CHAPTER

EXAMPLES OF DISCRETE PROBABILITY DISTRIBUTIONS

3

CHAPTER CONTENTS

In this chapter, popular discrete probability distributions and their properties such as the expectation, the variance, and the moment-generating functions are illustrated.

3.1 DISCRETE UNIFORM DISTRIBUTION

The *discrete uniform distribution* is the probability distribution for N events $\{1,\ldots,N\}$ that occur with equal probability. It is denoted by $U\{1,\ldots,N\}$, and its probability mass function is given by

$$f(x) = \frac{1}{N} \quad \text{for } x = 1,\ldots,N.$$

From the series formulas,

$$\sum_{x=1}^{N} x = \frac{N(N+1)}{2} \quad \text{and} \quad \sum_{x=1}^{N} x^2 = \frac{N(N+1)(2N+1)}{6},$$

the expectation and variance of $U\{1,\ldots,N\}$ can be computed as

$$E[x] = \frac{N+1}{2} \quad \text{and} \quad V[x] = \frac{N^2-1}{12}.$$

The probability mass function of $U\{1,\ldots,6\}$ is plotted in Fig. 2.2.

The probability mass function of the discrete uniform distribution $U\{a,a+1,\ldots,b\}$ for finite $a < b$ is given by

$$f(x) = \frac{1}{b-a+1} \quad \text{for } x = a,a+1,\ldots,b.$$

An Introduction to Statistical Machine Learning. DOI: 10.1016/B978-0-12-802121-7.00014-5

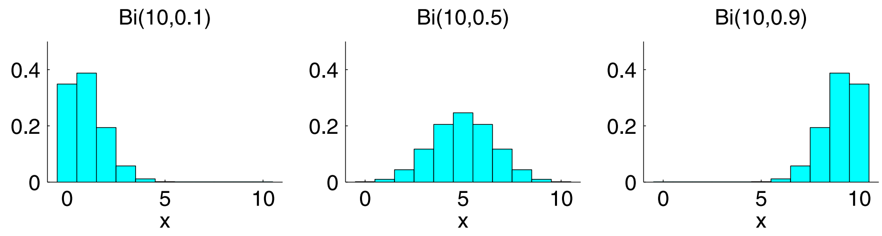


FIGURE 3.1

Probability mass functions of binomial distribution $\mathrm{Bi}(n, p)$.

Its expectation and variance are given by

$$E[x] = \frac{a+b}{2} \quad \text{and} \quad V[x] = \frac{(b-a+1)^2 - 1}{12}.$$

3.2 BINOMIAL DISTRIBUTION

Bernoulli trials are independent repeated trials of an experiment with two possible outcomes, say success and failure. Repeated independent tosses of the same coin are typical Bernoulli trials. Let p be the probability of success (getting heads in the coin toss) and q $(= 1 - p)$ be the probability of failure (getting tails in the coin toss). The *binomial distribution* is the probability distribution of the number x of successful trials in n Bernoulli trials and is denoted by $\mathrm{Bi}(n, p)$.

The probability of having x successful trials is given by p^x, while the probability of having $n - x$ unsuccessful trials is given by q^{n-x}. The number of combinations of x successful trials and $n - x$ unsuccessful trials is given by $\frac{n!}{x!(n-x)!}$, where

$$n! = n \times (n-1) \times \cdots \times 2 \times 1$$

denotes the *factorial*. $\frac{n!}{x!(n-x)!}$ is called the *binomial coefficient* and is denoted by $\binom{n}{x}$:

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

Putting them together based on the axioms of the probability provided in Section 2.2, the probability mass function of $\mathrm{Bi}(n, p)$ is given by

$$f(x) = p^x q^{n-x} \binom{n}{x} \quad \text{for } x = 0, 1, \ldots, N.$$

Probability mass functions of the binomial distribution for $n = 10$ are illustrated in Fig. 3.1.

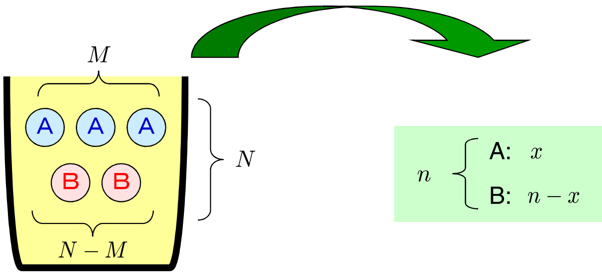


FIGURE 3.2

Sampling from a bag. The bag contains N balls which consist of $M < N$ balls labeled as "A" and $N - M$ balls labeled as "B." n balls are sampled from the bag, which consists of x balls labeled as "A" and $n - x$ balls labeled as "B."

The moment-generating function of $\mathrm{Bi}(n, p)$ can be obtained by using the *binomial theorem*,

$$(p+q)^n = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x},$$

as

$$\begin{aligned} M_x(t) &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x q^{n-x} \\ &= (pe^t + q)^n. \end{aligned}$$

From this, the expectation and variance of $\mathrm{Bi}(n, p)$ can be computed as

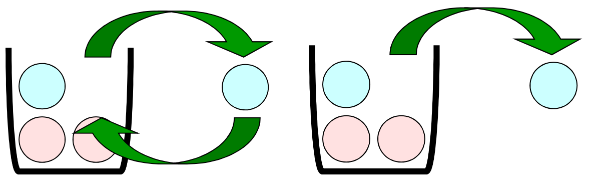
$$E[x] = np \quad \text{and} \quad V[x] = npq.$$

The expectation np would be intuitive because a Bernoulli trial with probability of success p is repeated n times. The variance npq is maximized when $p = 0.5$, while it is minimized when $p = 0$ or $p = 1$. This is also intuitive because it is difficult to predict the success or failure of a trial when the probability of success is 0.5.

The binomial distribution with $n = 1$, $\mathrm{Bi}(1, p)$, is specifically called the *Bernoulli distribution*.

3.3 HYPERGEOMETRIC DISTRIBUTION

Let us consider the situation illustrated in Fig. 3.2: n balls are sampled from the bag containing N balls, where M balls are labeled as "A" and $N - M$ balls are labeled



(a) Sampling with replacement (b) Sampling without replacement

FIGURE 3.3

Sampling with and without replacement. The sampled ball is returned to the bag before the next ball is sampled in sampling with replacement, while the next ball is sampled without returning the previously sampled ball in sampling without replacement.

as “B.” In this situation, there are two sampling schemes, as illustrated in Fig. 3.3: The first scheme is called *sampling with replacement*, which requires to return the sampled ball to the bag before the next ball is sampled. The other scheme is called *sampling without replacement*, where the next ball is sampled without returning the previously sampled ball to the bag.

In sampling with replacement, a ball is always sampled from the bag containing all N balls. This sampling process corresponds to the Bernoulli trials, and thus the probability distribution of obtaining x balls labeled as “A” when n balls are sampled with replacement is given by $\mathrm{Bi}(n, M/N)$.

On the other hand, in sampling without replacement, the number of balls contained in the bag is decreasing as the sampling process is progressed. Thus, the ratio of balls labeled as “A” and “B” in the bag depends on the history of sampling. The probability distribution of obtaining x balls labeled as “A” when n balls are sampled without replacement is called the *hypergeometric distribution* and denoted by $\mathrm{HG}(N, M, n)$.

The number of combinations of sampling x balls labeled as “A” from M balls is given by $\binom{M}{x}$, the number of combinations of sampling $n-x$ balls labeled as “B” from $N-M$ balls is given by $\binom{N-M}{n-x}$, and the number of combinations of sampling n balls from N balls is given by $\binom{N}{n}$. Putting them together, the probability mass function of $\mathrm{HG}(N, M, n)$ is given by

$$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad \text{for } x = 0, 1, \ldots, n.$$

Although the domain of x is $\{0, 1, \ldots, n\}$, the actual range of x is given by

$$\{\max(0, n-(N-M)), \ldots, \min(n, M)\},$$

because the number of balls in the bag is limited.

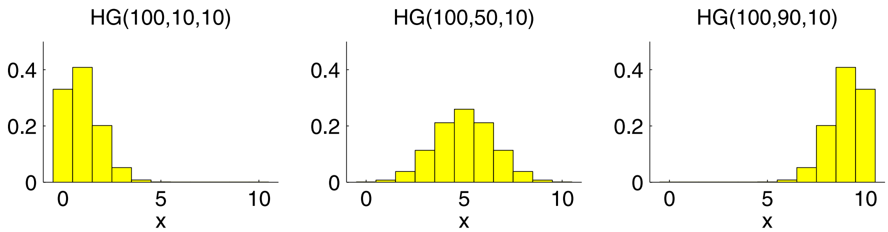


FIGURE 3.4

Probability mass functions of hypergeometric distribution $\mathrm{HG}(N, M, n)$.

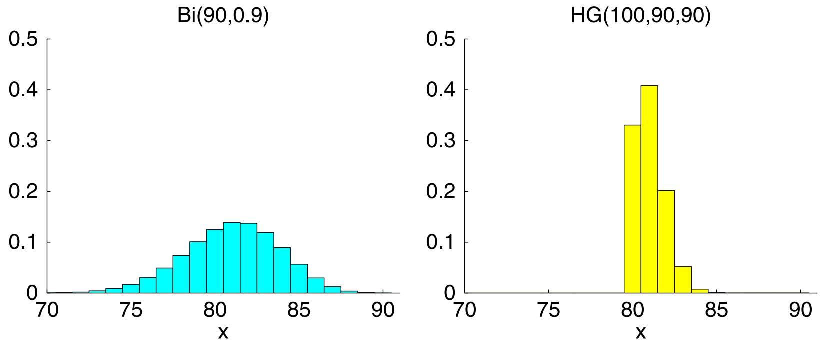


FIGURE 3.5

Probability mass functions of $\mathrm{Bi}(n, M/N)$ and $\mathrm{HG}(N, M, n)$ for $N = 100$, $M = 90$, and $n = 90$.

For $N = 100$ and $n = 10$, the probability mass functions of $\mathrm{HG}(N, M, n)$ and $\mathrm{Bi}(n, M/N)$ are plotted in Fig. 3.4 and Fig. 3.1, respectively. These graphs show that the probability mass functions are quite similar to each other, meaning that sampling with or without replacement does not affect that much when only $n = 10$ balls are sampled from $N = 100$ balls. Indeed, as N and M tend to infinity under n and the ratio N/M fixed, $\mathrm{HG}(N, M, n)$ is shown to agree with $\mathrm{Bi}(n, M/N)$.

Next, let us sample $n = 90$ from the bag of $N = 100$ balls where $M = 90$ balls are labeled as "A." In this situation, even though the probability is very low, sampling with replacement can result in selecting one of the 10 balls labeled as "B" in all 90 trials. On the other hand, sampling without replacement results in selecting balls labeled as "B" at most 10 times and balls labeled as "A" at least 80 times. Thus, in this situation, the probability mass functions of the hypergeometric distribution and the binomial distribution are quite different, as illustrated in Fig. 3.5.

The expectation and variance of $\mathrm{HG}(N, M, n)$ are given by

$$E[x] = \frac{nM}{N} \quad \text{and} \quad V[x] = \frac{nM(N-M)(N-n)}{N^2(N-1)},$$

which are proved below.

The expectation $E[x]$ can be expressed as

$$\begin{aligned}
E[x] &= \frac{1}{\binom{N}{n}} \sum_{x=0}^{n} x \binom{M}{x}\binom{N-M}{n-x} \\
&= \frac{1}{\binom{N}{n}} \sum_{x=1}^{n} x \binom{M}{x}\binom{N-M}{n-x} \quad \text{(the term with } x = 0 \text{ is zero)} \\
&= \frac{M}{\binom{N}{n}} \sum_{x=1}^{n} \binom{M-1}{x-1}\binom{N-M}{n-x} \quad \left(\binom{M}{x} = \frac{M}{x}\binom{M-1}{x-1}\right) \\
&= \frac{M}{\binom{N}{n}} \sum_{x=0}^{n-1} \binom{M-1}{x}\binom{N-M}{n-x-1} \quad (\text{let } x \leftarrow x-1) \\
&= \frac{nM}{N} \frac{1}{\binom{N-1}{n-1}} \sum_{x=0}^{n-1} \binom{M-1}{x}\binom{N-M}{n-x-1} \quad \left(\binom{N}{n} = \frac{N}{n}\binom{N-1}{n-1}\right).
\end{aligned} \tag{3.1}$$

Since the probability mass function satisfies $\sum_x f(x) = 1$,

$$\binom{N}{n} = \sum_{x=0}^{n} \binom{M}{x}\binom{N-M}{n-x}. \tag{3.2}$$

Letting $M \leftarrow M-1$, $N \leftarrow N-1$, and $n \leftarrow n-1$ in Eq. (3.2) yields

$$\binom{N-1}{n-1} = \sum_{x=0}^{n-1} \binom{M-1}{x}\binom{N-M}{n-x-1},$$

and substituting this into Eq. (3.1) gives $E[x] = \frac{nM}{N}$.

The variance $V[x]$ can be expressed as

$$V[x] = E[x(x-1)] + E[x] - (E[x])^2. \tag{3.3}$$

Similar derivation to the expectation and using Eq. (3.2) yield

$$E[x(x-1)] = \frac{n(n-1)M(M-1)}{N(N-1)},$$

and substituting this into Eq. (3.3) gives $V[x] = \frac{nM(N-M)(N-n)}{N^2(N-1)}$.

The moment-generating function of $\mathrm{HG}(N, M, n)$ is given by

$$M_x(t) = E[e^{tx}] = \frac{\binom{N-M}{n}}{\binom{N}{n}} F(-n, -M, N-M-n+1, e^t),$$

where

$$F(a,b,c,d) = \sum_{x=0}^{\infty} \frac{(a)_x(b)_x}{(c)_x} \frac{d^x}{x!},$$

$$(a)_x = \begin{cases} a(a+1)\cdots(a+x-1) & (x > 0) \\ 1 & (x = 0) \end{cases}$$

is the *hypergeometric series*. The name, the hypergeometric distribution, stems from the fact that its moment-generating function can be expressed using the hypergeometric series.

3.4 POISSON DISTRIBUTION

If the probability of success in Bernoulli trials is very small, every trial is almost always failed. However, even if the probability of success, p, is extremely small, the Bernoulli trials will succeed a small number of times, as long as the number of trials, n, is large enough. Indeed, given that the expectation of binomial distribution $\mathrm{Bi}(n,p)$ is np, repeating the Bernoulli trials $n = 10000000$ times with probability of success $p = 0.0000003$ yields three successful trials on average:

$$np = 10000000 \times 0.0000003 = 3.$$

This implies that the probability that the number of successful trials x is nonzero may not be that small, if the number of trials n is large enough. More precisely, given that the probability mass function of $\mathrm{Bi}(n,p)$ is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x},$$

the probability for $x = 5$ is

$$\binom{10000000}{5} (0.0000003)^5 (0.9999997)^{9999995}.$$

However, calculating $(0.9999997)^{9999995}$ requires 9999995-time multiplications of 0.9999997, which is computationally expensive. At a glance, the use of approximation $0.9999997 \approx 1$ works fine:

$$(0.9999997)^{9999995} \approx 1^{9999995} = 1.$$

However, the correct value is

$$(0.9999997)^{9999995} \approx 0.0498 \ll 1,$$

and thus the above approximation is very poor.

This problem can be overcome by applying *Poisson's law of small numbers*: for $p = \lambda/n$,

$$\lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} = \frac{e^{-\lambda}\lambda^x}{x!}.$$

Let us prove this. First, the left-hand side of the above equation can be expressed as

$$\begin{aligned}
&\lim_{n\to\infty} \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!} \lim_{n\to\infty} \frac{n!}{(n-x)!n^x} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-x}. \qquad (3.4)
\end{aligned}$$

Here, it holds that

$$\begin{aligned}
\lim_{n\to\infty} \frac{n!}{(n-x)!n^x} &= \lim_{n\to\infty} \frac{n}{n} \times \frac{n-1}{n} \times \cdots \times \frac{n-x+1}{n} \\
&= \lim_{n\to\infty} 1 \times \frac{1-\frac{1}{n}}{1} \times \cdots \times \frac{1-\frac{x}{n}+\frac{1}{n}}{1} = 1.
\end{aligned}$$

Also, setting $t = -\frac{\lambda}{n}$ in the definition of the *Euler number e*,

$$e = \lim_{t\to 0} (1+t)^{\frac{1}{t}}, \qquad (3.5)$$

yields

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}.$$

Furthermore,

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

holds. Putting them together yields

$$\lim_{n\to\infty} \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} = \frac{e^{-\lambda}\lambda^x}{x!}.$$

The *Poisson distribution*, denoted by $\text{Po}(\lambda)$, is the probability distribution whose probability mass function is given by

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}. \qquad (3.6)$$

Since this corresponds to the binomial distribution with $p = \lambda/n$, if $1/n$ is regarded as time and an event occurs λ times on average in unit time, $f(x)$ corresponds to the probability that the event occurs x times in unit time.

Eq. (3.6) is non-negative and the Taylor series expansion of the exponential function at the origin,

$$e^{\lambda} = 1 + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \cdots = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}, \tag{3.7}$$

yields

$$\sum_{x=0}^{\infty} f(x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1.$$

Thus, Eq. (3.6) is shown to be the probability mass function.

The moment-generating function of $\mathrm{Po}(\lambda)$ is given by

$$M_x(t) = E[e^{tx}] = \sum_{x=0}^{\infty} \frac{e^{tx}e^{-\lambda}\lambda^x}{x!} = \exp\left(\lambda(e^t - 1)\right),$$

where Eq. (3.7) is used for the derivation. From this, the expectation and variance of $\mathrm{Po}(\lambda)$ are obtained as

$$E[x] = \lambda \quad \text{and} \quad V[x] = \lambda.$$

Interestingly, the expectation and variance of the Poisson distribution are equal.

It was shown in Section 3.2 that the expectation and variance of the binomial distribution (i.e., without applying Poisson's law of small numbers) are given by

$$E[x] = np \quad \text{and} \quad V[x] = np(1 - p).$$

Setting $p = \lambda/n$ yields

$$\lim_{n\to\infty} np = \lambda \quad \text{and} \quad \lim_{n\to\infty} np(1 - p) = \lambda,$$

which imply that application of Poisson's law of small numbers does not essentially change the expectation and variance.

Probability mass functions of $\mathrm{Po}(\lambda)$ are illustrated in Fig. 3.6, showing that the expectation and variance grow as λ is increased.

3.5 NEGATIVE BINOMIAL DISTRIBUTION

Let us consider Bernoulli trials with probability of success p. Then the number of unsuccessful trials x until the kth success is obtained follows the *negative binomial distribution*, which is denoted by $\mathrm{NB}(k, p)$.

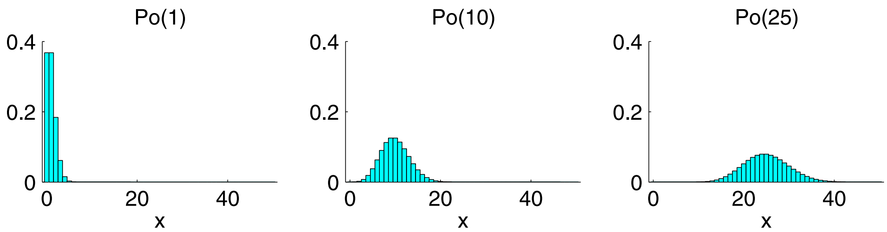


FIGURE 3.6

Probability mass functions of Poisson distribution Po(λ).

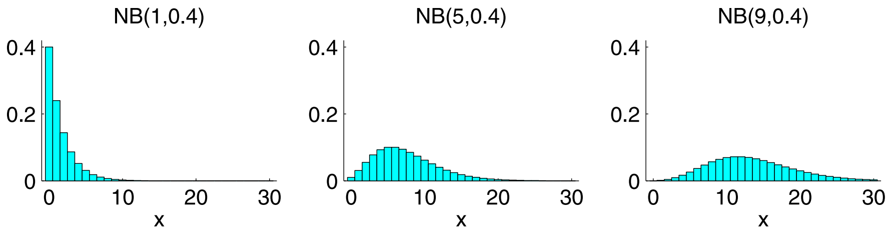


FIGURE 3.7

Probability mass functions of negative binomial distribution NB(k,p).

Since the kth success is obtained at the $(k+x)$th trial, the $(k+x)$th trial is always successful. On the other hand, the number of combinations of x unsuccessful trials in the first $(k+x-1)$ trials is given by $\binom{k+x-1}{x}$. Putting them together, the probability mass function of NB(k,p) is given by

$$f(x) = \binom{k+x-1}{x} p^k (1-p)^x. \tag{3.8}$$

Probability mass functions of NB(k,p) for $p = 0.4$ are illustrated in Fig. 3.7.

The binomial coefficient $\binom{r}{x}$ can be extended to negative number $r = -k < 0$:

$$\binom{-k}{x} = \frac{(-k-x+1)(-k-x+2)\cdots(-k-1)(-k)}{x(x-1)\cdots 2\cdot 1}.$$

With this *negative binomial coefficient*, the probability mass function defined in (3.8) can be expressed as

$$\begin{aligned} f(x) &= \frac{(k+x-1)(k+x-2)\cdots(k+1)k}{x(x-1)\cdots 2\cdot 1} p^k (1-p)^x \\ &= (-1)^x \binom{-k}{x} p^k (1-p)^x. \end{aligned} \tag{3.9}$$

The name, the negative binomial distribution, stems from this fact.

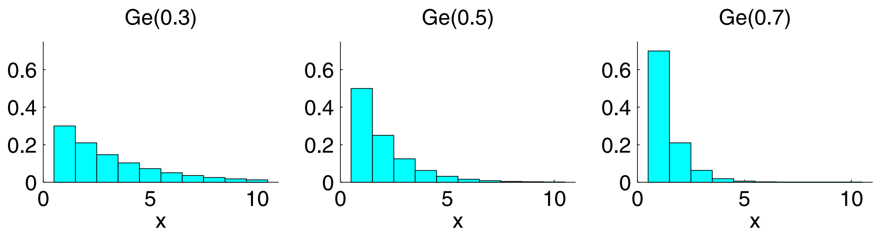


FIGURE 3.8

Probability mass functions of geometric distribution Ge(p).

The *binomial theorem* explained in Section 3.2 can also be generalized to negative numbers as

$$\sum_{x=0}^{\infty} \binom{-k}{x} t^x = (1+t)^{-k}.$$

Setting $t = p - 1$ in the above equation yields that Eq. (3.9) satisfies

$$\sum_{x=0}^{\infty} f(x) = p^k \sum_{x=0}^{\infty} \binom{-k}{x} (p-1)^x = 1.$$

This generalized binomial theorem allows us to obtain the moment-generating function of NB(k,p) as

$$\begin{aligned} M_x(t) = E[e^{tx}] &= \sum_{x=0}^{\infty} e^{tx} \binom{-k}{x} p^k (p-1)^x \\ &= p^k \sum_{x=0}^{\infty} \binom{-k}{x} \{(p-1)e^t\}^x = \left(\frac{p}{1-(1-p)e^t}\right)^k. \end{aligned}$$

From this, the expectation and variance of NB(k,p) are obtained as

$$E[x] = \frac{k(1-p)}{p} \quad \text{and} \quad V[x] = \frac{k(1-p)}{p^2}.$$

The negative binomial distribution is also referred to as the *Pascal distribution*.

3.6 GEOMETRIC DISTRIBUTION

Let us consider Bernoulli trials with probability of success p. Then the number of unsuccessful trials x until the first success is obtained follows the *geometric distribution*, which is denoted by Ge(p). The geometric distribution Ge(p) is equivalent to the negative binomial distribution $\mathrm{NB}(k, p)$ with $k = 1$. Thus, its probability mass function is given by

$$f(x) = p(1-p)^x,$$

where the probability decreases exponentially as x increases. Probability mass functions of $\mathrm{Ge}(p)$ are illustrated in Fig. 3.8.

Given that $\mathrm{Ge}(p)$ is equivalent to $\mathrm{NB}(1, p)$, its moment-generating function is given by

$$M_x(t) = \frac{p}{1-(1-p)e^t},$$

and the expectation and variance are given by

$$E[x] = \frac{1-p}{p} \quad \text{and} \quad V[x] = \frac{1-p}{p^2}.$$