually suffer from diminished power over other methods. In Section 6.1, there is additional discussion of the settings where exact methods are useful.

One method of increasing the power of the exact test of significance is to apply a *continuity correction*. In calculating  $\Pr[Y \geq 4]$  in the current example, the value of this expression is dominated by the first term, namely  $\Pr[Y = 4]$ . To improve the power, we sometimes apply a correction that diminishes the effect of this first term and calculate

$$\frac{1}{2} \Pr[Y = 4] + \Pr[Y > 4] = .0236 \quad (1.24)$$

and report this value as the exact significance level. This continuity corrected significance level is very close to the corresponding value obtained by the  $\chi^2$  statistic for this data.

The examination of Table 1.4 using both exact and asymptotic  $\chi^2$  methods provides evidence that tumor incidence is increased with exposure to the fungicide in terms of their statistical significance levels that are rather small. We can also measure the effect using the extended hypergeometric distribution given in (1.22).

The empirical log-odds ratio for this table is

$$\log[ (4 \times 74)/(5 \times 12) ] = 1.596.$$

The maximum likelihood estimate  $\hat{\lambda}$  of  $\lambda$  is 1.572, which is close to the empirical ratio. The maximum likelihood estimate is the value of  $\lambda$  that maximizes the probability

$$\Pr[ X = 4 \mid \lambda, N = 95, m = 16, n = 9 ]$$

of the observed data using the extended hypergeometric distribution (1.22).

An exact 95% confidence interval for  $\lambda$  is obtained by solving the equations

$$\Pr[ X \geq 4 \mid \lambda_1, N = 95, m = 16, n = 9 ] = .025$$

and

$$\Pr[ X \leq 4 \mid \lambda_2, N = 95, m = 16, n = 9 ] = .025$$

for  $(\lambda_1, \lambda_2)$  using distribution (1.22). The resulting 95% confidence interval for  $\lambda$  is the solution of these two equations, namely  $(-.1814, 3.264)$ .

This exact 95% confidence interval for  $\lambda$  contains zero, corresponding to the null hypothesis of independence of exposure and tumor development. This confidence interval is two sided. The exact significance level of  $\lambda$  given in (1.23) and (1.24) is less than 0.05 but these also represent one-sided tests. In contrast, we note that the  $\chi^2$  test is two-sided because it squares the difference between the observed and expected counts in its numerator.

## 1.8 Stirling's Approximation

This section contains a number of mathematical results that will be used in subsequent chapters.

The *gamma function* is defined by

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

for  $x > 0$ .

The gamma function satisfies the recurrence relation

$$\Gamma(x + 1) = x \Gamma(x).$$

When the argument  $x$  is a positive integer, then the gamma function is equal to the factorial

$$\Gamma(x + 1) = x!.$$

Stirling's approximation to the gamma function for large values of the argument is given by

$$\Gamma(x) = e^{-x} x^{x-1/2} (2\pi)^{1/2} \left[ 1 + 1/12x + 1/288x^2 + \dots \right]. \quad (1.25)$$

Other useful forms of Stirling's formula include

$$x! = (2\pi)^{1/2} x^{x+1/2} \exp[-x + \theta/12x] \quad (1.26)$$

for  $0 < \theta(x) < 1$  and

$$\log \Gamma(x) = -x + (x - 1/2) \log(x) + 1/2 \log(2\pi) + 1/12x - 1/360x^3 + \dots \quad (1.27)$$

These relations and many other useful approximations and properties of the gamma function are given in (Abramowitz and Stegun (1972, Ch. 6)).