Chapter 4 Important discrete distributions

Bibliografie: Riley et al. (2006), Mihoc and Micu (1980)

Having discussed some general properties of distributions, we now consider the more important discrete distributions encountered in physical applications. These are discussed in detail below, and summarised in table 4.1.

Table 4.1. Important discrete distributions

Distribution	Probability law $f(x)$	MGF	$E[X]$	$V[X]$
Binomial	$C_n^x p^x q^{n-x}$	$\left(pe^t+q\right)^n$	np	npq
Negative binomial	$C_{r+x-1}^x p^r q^x$	$\left(\frac{p}{1-qe^t}\right)^r$	$\frac{rq}{p}$	$\frac{rq}{p^2}$
Geometric	$q^{x-1}p$	$\frac{pe^t}{1-qe^t}$	$\frac{1}{p}$	$\frac{q}{p^2}$
Hypergeometric	$\frac{(Np)!(Nq)!n!(N-n)!}{x!(Np-x)!(n-x)!(Nq-n+x)!N!}$		np	$\frac{N-n}{N-1}npq$
Poisson	$\frac{\lambda^x}{x!}e^{-\lambda}$	$e^{\lambda\left(e^t-1\right)}$	λ	λ

4.1 The binomial distribution

Perhaps the most important discrete probability distribution is the *binomial distribution*. This distribution describes processes that consist of a number of independent identical *trials* with two possible outcomes, A and $B=\bar{A}$. We may call these outcomes 'success' and 'failure' respectively. If the probability of a success is $\Pr(A)=p$ then the probability of a failure is $\Pr(B)=q=1-p$. If we perform n trials then the discrete random variable

$$X = \text{ number of times } \text{A occurs} \tag{4.1}$$

can take the values $0,1,2,\ldots,n$; its distribution amongst these values is described by the *binomial distribution*.

We now calculate the probability that in *n* trials we obtain *x* successes (and so $n-x$ failures). One way of obtaining such a result is to have *x* successes followed by $n-x$ failures. Since the trials are assumed independent, the probability of this is

$$\underbrace{pp\cdots p}_{x\text{ times}}\times\underbrace{qq\cdots q}_{n-x\text{ times}}=p^x q^{n-x}$$

This is, however, just one permutation of *x* successes and $n-x$ failures. The total number of permutations of *n* objects, of which *x* are identical and of type *1* and $n-x$ are identical and of type *2*, is given by

$$\frac{n!}{x!(n-x)!}=C_n^x$$

Therefore, the total probability of obtaining *x* successes from *n* trials is

$$f(x)=\Pr(X=x)=C_n^x p^x q^{n-x}=C_n^x p^x (1-p)^{n-x} \tag{4.2}$$

which is the *binomial probability distribution formula*. When a random variable *X* follows the binomial distribution for *n* trials, with a probability of success *p*, we write $X\sim Bin(n,p)$. Some typical binomial distributions are shown in figure 4.1.

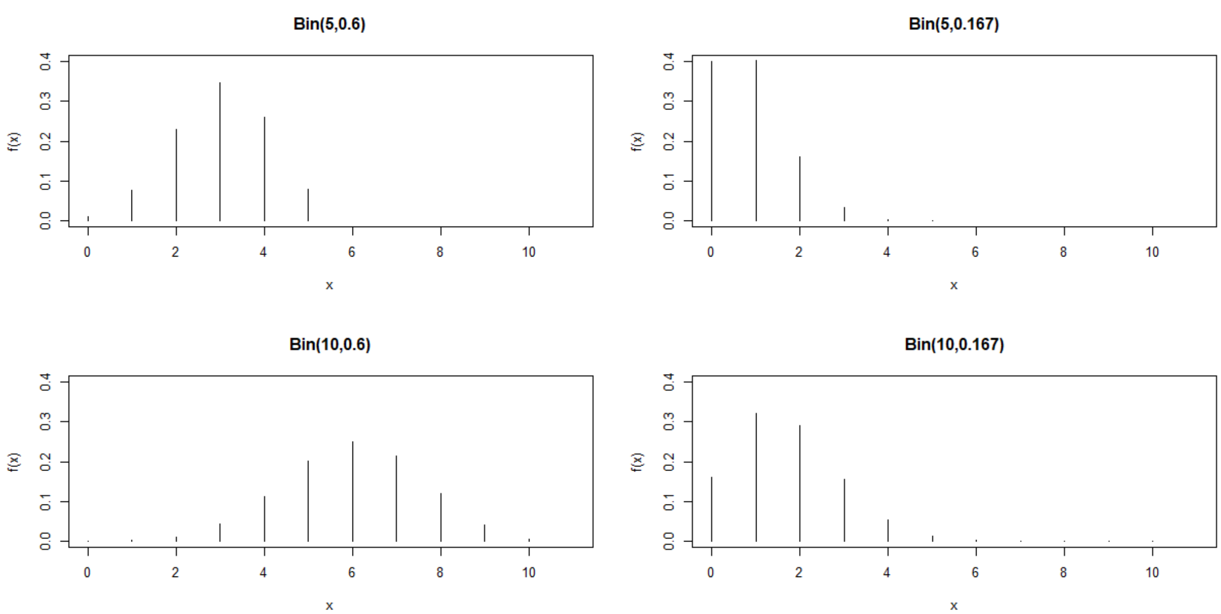

Figure 4.1 Some typical binomial distributions with various combinations of parameters *n* and *p*.

Exercise 1. If a single six-sided die is rolled five times, what is the probability that a six is thrown exactly three times?

Here the number of 'trials' n = 5, and we are interested in the random variable

$$X = \text{number of sixes thrown}$$

Since the probability of a 'success' is $p = 1/6$, the probability of obtaining exactly three sixes in five throws is given by (4.2) as

$$\Pr(X = 3) = C_5^3 p^3 q^{5-3} = \frac{5!}{3!(5-3)!}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = 0.032$$

For evaluating binomial probabilities a useful result is the binomial recurrence formula

$$\Pr(X = x+1) = C_n^{x+1} p^{x+1} q^{n-x-1} = \frac{n!}{(x+1)!(n-x-1)!} p^{x+1} q^{n-x-1}$$

$$= \frac{n!}{x!(n-x)!}\frac{n-x}{x+1} p^x p q^{n-x} q^{-1} = \frac{p}{q}\frac{n-x}{x+1}\frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$\Pr(X = x+1) = \frac{p}{q}\frac{n-x}{x+1}\Pr(X = x) \qquad (4.3)$$

which enables successive probabilities $\Pr(X = x+k)$, $k = 1, 2, \dots$, to be calculated once $\Pr(X = x)$ is known; it is often quicker to use than (4.2).

Exercise 2. The random variable X is distributed as $X \sim Bin\left(3, \frac{1}{2}\right)$. Evaluate the probability function $f(x)$ using the binomial recurrence formula.

The probability $\Pr(X = 0)$ may be calculated using (4.2) and is

$$\Pr(X = 0) = C_3^0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

The ratio $p/q = \frac{1}{2}/\frac{1}{2} = 1$ in this case and so, using the binomial recurrence formula (4.3), we find

$$\Pr(X = 1) = 1 \times \frac{3-0}{0+1} \times \frac{1}{8} = \frac{3}{8}$$

$$\Pr(X = 2) = 1 \times \frac{3-1}{1+1} \times \frac{3}{8} = \frac{3}{8}$$

$$\Pr(X = 3) = 1 \times \frac{3-2}{2+1} \times \frac{3}{8} = \frac{1}{8}$$

results which may be verified by direct application of (4.2).

As required, the binomial distribution satisfies

$$\sum_{x=0}^{n} f(x) = \sum_{x=0}^{n} C_n^x p^x q^{n-x} = (p+q)^n = 1$$

From the definitions of $E[X]$ and $V[X]$ for a discrete distribution, we may show that for the binomial distribution $E[X] = np$ and $V[X] = npq$. The direct summations involved are, however, rather cumbersome and these results are obtained much more simply using the moment generating function.

The moment generating function for the binomial distribution

To find the MGF for the binomial distribution we consider the binomial random variable *X* to be the sum of the random variables X_i, $i = 1, 2, \ldots, n$, which are defined by

$$X_i = \begin{cases} 1 & \text{if a 'success' occurs on the } i\text{th trial} \\ 0 & \text{if a 'failure' occurs on the } i\text{th trial} \end{cases} \tag{4.4}$$

X_i are known as *Bernoulli* random variables. Thus,

$$M_i(t) = E\left[e^{tX_i}\right] = e^{0t} \times \Pr(X_i = 0) + e^{1t} \times \Pr(X_i = 1)$$

$$= 1 \times q + e^t \times p = pe^t + q$$

If X_1, X_2, … , X_n are independent random variables then MGF for the sum $S_n = X_1 + X_2 + \ldots + X_n$ is

$$M_{S_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) \tag{4.5}$$

It follows that the MGF for the binomial distribution is given by

$$M(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(pe^t + q\right)^n \tag{4.6}$$

We can now use the moment generating function to derive the mean and variance of the binomial distribution.

$$M'(t) = n\left(pe^t + q\right)^{n-1} pe^t$$

$$E[X] = M'(0) = np(p+q)^{n-1} = np \tag{4.7}$$

The last equality follows from $p + q = 1$.

$$M''(t) = npe^t(n-1)\left(pe^t + q\right)^{n-2} pe^t + npe^t\left(pe^t + q\right)^{n-1}$$

$$E\left[X^2\right] = M''(0) = np^2(n-1) + np$$

Thus,

$$V[X] = E\left[X^2\right] - \left(E[X]\right)^2 = np^2(n-1) + np - (np)^2 = np(1-p)$$

$$V[X] = npq \tag{4.8}$$

Multiple binomial distributions

Suppose X and Y are two independent random variables, both of which are described by binomial distributions with a common probability of success p, but with different numbers of trials n_1 and n_2, so that $X \sim Bin(n_1, p)$ and $Y \sim Bin(n_2, p)$. Now consider the random variable $Z = X + Y$. We could calculate the probability distribution of Z directly using (3.41), but it is much easier to use the MGF (4.6).

Since X and Y are independent random variables, the MGF $M_Z(t)$ of the new variable $Z = X + Y$ is given simply by the product of the individual MGFs $M_X(t)$ and $M_Y(t)$. Thus,

$$M_Z(t) = M_X(t)M_Y(t) = \left(pe^t + q\right)^{n_1}\left(pe^t + q\right)^{n_2}$$

$$M_Z(t) = \left(pe^t + q\right)^{n_1+n_2} \qquad (4.9)$$

which we recognize as the MGF of $Z \sim Bin(n_1 + n_2, p)$. Hence Z is also described by a binomial distribution.

This result may be extended to any number of binomial distributions. If X_i, $i = 1, 2, \ldots, N$, is distributed as $X_i \sim Bin(n_i, p)$ then $Z = X_1 + X_2 + \ldots + X_N$ is distributed as $Z \sim Bin(n_1 + n_2 + \ldots + n_N, p)$, as would be expected since the result of $\sum_i n_i$ trials cannot depend on how they are split up.

4.2 The geometric and negative binomial distributions

A special case of the binomial distribution occurs when instead of the number of successes we consider the discrete random variable

$$X = \text{ number of trials required to obtain the first success} \qquad (4.10)$$

The probability that *x* trials are required in order to obtain the first success, is simply the probability of obtaining $x-1$ failures followed by one success. If the probability of a success on each trial is *p*, then for $x > 0$

$$f(x) = \Pr(X = x) = (1-p)^{x-1}\,p = q^{x-1}p \qquad (4.11)$$

where $q = 1 - p$. This distribution is called the *geometric distribution* (for waiting time problems). The *probability generating function* for this distribution is:

$$\Phi_X(t) = \sum_{n=0}^{\infty} f_n\, t^n = \sum_{n=1}^{\infty}\left(q^{n-1}p\right)\, t^n$$

$$= \frac{p}{q}\sum_{n=1}^{\infty}(qt)^n = \frac{p}{q}\frac{qt}{1-qt} = \frac{pt}{1-qt} \qquad (4.12)$$

By replacing t by e^t in (4.12) we immediately obtain the MGF of the geometric distribution

$$M(t) = \frac{pe^t}{1-qe^t} \qquad (4.13)$$

from which its mean and variance are found to be

$$M'(t) = \frac{pe^t(1-qe^t) - pe^t(-qe^t)}{(1-qe^t)^2} = \frac{pe^t}{(1-qe^t)^2}$$

$$E[X] = M'(0) = \frac{p}{(1-q)^2} = \frac{1}{p} \quad \Rightarrow \quad E[X] = \frac{1}{p} \qquad (4.14)$$

$$M''(t) = \frac{pe^t(1-qe^t)^2 - pe^t 2(1-qe^t)(-qe^t)}{(1-qe^t)^4}$$

$$M''(t) = \frac{pe^t}{(1-qe^t)^4}\left(1 - 2qe^t + q^2e^{2t} + 2qe^t - 2q^2e^{2t}\right)$$

$$M''(t) = \frac{pe^t}{(1-qe^t)^4}\left(1 - q^2e^{2t}\right)$$

$$E\left[X^2\right] = M''(0) = \frac{p}{(1-q)^4}(1-q^2) = \frac{p(1+q)}{(1-q)^3} = \frac{1+q}{p^2}$$

$$V[X] = E\left[X^2\right] - (E[X])^2 = \frac{1+q}{p^2} - \frac{1}{p^2} = \frac{q}{p^2} \quad \Rightarrow \quad V[X] = \frac{q}{p^2} \qquad (4.15)$$

Exercise 3: Wafer Contamination. The probability that a wafer contains a large particle of contamination is 0.01. If it is assumed that the wafers are independent, what is the probability that exactly 125 wafers need to be analyzed before a large particle is detected?

Let *X* denote the number of samples analyzed until a large particle is detected. Then *X* is a geometric random variable with $p = 0.01$. The requested probability is

$$P(X = 125) = (0.99)^{124}\, 0.01 = 0.0029$$

Another distribution closely related to the binomial is the *negative binomial distribution*. This describes the probability distribution of the random variable

$$X = \text{ number of failures before the } r\text{th success.} \qquad (4.16)$$

One way of obtaining x failures before the r-th success is to have $r-1$ successes followed by x failures followed by the r-th success, for which the probability is

$$\underbrace{pp\cdots p}_{r-1\text{ times}}\times\underbrace{qq\cdots q}_{x\text{ times}}\times p = p^r q^x \qquad (4.17)$$

However, the first $r+x-1$ factors constitute just one permutation of $r-1$ successes and x failures. The total number of permutations of these $r+x-1$ objects, of which $r-1$ are identical and of type *1* and x are identical and of type *2*, is

$$\frac{(r+x-1)!}{(r-1)!x!} = C_{r+x-1}^{x}$$

Therefore, the total probability of obtaining x failures before the rth success is

$$f(x) = \Pr(X = x) = C_{r+x-1}^{x} p^r q^x \qquad (4.18)$$

which is called the *negative binomial distribution*.

The PGF of this distribution is:

$$\Phi_X(t) = \sum_{n=0}^{\infty} f_n\, t^n = \sum_{n=0}^{\infty}\left(C_{r+n-1}^{n} p^r q^n\right) t^n = p^r \sum_{n=0}^{\infty}\left(C_{r+n-1}^{n} q^n\right) t^n \qquad (4.19)$$

For negative value of n in the binomial expansion, we have:

$$(x+y)^{-m} = x^{-m}\sum_{k=0}^{\infty} C_{-m}^{k}\left(\frac{y}{x}\right)^k \quad \text{where } m>0 \text{ and } |x|>|y| \qquad (4.20)$$

and $C_{-m}^{k} = (-1)^k C_{m+k-1}^{k}$

So, $$\Phi_X(t) = p^r \sum_{n=1}^{\infty}(-1)^n \mathrm{C}_{-r}^{n}\; q^n t^n = p^r (1-qt)^{-r}$$

$$\Phi_X(t) = \left(\frac{p}{1-qt}\right)^r \qquad (4.21)$$

By replacing t by e^t, we immediately obtain the MGF of the negative binomial distribution

$$M(t) = \left(\frac{p}{1-qe^t}\right)^r \qquad (4.22)$$

and that its mean and variance are given by

$$M'(t) = p^r(-r)\left(1-qe^t\right)^{-r-1}\left(-qe^t\right) = p^r rqe^t\left(1-qe^t\right)^{-r-1}$$

$$E[X] = M'(0) = p^r rq(1-q)^{-r-1} = \frac{rq}{p} \qquad (4.23)$$

$$M''(t) = p^r rqe^t\left(1-qe^t\right)^{-r-1} + p^r rqe^t(-r-1)\left(1-qe^t\right)^{-r-2}\left(-qe^t\right)$$

$$E\left[X^2\right] = M''(0) = p^r rq(1-q)^{-r-1} + p^r rq(-r-1)(1-q)^{-r-2}(-q)$$

$$E\left[X^2\right] = p^r rq\left(\frac{1}{p^{r+1}} + q\frac{r+1}{p^{r+2}}\right) = rq\left(\frac{1}{p} + q\frac{r+1}{p^2}\right) = \frac{rq}{p^2}(p+qr+q) = \frac{rq}{p^2}(1+qr)$$

$$V[X] = E\left[X^2\right] - \left(E[X]\right)^2 = \frac{rq}{p^2}(1+qr) - \frac{r^2q^2}{p^2} = \frac{rq}{p^2}$$

$$E[X] = \frac{rq}{p} \quad \text{and} \quad V[X] = \frac{rq}{p^2} \qquad (4.24)$$

4.3 The hypergeometric distribution

In subsection 4.1 we saw that the probability of obtaining x successes in n independent trials was given by the *binomial distribution*. Suppose that these n 'trials' actually consist of drawing at random n balls, from a set of N such balls of which M are red and the rest white. Let us consider the random variable

X = number of red balls drawn. (4.25)

On the one hand, if the balls are drawn with *replacement* then the trials are <u>independent</u> and the probability of drawing a red ball is $p = M/N$ each time. Therefore, the probability of drawing x red balls in n trials is given by the binomial distribution as

$$\Pr(X = x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \qquad (4.26)$$

On the other hand, if the balls are drawn *without replacement* the trials are <u>not independent</u> and the probability of drawing a red ball depends on how many red balls have already been drawn. We can derive a general formula for the probability of drawing x red balls in n trials, as follows.

The number of ways of drawing *x* red balls from *M* is C_M^x, and the number of ways of drawing $n-x$ white balls from $N-M$ is C_{N-M}^{n-x}. Therefore, the total number of ways to obtain *x* red balls in *n* trials is $C_M^x C_{N-M}^{n-x}$. However, the total number of ways of drawing *n* objects from *N* is simply C_N^n. Hence the probability of obtaining *x* red balls in *n* trials is

$$\Pr(X=x)=\frac{C_M^x C_{N-M}^{n-x}}{C_N^n} \tag{4.27}$$

$$\Pr(X=x)=\frac{M!}{x!(M-x)!}\frac{(N-M)!}{(n-x)!(N-M-n+x)!}\frac{n!(N-n)!}{N!} \tag{4.28}$$

$$p\overset{not}{=}\frac{M}{N} \qquad q=1-p$$

$$=\frac{(Np)!}{x!(Np-x)!}\frac{\left(N\left(1-\frac{M}{N}\right)\right)!}{(n-x)!\left(N\left(1-\frac{M}{N}\right)-n+x\right)!}\frac{n!(N-n)!}{N!}$$

$$\Pr(X=x)=\frac{(Np)!(Nq)!n!(N-n)!}{x!(Np-x)!(n-x)!(Nq-n+x)!N!} \tag{4.29}$$

This is called the *hypergeometric distribution*.

It may be shown that the hypergeometric distribution has mean

$$E[X]=n\frac{M}{N}=np \tag{4.30}$$

and variance

$$V[X]=\frac{nM(N-M)(N-n)}{N^2(N-1)}=\frac{N-n}{N-1}npq \tag{4.31}$$

The argumentation is based on the identity

$$\sum_{x=0}^{n}C_M^x C_{N-M}^{n-x}=C_N^n \tag{4.32}$$

resulting from the identification of the coefficients of y^n in

$$(1+y)^M (1+y)^{N-M} = (1+y)^N$$

$$E[X] = \sum_{x=0}^{n} x \frac{C_M^x C_{N-M}^{n-x}}{C_N^n} = \frac{M}{C_N^n} \sum_{x=0}^{n} C_{M-1}^{x-1} C_{N-M}^{n-x} = \frac{M}{C_N^n} C_{N-1}^{n-1} = \frac{M}{\frac{N}{n} C_{N-1}^{n-1}} C_{N-1}^{n-1} = \frac{Mn}{N} = np$$

$$\left(C_M^x = \frac{M!}{x!(M-x)!} = \frac{M(M-1)!}{x(x-1)!(M-x)!} = \frac{M}{x} C_{M-1}^{x-1} \right)$$

$$E\left[X^2\right] = \sum_{x=0}^{n} x^2 \frac{C_M^x C_{N-M}^{n-x}}{C_N^n} = \sum_{x=0}^{n} x(x-1) \frac{C_M^x C_{N-M}^{n-x}}{C_N^n} + \sum_{x=0}^{n} x \frac{C_M^x C_{N-M}^{n-x}}{C_N^n}$$

$$= \frac{M}{C_N^n} \sum_{x=0}^{n} (x-1) C_{M-1}^{x-1} C_{N-M}^{n-x} + E[X] = \frac{M(M-1)}{C_N^n} \sum_{x=0}^{n} C_{M-2}^{x-2} C_{N-M}^{n-x} + E[X]$$

$$E\left[X^2\right] = \frac{M(M-1)}{C_N^n} C_{N-2}^{n-2} + n\frac{M}{N}$$

$$V[X] = E\left[X^2\right] - \left(E[X]\right)^2 = \frac{M(M-1)}{C_N^n} C_{N-2}^{n-2} + n\frac{M}{N} - \left(n\frac{M}{N}\right)^2$$

Also $\quad C_N^n = \dfrac{N!}{n!(N-n)!} = \dfrac{N(N-1)(N-2)!}{n(n-1)(n-2)!(N-n)!} = \dfrac{N(N-1)}{n(n-1)} C_{N-2}^{n-2}$

$$V[X] = \frac{M(M-1)}{C_N^n} \frac{n(n-1)}{N(N-1)} C_N^n + n\frac{M}{N} - \left(n\frac{M}{N}\right)^2$$

$$V[X] = \frac{M(M-1)n(n-1)N + nMN(N-1) - n^2M^2(N-1)}{N^2(N-1)}$$

$$V[X] = \frac{nM(N-M)(N-n)}{N^2(N-1)} = \frac{N-n}{N-1} npq$$

Exercise 4. In the UK National Lottery each participant chooses six different numbers between 1 and 49. In each weekly draw six numbered winning balls are subsequently drawn. Find the probabilities that a participant chooses 0, 1, 2, 3, 4, 5, 6 winning numbers correctly.

The probabilities are given by a hypergeometric distribution with $N = 49$ (the total number of balls), $M = 6$ (the number of winning balls drawn), and $n = 6$ (the number of numbers chosen by each participant). Thus, substituting in (4.27), we find

$$\Pr(0) = \frac{C_6^0 C_{43}^6}{C_{49}^6} = \frac{1}{2.29} \qquad \Pr(1) = \frac{C_6^1 C_{43}^5}{C_{49}^6} = \frac{1}{2.42}$$

$$\Pr(2) = \frac{C_6^2 C_{43}^4}{C_{49}^6} = \frac{1}{7.55} \qquad \Pr(3) = \frac{C_6^3 C_{43}^3}{C_{49}^6} = \frac{1}{56.6}$$

$$\Pr(4) = \frac{C_6^4 C_{43}^2}{C_{49}^6} = \frac{1}{1032} \qquad \Pr(5) = \frac{C_6^5 C_{43}^1}{C_{49}^6} = \frac{1}{54200}$$

$$\Pr(6) = \frac{C_6^6 C_{43}^0}{C_{49}^6} = \frac{1}{13.98 \times 10^6}$$

It can easily be seen that

$$\sum_{i=1}^{6} \Pr(i) = 0.44 + 0.41 + 0.13 + 0.02 + O\left(10^{-3}\right) = 1$$

Note that if the number of trials (balls drawn) is small compared with *N*, *M* and $N - M$ then not replacing the balls is of little consequence, and we may approximate the hypergeometric distribution by the binomial distribution (with $p = M / N$); this is much easier to evaluate.

4.4 The Poisson distribution

We have seen that the binomial distribution describes the number of successful outcomes in a certain number of trials *n*. The Poisson distribution also describes the probability of obtaining a given number of successes but for situations in which the number of ‘trials’ cannot be enumerated; rather it describes the situation in which *discrete events occur in a continuum*. Typical examples of discrete random variables *X* described by a Poisson distribution are the number of telephone calls received by a switchboard in a given interval, or the *number of stars* above a certain brightness in a particular area of the sky. Given a mean rate of occurrence λ of these events in the relevant interval or area, the Poisson distribution gives the probability $\Pr(X = x)$ that exactly *x* events will occur.

We may derive the form of the Poisson distribution as the limit of the binomial distribution when the number of trials $n \to \infty$ and the probability of 'success' $p \to 0$, in such a way that $np = \lambda$ remains finite. Thus, in our example of a telephone switchboard, suppose we wish to find the probability that exactly *x* calls are received during some time interval, given that the mean number of calls in such an interval is λ. Let us begin by dividing the time interval into a large number, *n*, of equal shorter intervals, in each of which the probability of receiving a call is *p*. As we let $n \to \infty$ then $p \to 0$, but since we require the mean number of calls in the interval to equal λ, we must have $np = \lambda$. The probability of *x* successes in *n* trials is given by the binomial formula as

$$\Pr\left(X = x\right) = \frac{n!}{x!\left(n-x\right)!} p^x \left(1-p\right)^{n-x} \qquad (4.33)$$

Now as $n \to \infty$, with *x* finite, the ratio of the *n*-dependent factorials in (4.33) behaves asymptotically as a power of *n*, i.e.

$$\lim_{n\to\infty} \frac{n!}{\left(n-x\right)!} = \lim_{n\to\infty} n\left(n-1\right)\left(n-2\right)\cdots\left(n-x+1\right) \sim n^x$$

$$\lim_{n\to\infty} \lim_{p\to 0} \left(1-p\right)^{n-x} = \lim_{p\to 0} \frac{\left(1-p\right)^{\lambda/p}}{\left(1-p\right)^x} = \frac{e^{-\lambda}}{1}$$

Thus, using $\lambda = np$, (4.33) tends to the *Poisson distribution*

$$f\left(x\right) = \Pr\left(X = x\right) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad (4.34)$$

which gives the probability of obtaining exactly *x* calls in the given time interval. As we shall show below, λ is the mean of the distribution. Events following a Poisson distribution are usually said to occur randomly in time.

If a discrete random variable is described by a Poisson distribution of mean λ then we write $X \sim Po\left(\lambda\right)$. As it must be, the sum of the probabilities is unity:

$$\sum_{x=0}^{\infty} \Pr\left(X = x\right) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1$$

From (4.34) we may also derive the Poisson recurrence formula,

$$\Pr(X = x+1) = \frac{\lambda^{x+1}}{(x+1)!}e^{-\lambda} = \frac{\lambda}{x+1}\frac{\lambda^x}{x!}e^{-\lambda} = \frac{\lambda}{x+1}\Pr(X = x) \quad \text{for } x = 0,1,2,\ldots \quad (4.35)$$

which enables successive probabilities to be calculated easily once one is known.

Exercise 5. A person receives on average one e-mail message per half-hour interval. Assuming that the e-mails are received randomly in time, find the probabilities that in any particular hour 0, 1, 2, 3, 4, 5 messages are received.

Let X = number of e-mails received per hour. Clearly the mean number of e-mails per hour is two, and so X follows a Poisson distribution with $\lambda = 2$, i.e.

$$\Pr(X = x) = \frac{2^x}{x!}e^{-2}$$

Thus $\Pr(X = 0) = e^{-2} = 0.135$

$\Pr(X = 1) = 2e^{-2} = 0.271$

$\Pr(X = 2) = 2^2 e^{-2} / 2! = 0.271$

$\Pr(X = 3) = 2^3 e^{-2} / 3! = 0.180$

$\Pr(X = 4) = 2^4 e^{-2} / 4! = 0.090$

$\Pr(X = 5) = 2^5 e^{-2} / 5! = 0.036$

Or using the recurrence formula (4.35): $\Pr(X = 1) = \frac{2}{0+1} \times 0.135 = 0.270$,

$\Pr(X = 2) = \frac{2}{1+1} \times 0.270 = 0.270$, $\Pr(X = 3) = \frac{2}{2+1} \times 0.270 = 0.180 \quad \ldots$

The above example illustrates the point that a Poisson distribution typically rises and then falls. It either has a maximum when x is equal to the integer part of λ or, if λ happens to be an integer, has equal maximal values at $x = \lambda - 1$ and $x = \lambda$. The Poisson distribution always has a long ‘tail’ towards higher values of X but the higher the value of the mean the more symmetric the distribution becomes. Typical Poisson distributions are shown in figure 4.2.

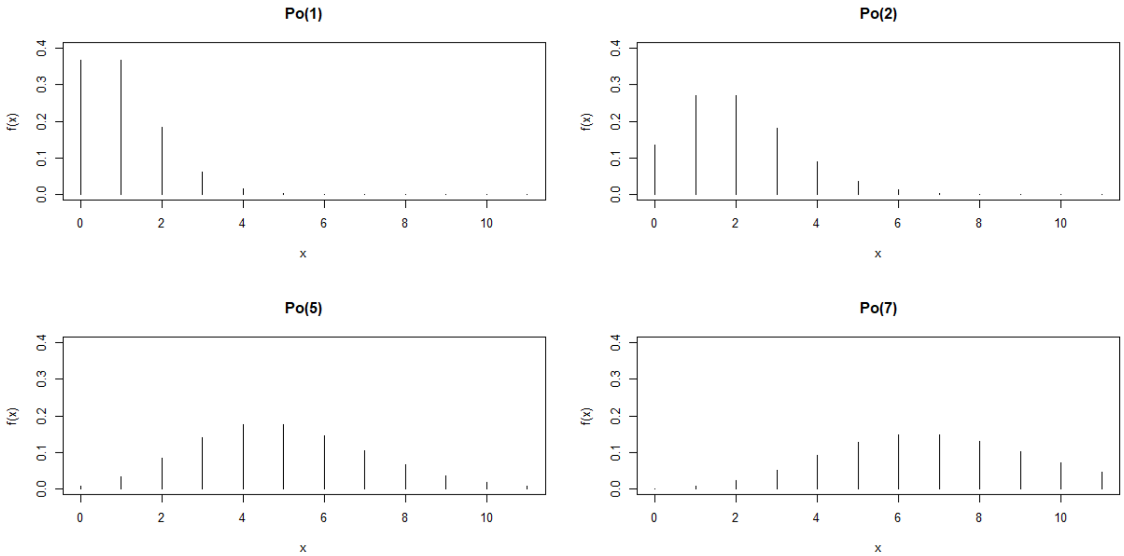

Figure 4.2 Poisson distributions for different values of the parameter λ.

Using the definitions of mean and variance, we may show that, for the Poisson distribution, $E[X]=\lambda$ and $V[X]=\lambda$. Nevertheless, as in the case of the binomial distribution, these results are much more easily proved using the MGF.

The moment generating function for the Poisson distribution

The MGF of the Poisson distribution is given by

$$M_X(t)=E\left[e^{tX}\right]=\sum_{x=0}^{\infty}e^{tx}\frac{e^{-\lambda}\lambda^x}{x!}=e^{-\lambda}\sum_{x=0}^{\infty}\frac{\left(\lambda e^t\right)^x}{x!}=e^{-\lambda}e^{\lambda e^t}$$

$$M_X(t)=e^{\lambda\left(e^t-1\right)} \tag{4.36}$$

from which we obtain

$$M_X'(t)=\lambda e^t e^{\lambda\left(e^t-1\right)}$$

$$M_X''(t)=\lambda e^t e^{\lambda\left(e^t-1\right)}+\left(\lambda e^t\right)^2 e^{\lambda\left(e^t-1\right)}$$

$$= \left(\lambda^2 e^{2t} + \lambda e^t\right) e^{\lambda\left(e^t - 1\right)}$$

Thus, the mean and variance of the Poisson distribution are given by

$$E\left[X\right] = M'_X\left(0\right) = \lambda \tag{4.37}$$

$$E\left[X^2\right] = M''_X\left(0\right) = \lambda + \lambda^2$$

$$V\left[X\right] = E\left[X^2\right] - \left(E\left[X\right]\right)^2 = \lambda + \lambda^2 - \lambda^2 = \lambda \tag{4.38}$$

The Poisson approximation to the binomial distribution

Earlier we derived the Poisson distribution as the limit of the binomial distribution when $n \to \infty$ and $p \to 0$ in such a way that $np = \lambda$ remains finite, where λ is the mean of the Poisson distribution. It is not surprising, therefore, that the Poisson distribution is a very good approximation to the binomial distribution for large n (≥ 50, say) and small p (≤ 0.1, say). Moreover, it is easier to calculate as it involves fewer factorials.

Exercise 6. In a large batch of light bulbs, the probability that a bulb is defective is 0.5%. For a sample of 200 bulbs taken at random, find the approximate probabilities that 0, 1 and 2 of the bulbs respectively are defective.

Let the random variable X = number of defective bulbs in a sample. This is distributed as $X \sim Bin\left(200, 0.005\right)$, implying that $\lambda = np = 1.0$. Since n is large and p small, we may approximate the distribution as $X \sim Po\left(1\right)$, giving

$$\Pr\left(X = x\right) = e^{-1}\frac{1^x}{x!}$$

from which $\Pr\left(X = 0\right) \approx e^{-1} \approx 0.37$, $\Pr\left(X = 1\right) \approx e^{-1} \approx 0.37$, $\Pr\left(X = 2\right) \approx e^{-1}/2 \approx 0.18$. For comparison, it may be noted that the exact values calculated from the binomial distribution are identical to those found here to two decimal places.

Multiple Poisson distributions

Let us suppose X and Y are two independent random variables, both of which are described by Poisson distributions with (in general) different means, so that $X \sim Po(\lambda_1)$ and $Y \sim Po(\lambda_2)$. Now consider the random variable $Z = X + Y$.

Since X and Y are independent RVs, the MGF for Z is simply the product of the individual MGFs for X and Y. Thus,

$$M_Z(t) = M_X(t) M_Y(t) = e^{\lambda_1(e^t-1)} e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}$$

which we recognise as the MGF of $Z \sim Po(\lambda_1 + \lambda_2)$. Hence Z is also Poisson distributed and has mean $\lambda_1 + \lambda_2$.

Exercise 7. Two types of e-mail arrive independently and at random: external e-mails at a mean rate of one every five minutes and internal e-mails at a rate of two every five minutes. Calculate the probability of receiving two or more e-mails in any two-minute interval.

Let X = number of external e-mails per two-minute interval,

Y = number of internal e-mails per two-minute interval.

Since we expect on average one external e-mail and two internal e-mails every five minutes we have $X \sim Po(0.4)$ and $Y \sim Po(0.8)$. Letting $Z = X + Y$ we have $Z \sim Po(0.4 + 0.8) = Po(1.2)$. Now

$$\Pr(Z \geq 2) = 1 - \Pr(Z < 2) = 1 - \Pr(Z = 0) - \Pr(Z = 1)$$

$$\Pr(Z = 0) = e^{-1.2} = 0.301$$

$$\Pr(Z = 1) = e^{-1.2}\frac{1.2}{1} = 0.361$$

Hence $\Pr(Z \geq 2) = 1 - \Pr(Z = 0) - \Pr(Z = 1) = 1 - 0.301 - 0.361 = 0.338$

The above result can be extended, to any number of Poisson processes, so that if $X_i \sim Po(\lambda_i)$, $i = 1, 2, \ldots, n$ then the random variable $Z = X_1 + X_2 + \ldots + X_n$ is distributed as $Z \sim Po(\lambda_1 + \lambda_2 + \ldots + \lambda_n)$.

Un site bun pentru reprezentarea functiilor de probabilitate The applet at http://mathlets.org/mathlets/probability-distributions/ gives a dynamic view of some discrete distributions.