Chapter 5

Some Discrete Probability Distributions

5.1 Introduction and Motivation

No matter whether a discrete probability distribution is represented graphically by a histogram, in tabular form, or by means of a formula, the behavior of a random variable is described. Often, the observations generated by different statistical experiments have the same general type of behavior. Consequently, discrete random variables associated with these experiments can be described by essentially the same probability distribution and therefore can be represented by a single formula. In fact, one needs only a handful of important probability distributions to describe many of the discrete random variables encountered in practice.

Such a handful of distributions describe several real-life random phenomena. For instance, in a study involving testing the effectiveness of a new drug, the number of cured patients among all the patients who use the drug approximately follows a binomial distribution (Section 5.2). In an industrial example, when a sample of items selected from a batch of production is tested, the number of defective items in the sample usually can be modeled as a hypergeometric random variable (Section 5.3). In a statistical quality control problem, the experimenter will signal a shift of the process mean when observational data exceed certain limits. The number of samples required to produce a false alarm follows a geometric distribution which is a special case of the negative binomial distribution (Section 5.4). On the other hand, the number of white cells from a fixed amount of an individual's blood sample is usually random and may be described by a Poisson distribution (Section 5.5). In this chapter, we present these commonly used distributions with various examples.

5.2 Binomial and Multinomial Distributions

An experiment often consists of repeated trials, each with two possible outcomes that may be labeled **success** or **failure**. The most obvious application deals with

the testing of items as they come off an assembly line, where each trial may indicate a defective or a nondefective item. We may choose to define either outcome as a success. The process is referred to as a **Bernoulli process**. Each trial is called a **Bernoulli trial**. Observe, for example, if one were drawing cards from a deck, the probabilities for repeated trials change if the cards are not replaced. That is, the probability of selecting a heart on the first draw is 1/4, but on the second draw it is a conditional probability having a value of 13/51 or 12/51, depending on whether a heart appeared on the first draw: this, then, would no longer be considered a set of Bernoulli trials.

The Bernoulli Process

Strictly speaking, the Bernoulli process must possess the following properties:

1. The experiment consists of repeated trials.
2. Each trial results in an outcome that may be classified as a success or a failure.
3. The probability of success, denoted by p, remains constant from trial to trial.
4. The repeated trials are independent.

Consider the set of Bernoulli trials where three items are selected at random from a manufacturing process, inspected, and classified as defective or nondefective. A defective item is designated a success. The number of successes is a random variable X assuming integral values from 0 through 3. The eight possible outcomes and the corresponding values of X are

Outcome	NNN	NDN	NND	DNN	NDD	DND	DDN	DDD
$\boldsymbol{x}$	0	1	1	1	2	2	2	3

Since the items are selected independently and we assume that the process produces 25% defectives, we have

$$P(NDN) = P(N)P(D)P(N) = \left(\frac{3}{4}\right)\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = \frac{9}{64}.$$

Similar calculations yield the probabilities for the other possible outcomes. The probability distribution of X is therefore

x	0	1	2	3
$f(x)$	$\frac{27}{64}$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{1}{64}$

Binomial Distribution

The number X of successes in n Bernoulli trials is called a **binomial random variable**. The probability distribution of this discrete random variable is called the **binomial distribution**, and its values will be denoted by $b(x; n, p)$ since they depend on the number of trials and the probability of a success on a given trial. Thus, for the probability distribution of X, the number of defectives is

$$P(X = 2) = f(2) = b\left(2; 3, \frac{1}{4}\right) = \frac{9}{64}.$$

Let us now generalize the above illustration to yield a formula for $b(x; n, p)$. That is, we wish to find a formula that gives the probability of x successes in n trials for a binomial experiment. First, consider the probability of x successes and $n - x$ failures in a specified order. Since the trials are independent, we can multiply all the probabilities corresponding to the different outcomes. Each success occurs with probability p and each failure with probability $q = 1 - p$. Therefore, the probability for the specified order is $p^x q^{n-x}$. We must now determine the total number of sample points in the experiment that have x successes and $n-x$ failures. This number is equal to the number of partitions of n outcomes into two groups with x in one group and $n-x$ in the other and is written $\binom{n}{x}$ as introduced in Section 2.3. Because these partitions are mutually exclusive, we add the probabilities of all the different partitions to obtain the general formula, or simply multiply $p^x q^{n-x}$ by $\binom{n}{x}$.

Binomial Distribution

A Bernoulli trial can result in a success with probability p and a failure with probability $q = 1 - p$. Then the probability distribution of the binomial random variable X, the number of successes in n independent trials, is

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

Note that when $n = 3$ and $p = 1/4$, the probability distribution of X, the number of defectives, may be written as

$$b\left(x; 3, \frac{1}{4}\right) = \binom{3}{x}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{3-x}, \quad x = 0, 1, 2, 3,$$

rather than in the tabular form on page 144.

Example 5.1: The probability that a certain kind of component will survive a shock test is 3/4. Find the probability that exactly 2 of the next 4 components tested survive.

***Solution*:** Assuming that the tests are independent and $p = 3/4$ for each of the 4 tests, we obtain

$$b\left(2; 4, \frac{3}{4}\right) = \binom{4}{2}\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)^2 = \left(\frac{4!}{2!\,2!}\right)\left(\frac{3^2}{4^4}\right) = \frac{27}{128}.$$

Where Does the Name *Binomial* Come From?

The binomial distribution derives its name from the fact that the $n + 1$ terms in the binomial expansion of $(q + p)^n$ correspond to the various values of $b(x; n, p)$ for $x = 0, 1, 2, \ldots, n$. That is,

$$\begin{aligned}(q+p)^n &= \binom{n}{0}q^n + \binom{n}{1}pq^{n-1} + \binom{n}{2}p^2q^{n-2} + \cdots + \binom{n}{n}p^n \\ &= b(0; n, p) + b(1; n, p) + b(2; n, p) + \cdots + b(n; n, p).\end{aligned}$$

Since $p + q = 1$, we see that

$$\sum_{x=0}^{n} b(x; n, p) = 1,$$

a condition that must hold for any probability distribution.

Frequently, we are interested in problems where it is necessary to find $P(X < r)$ or $P(a \le X \le b)$. Binomial sums

$$B(r; n, p) = \sum_{x=0}^{r} b(x; n, p)$$

are given in Table A.1 of the Appendix for $n = 1, 2, \ldots, 20$ for selected values of p from 0.1 to 0.9. We illustrate the use of Table A.1 with the following example.

Example 5.2: The probability that a patient recovers from a rare blood disease is 0.4. If 15 people are known to have contracted this disease, what is the probability that (a) at least 10 survive, (b) from 3 to 8 survive, and (c) exactly 5 survive?

Solution: Let X be the number of people who survive.

(a) $$P(X \ge 10) = 1 - P(X < 10) = 1 - \sum_{x=0}^{9} b(x; 15, 0.4) = 1 - 0.9662$$
$$= 0.0338$$

(b) $$P(3 \le X \le 8) = \sum_{x=3}^{8} b(x; 15, 0.4) = \sum_{x=0}^{8} b(x; 15, 0.4) - \sum_{x=0}^{2} b(x; 15, 0.4)$$
$$= 0.9050 - 0.0271 = 0.8779$$

(c) $$P(X = 5) = b(5; 15, 0.4) = \sum_{x=0}^{5} b(x; 15, 0.4) - \sum_{x=0}^{4} b(x; 15, 0.4)$$
$$= 0.4032 - 0.2173 = 0.1859$$

Example 5.3: A large chain retailer purchases a certain kind of electronic device from a manufacturer. The manufacturer indicates that the defective rate of the device is 3%.

(a) The inspector randomly picks 20 items from a shipment. What is the probability that there will be at least one defective item among these 20?

(b) Suppose that the retailer receives 10 shipments in a month and the inspector randomly tests 20 devices per shipment. What is the probability that there will be exactly 3 shipments each containing at least one defective device among the 20 that are selected and tested from the shipment?

Solution: (a) Denote by X the number of defective devices among the 20. Then X follows a $b(x; 20, 0.03)$ distribution. Hence,

$$P(X \ge 1) = 1 - P(X = 0) = 1 - b(0; 20, 0.03)$$
$$= 1 - (0.03)^0(1 - 0.03)^{20-0} = 0.4562.$$

(b) In this case, each shipment can either contain at least one defective item or not. Hence, testing of each shipment can be viewed as a Bernoulli trial with $p = 0.4562$ from part (a). Assuming independence from shipment to shipment

and denoting by Y the number of shipments containing at least one defective item, Y follows another binomial distribution $b(y; 10, 0.4562)$. Therefore,

$$P(Y = 3) = \binom{10}{3} 0.4562^3 (1 - 0.4562)^7 = 0.1602.$$

Areas of Application

From Examples 5.1 through 5.3, it should be clear that the binomial distribution finds applications in many scientific fields. An industrial engineer is keenly interested in the "proportion defective" in an industrial process. Often, quality control measures and sampling schemes for processes are based on the binomial distribution. This distribution applies to any industrial situation where an outcome of a process is dichotomous and the results of the process are independent, with the probability of success being constant from trial to trial. The binomial distribution is also used extensively for medical and military applications. In both fields, a success or failure result is important. For example, "cure" or "no cure" is important in pharmaceutical work, and "hit" or "miss" is often the interpretation of the result of firing a guided missile.

Since the probability distribution of any binomial random variable depends only on the values assumed by the parameters n, p, and q, it would seem reasonable to assume that the mean and variance of a binomial random variable also depend on the values assumed by these parameters. Indeed, this is true, and in the proof of Theorem 5.1 we derive general formulas that can be used to compute the mean and variance of any binomial random variable as functions of n, p, and q.

Theorem 5.1: The mean and variance of the binomial distribution $b(x; n, p)$ are

$$\mu = np \text{ and } \sigma^2 = npq.$$

***Proof*:** Let the outcome on the jth trial be represented by a Bernoulli random variable I_j, which assumes the values 0 and 1 with probabilities q and p, respectively. Therefore, in a binomial experiment the number of successes can be written as the sum of the n independent indicator variables. Hence,

$$X = I_1 + I_2 + \cdots + I_n.$$

The mean of any I_j is $E(I_j) = (0)(q) + (1)(p) = p$. Therefore, using Corollary 4.4 on page 131, the mean of the binomial distribution is

$$\mu = E(X) = E(I_1) + E(I_2) + \cdots + E(I_n) = \underbrace{p + p + \cdots + p}_{n \text{ terms}} = np.$$

The variance of any I_j is $\sigma^2_{I_j} = E(I_j^2) - p^2 = (0)^2(q) + (1)^2(p) - p^2 = p(1-p) = pq$. Extending Corollary 4.11 to the case of n independent Bernoulli variables gives the variance of the binomial distribution as

$$\sigma^2_X = \sigma^2_{I_1} + \sigma^2_{I_2} + \cdots + \sigma^2_{I_n} = \underbrace{pq + pq + \cdots + pq}_{n \text{ terms}} = npq.$$

Example 5.4: It is conjectured that an impurity exists in 30% of all drinking wells in a certain rural community. In order to gain some insight into the true extent of the problem, it is determined that some testing is necessary. It is too expensive to test all of the wells in the area, so 10 are randomly selected for testing.

(a) Using the binomial distribution, what is the probability that exactly 3 wells have the impurity, assuming that the conjecture is correct?

(b) What is the probability that more than 3 wells are impure?

Solution: (a) We require

$$b(3; 10, 0.3) = \sum_{x=0}^{3} b(x; 10, 0.3) - \sum_{x=0}^{2} b(x; 10, 0.3) = 0.6496 - 0.3828 = 0.2668.$$

(b) In this case, $P(X > 3) = 1 - 0.6496 = 0.3504$.

Example 5.5: Find the mean and variance of the binomial random variable of Example 5.2, and then use Chebyshev's theorem (on page 137) to interpret the interval $\mu \pm 2\sigma$.

Solution: Since Example 5.2 was a binomial experiment with $n = 15$ and $p = 0.4$, by Theorem 5.1, we have

$$\mu = (15)(0.4) = 6 \text{ and } \sigma^2 = (15)(0.4)(0.6) = 3.6.$$

Taking the square root of 3.6, we find that $\sigma = 1.897$. Hence, the required interval is $6 \pm (2)(1.897)$, or from 2.206 to 9.794. Chebyshev's theorem states that the number of recoveries among 15 patients who contracted the disease has a probability of at least 3/4 of falling between 2.206 and 9.794 or, because the data are discrete, between 2 and 10 inclusive.

There are solutions in which the computation of binomial probabilities may allow us to draw a scientific inference about population after data are collected. An illustration is given in the next example.

Example 5.6: Consider the situation of Example 5.4. The notion that 30% of the wells are impure is merely a conjecture put forth by the area water board. Suppose 10 wells are randomly selected and 6 are found to contain the impurity. What does this imply about the conjecture? Use a probability statement.

Solution: We must first ask: "If the conjecture is correct, is it likely that we would find 6 or more impure wells?"

$$P(X \geq 6) = \sum_{x=0}^{10} b(x; 10, 0.3) - \sum_{x=0}^{5} b(x; 10, 0.3) = 1 - 0.9527 = 0.0473.$$

As a result, it is very unlikely (4.7% chance) that 6 or more wells would be found impure if only 30% of all are impure. This casts considerable doubt on the conjecture and suggests that the impurity problem is much more severe.

As the reader should realize by now, in many applications there are more than two possible outcomes. To borrow an example from the field of genetics, the color of guinea pigs produced as offspring may be red, black, or white. Often the "defective" or "not defective" dichotomy is truly an oversimplification in engineering situations. Indeed, there are often more than two categories that characterize items or parts coming off an assembly line.

Multinomial Experiments and the Multinomial Distribution

The binomial experiment becomes a **multinomial experiment** if we let each trial have more than two possible outcomes. The classification of a manufactured product as being light, heavy, or acceptable and the recording of accidents at a certain intersection according to the day of the week constitute multinomial experiments. The drawing of a card from a deck *with replacement* is also a multinomial experiment if the 4 suits are the outcomes of interest.

In general, if a given trial can result in any one of k possible outcomes $E_1, E_2, \ldots, E_k$ with probabilities $p_1, p_2, \ldots, p_k$, then the **multinomial distribution** will give the probability that E_1 occurs x_1 times, E_2 occurs x_2 times, ..., and E_k occurs x_k times in n independent trials, where

$$x_1 + x_2 + \cdots + x_k = n.$$

We shall denote this joint probability distribution by

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n).$$

Clearly, $p_1 + p_2 + \cdots + p_k = 1$, since the result of each trial must be one of the k possible outcomes.

To derive the general formula, we proceed as in the binomial case. Since the trials are independent, any specified order yielding x_1 outcomes for E_1, x_2 for $E_2, \ldots, x_k$ for E_k will occur with probability $p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$. The total number of orders yielding similar outcomes for the n trials is equal to the number of partitions of n items into k groups with x_1 in the first group, x_2 in the second group, ..., and x_k in the kth group. This can be done in

$$\binom{n}{x_1, x_2, \ldots, x_k} = \frac{n!}{x_1!\, x_2! \cdots x_k!}$$

ways. Since all the partitions are mutually exclusive and occur with equal probability, we obtain the multinomial distribution by multiplying the probability for a specified order by the total number of partitions.

Multinomial Distribution

If a given trial can result in the k outcomes $E_1, E_2, \ldots, E_k$ with probabilities $p_1, p_2, \ldots, p_k$, then the probability distribution of the random variables $X_1, X_2, \ldots, X_k$, representing the number of occurrences for $E_1, E_2, \ldots, E_k$ in n independent trials, is

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1}\, p_2^{x_2} \cdots p_k^{x_k},$$

with

$$\sum_{i=1}^{k} x_i = n \text{ and } \sum_{i=1}^{k} p_i = 1.$$

The multinomial distribution derives its name from the fact that the terms of the multinomial expansion of $(p_1 + p_2 + \cdots + p_k)^n$ correspond to all the possible values of $f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n)$.

Example 5.7: The complexity of arrivals and departures of planes at an airport is such that computer simulation is often used to model the "ideal" conditions. For a certain airport with three runways, it is known that in the ideal setting the following are the probabilities that the individual runways are accessed by a randomly arriving commercial jet:

Runway 1: $p_1 = 2/9$,
Runway 2: $p_2 = 1/6$,
Runway 3: $p_3 = 11/18$.

What is the probability that 6 randomly arriving airplanes are distributed in the following fashion?

Runway 1: 2 airplanes,
Runway 2: 1 airplane,
Runway 3: 3 airplanes

***Solution*:** Using the multinomial distribution, we have

$$f\left(2,1,3;\frac{2}{9},\frac{1}{6},\frac{11}{18},6\right) = \binom{6}{2,1,3}\left(\frac{2}{9}\right)^2\left(\frac{1}{6}\right)^1\left(\frac{11}{18}\right)^3$$
$$= \frac{6!}{2!\,1!\,3!}\cdot\frac{2^2}{9^2}\cdot\frac{1}{6}\cdot\frac{11^3}{18^3} = 0.1127.$$

Exercises

5.1 A random variable X that assumes the values $x_1, x_2, \ldots, x_k$ is called a discrete uniform random variable if its probability mass function is $f(x) = \frac{1}{k}$ for all of $x_1, x_2, \ldots, x_k$ and 0 otherwise. Find the mean and variance of X.

5.2 Twelve people are given two identical speakers, which they are asked to listen to for differences, if any. Suppose that these people answer simply by guessing. Find the probability that three people claim to have heard a difference between the two speakers.

5.3 An employee is selected from a staff of 10 to supervise a certain project by selecting a tag at random from a box containing 10 tags numbered from 1 to 10. Find the formula for the probability distribution of X representing the number on the tag that is drawn. What is the probability that the number drawn is less than 4?

5.4 In a certain city district, the need for money to buy drugs is stated as the reason for 75% of all thefts. Find the probability that among the next 5 theft cases reported in this district,

(a) exactly 2 resulted from the need for money to buy drugs;

(b) at most 3 resulted from the need for money to buy drugs.

5.5 According to *Chemical Engineering Progress* (November 1990), approximately 30% of all pipework failures in chemical plants are caused by operator error.

(a) What is the probability that out of the next 20 pipework failures at least 10 are due to operator error?

(b) What is the probability that no more than 4 out of 20 such failures are due to operator error?

(c) Suppose, for a particular plant, that out of the random sample of 20 such failures, exactly 5 are due to operator error. Do you feel that the 30% figure stated above applies to this plant? Comment.

5.6 According to a survey by the Administrative Management Society, one-half of U.S. companies give employees 4 weeks of vacation after they have been with the company for 15 years. Find the probability that among 6 companies surveyed at random, the number that give employees 4 weeks of vacation after 15 years of employment is

(a) anywhere from 2 to 5;

(b) fewer than 3.

5.7 One prominent physician claims that 70% of those with lung cancer are chain smokers. If his assertion is correct,

(a) find the probability that of 10 such patients