

## Random Variable and Probability Distributions

### Random variable:

Statistics is concerned with making inferences about populations and population characteristics. Experiments are conducted with results that are subject to chance. The testing of a number of electronic components is an example of a statistical experiment, a term that is used to describe any process by which several chance observations are generated. It is often important to allocate a numerical description to the outcome. For example, the sample space giving a detailed description of each possible outcome when three electronic components are tested may be written

$$S = \{NNN, NND, NDN, DNN, NDD, DND, DDN, DDD\},$$

where N denotes non-defective and D denotes defective. One is naturally concerned with the number of defectives that occur. Thus, each point in the sample space will be assigned a numerical value of 0, 1, 2, or 3. These values are, of course, random quantities determined by the outcome of the experiment. They may be viewed as values assumed by the random variable  $X$ , the number of defective items when three electronic components are tested.

A **random variable** is a function that associates a real number with each element in the sample space.

### Example:

Two balls are drawn in succession without replacement from an urn containing 4 red balls and 3 black balls. The possible outcomes and the values  $y$  of the random variable  $Y$ , where  $Y$  is the number of red balls, are

Sample Space	$y$
$RR$	2
$RB$	1
$BR$	1
$BB$	0

### Example:

Consider the simple condition in which components are arriving from the production line and they are stipulated to be defective or not defective. Define the random variable  $X$  by

$$X = \begin{cases} 1, & \text{if the component is defective,} \\ 0, & \text{if the component is not defective.} \end{cases}$$

Clearly the assignment of 1 or 0 is arbitrary though quite convenient. This will become clear in later chapters. The random variable for which 0 and 1 are chosen to describe the two possible values is called a **Bernoulli random variable**.

### Example:

A stockroom clerk returns three safety helmets at random to three steel mill employees who had previously checked them. If Smith, Jones, and Brown, in that order, receive one of the three hats, list the sample points for the possible orders of returning the helmets, and find the value  $m$  of the random variable  $M$  that represents the number of correct matches.

If  $S$ ,  $J$ , and  $B$  stand for Smith's, Jones's, and Brown's helmets, respectively, then the possible arrangements in which the helmets may be returned and the number of correct matches are

Sample Space	$m$
$SJB$	3
$SBJ$	1
$BJS$	1
$JSB$	1
$JBS$	0
$BSJ$	0

### Discrete Sample Space:

If a sample space contains a finite number of possibilities or an unending sequence with as many elements as there are whole numbers, it is called a **discrete sample space**.

### Continuous Sample Space:

If a sample space contains an infinite number of possibilities equal to the number of points on a line segment, it is called a **continuous sample space**.

### Discrete Probability Distribution:

A discrete random variable assumes each of its values with a certain probability. In the case of tossing a coin three times, the variable  $X$ , representing the number of heads, assumes the value 2 with probability  $3/8$ , since 3 of the 8 equally likely sample points result in two heads and one tail. If one assumes equal weights for the simple events, the probability that no employee gets back the right helmet, that is, the probability that  $M$  assumes the value 0, is  $1/3$ . The possible values  $m$  of  $M$  and their probabilities are

$m$	0	1	3
$P(M = m)$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$

Note that the values of  $m$  exhaust all possible cases and hence the probabilities add to 1.

Frequently, it is convenient to represent all the probabilities of a random variable  $X$  by a formula. Such a formula would necessarily be a function of the numerical values  $x$  that we shall denote by  $f(x)$ ,  $g(x)$ ,  $r(x)$ , and so forth. Therefore, we write  $f(x) = P(X = x)$ ; that is,  $f(3) = P(X = 3)$ . The set of ordered pairs  $(x, f(x))$  is called the **probability function**, **probability mass function**, or **probability distribution** of the discrete random variable  $X$ .

## Probability Mass Function:

The set of ordered pairs  $(x, f(x))$  is a **probability function**, **probability mass function**, or **probability distribution** of the discrete random variable  $X$  if, for each possible outcome  $x$ ,

1.  $f(x) \geq 0$ ,
2.  $\sum_x f(x) = 1$ ,
3.  $P(X = x) = f(x)$ .

### Example:

A shipment of 20 similar laptop computers to a retail outlet contains 3 that are defective. If a school makes a random purchase of 2 of these computers, find the probability distribution for the number of defectives.

Let  $X$  be a random variable whose values  $x$  are the possible numbers of defective computers purchased by the school. Then  $x$  can only take the numbers 0, 1, and

2. Now

$$f(0) = P(X = 0) = \frac{\binom{3}{0}\binom{17}{2}}{\binom{20}{2}} = \frac{68}{95}, \quad f(1) = P(X = 1) = \frac{\binom{3}{1}\binom{17}{1}}{\binom{20}{2}} = \frac{51}{190},$$
$$f(2) = P(X = 2) = \frac{\binom{3}{2}\binom{17}{0}}{\binom{20}{2}} = \frac{3}{190}.$$

Thus, the probability distribution of  $X$  is

$x$	0	1	2
$f(x)$	$\frac{68}{95}$	$\frac{51}{190}$	$\frac{3}{190}$



### Example 3.9:

If a car agency sells 50% of its inventory of a certain foreign car equipped with side airbags, find a formula for the probability distribution of the number of cars with side airbags among the next 4 cars sold by the agency.

Since the probability of selling an automobile with side airbags is 0.5, the  $2^4 = 16$  points in the sample space are equally likely to occur. Therefore, the denominator for all probabilities, and also for our function, is 16. To obtain the number of ways of selling 3 cars with side airbags, we need to consider the number of ways of partitioning 4 outcomes into two cells, with 3 cars with side airbags assigned to one cell and the model without side airbags assigned to the other. This can be done in  $\binom{4}{3} = 4$  ways. In general, the event of selling  $x$  models with side airbags and  $4 - x$  models without side airbags can occur in  $\binom{4}{x}$  ways, where  $x$  can be 0, 1, 2, 3, or 4. Thus, the probability distribution  $f(x) = P(X = x)$  is

$$f(x) = \frac{1}{16} \binom{4}{x}, \quad \text{for } x = 0, 1, 2, 3, 4.$$



## Cumulative Distribution Function

There are many problems where we may wish to compute the probability that the observed value of a random variable  $X$  will be less than or equal to some real number  $x$ . Writing  $F(x) = P(X \leq x)$  for every real number  $x$ , we define  $F(x)$  to be the **cumulative distribution function** of the random variable  $X$ .

The **cumulative distribution function**  $F(x)$  of a discrete random variable  $X$  with probability distribution  $f(x)$  is

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), \quad \text{for } -\infty < x < \infty.$$

### Example:

Find the cumulative distribution function of the random variable  $X$  in Example 3.9. Using  $F(x)$ , verify that  $f(2) = 3/8$ .

Direct calculations of the probability distribution of Example 3.9 give  $f(0) = 1/16$ ,  $f(1) = 1/4$ ,  $f(2) = 3/8$ ,  $f(3) = 1/4$ , and  $f(4) = 1/16$ . Therefore,

$$F(0) = f(0) = \frac{1}{16},$$

$$F(1) = f(0) + f(1) = \frac{5}{16},$$

$$F(2) = f(0) + f(1) + f(2) = \frac{11}{16},$$

$$F(3) = f(0) + f(1) + f(2) + f(3) = \frac{15}{16},$$

$$F(4) = f(0) + f(1) + f(2) + f(3) + f(4) = 1.$$

Hence,

$$F(x) = \begin{cases} 0, & \text{for } x < 0, \\ \frac{1}{16}, & \text{for } 0 \leq x < 1, \\ \frac{5}{16}, & \text{for } 1 \leq x < 2, \\ \frac{11}{16}, & \text{for } 2 \leq x < 3, \\ \frac{15}{16}, & \text{for } 3 \leq x < 4, \\ 1 & \text{for } x \geq 4. \end{cases}$$

Now

$$f(2) = F(2) - F(1) = \frac{11}{16} - \frac{5}{16} = \frac{3}{8}.$$





## Probability Histogram:

It is often helpful to look at a probability distribution in graphic form. One might plot the points  $(x, f(x))$  of Example 3.9 to obtain Figure 3.1. By joining the points to the  $x$  axis either with a dashed or with a solid line, we obtain a probability mass function plot. Figure 3.1 makes it easy to see what values of  $X$  are most likely to occur, and it also indicates a perfectly symmetric situation in this case.

Instead of plotting the points  $(x, f(x))$ , we more frequently construct rectangles, as in Figure 3.2. Here the rectangles are constructed so that their bases of equal width are centered at each value  $x$  and their heights are equal to the corresponding probabilities given by  $f(x)$ . The bases are constructed so as to leave no space between the rectangles. Figure 3.2 is called a **probability histogram**.

Since each base in Figure 3.2 has unit width,  $P(X = x)$  is equal to the area of the rectangle centered at  $x$ . Even if the bases were not of unit width, we could adjust the heights of the rectangles to give areas that would still equal the probabilities of  $X$  assuming any of its values  $x$ . This concept of using areas to represent probabilities is necessary for our consideration of the probability distribution of a continuous random variable.

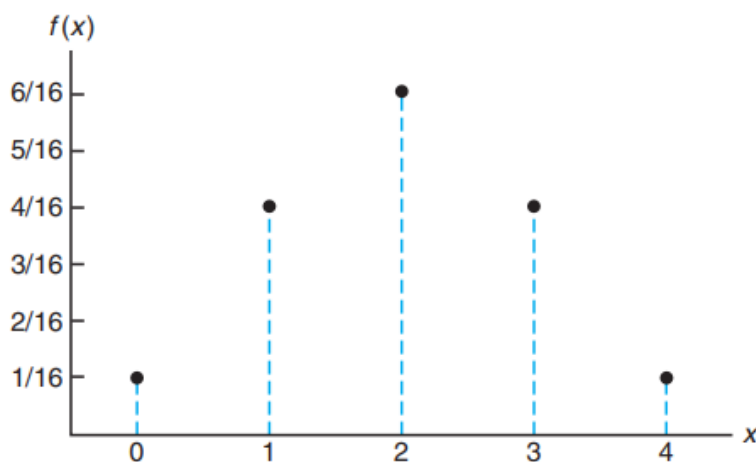


Figure 3.1: Probability mass function plot.

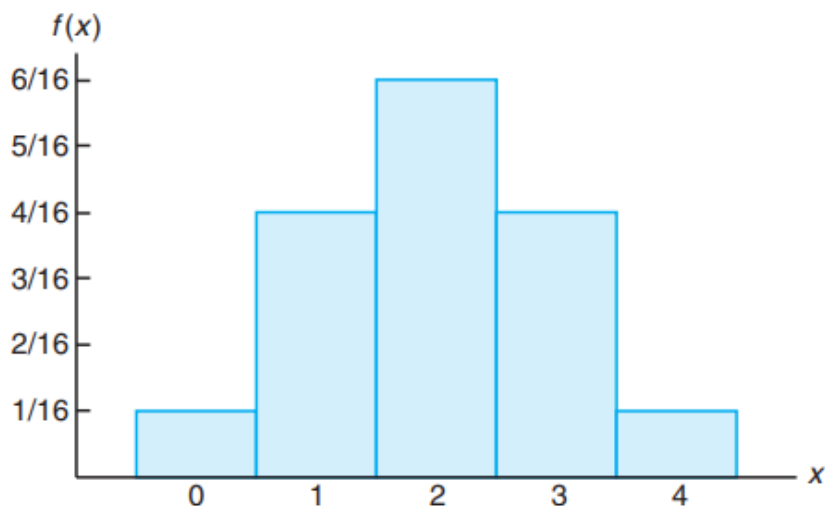


Figure 3.2: Probability histogram.

The graph of the cumulative distribution function of Example 3.9, which appears as a step function in Figure 3.3, is obtained by plotting the points  $(x, F(x))$ .

Certain probability distributions are applicable to more than one physical situation. The probability distribution of Example 3.9, for example, also applies to the random variable  $Y$ , where  $Y$  is the number of heads when a coin is tossed 4 times, or to the random variable  $W$ , where  $W$  is the number of red cards that occur when 4 cards are drawn at random from a deck in succession with each card replaced and the deck shuffled before the next drawing. Special discrete distributions that can be applied to many different experimental situations will be considered in Chapter 5.

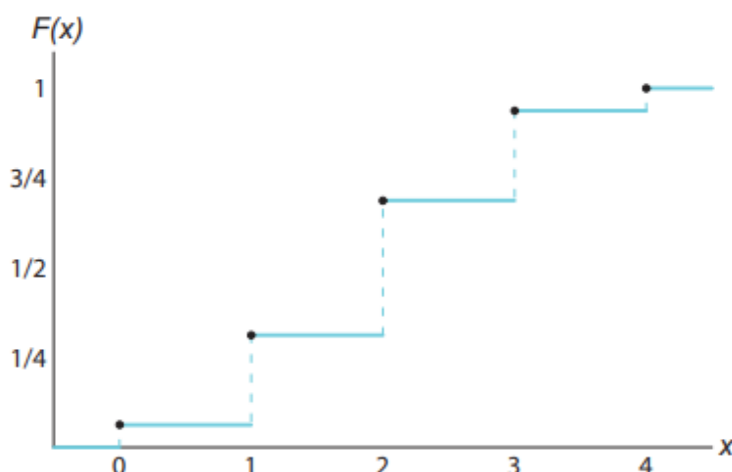


Figure 3.3: Discrete cumulative distribution function.

## Continuous Probability Distribution:

A continuous random variable has a probability of 0 of assuming *exactly* any of its values. Consequently, its probability distribution cannot be given in tabular form.

At first this may seem startling, but it becomes more plausible when we consider a particular example. Let us discuss a random variable whose values are the heights of all people over 21 years of age. Between any two values, say 163.5 and 164.5 centimeters, or even 163.99 and 164.01 centimeters, there are an infinite number of heights, one of which is 164 centimeters. The probability of selecting a person at random who is exactly 164 centimeters tall and not one of the infinitely large set of heights so close to 164 centimeters that you cannot humanly measure the difference is remote, and thus we assign a probability of 0 to the event. This is not the case, however, if we talk about the probability of selecting a person who is at least 163 centimeters but not more than 165 centimeters tall. Now we are dealing with an interval rather than a point value of our random variable.

We shall concern ourselves with computing probabilities for various intervals of continuous random variables such as  $P(a < X < b)$ ,  $P(W \geq c)$ , and so forth. Note that when  $X$  is continuous,

$$P(a < X \leq b) = P(a < X < b) + P(X = b) = P(a < X < b).$$

That is, it does not matter whether we include an endpoint of the interval or not. This is not true, though, when  $X$  is discrete.

Although the probability distribution of a continuous random variable cannot be presented in tabular form, it can be stated as a formula. Such a formula would necessarily be a function of the numerical values of the continuous random variable  $X$  and as such will be represented by the functional notation  $f(x)$ . In dealing with continuous variables,  $f(x)$  is usually called the **probability density function**, or simply the **density function**, of  $X$ . Since  $X$  is defined over a continuous sample space, it is possible for  $f(x)$  to have a finite number of discontinuities. However, most density functions that have practical applications in the analysis of statistical data are continuous and their graphs may take any of several forms, some of which are shown in Figure 3.4. Because areas will be used to represent probabilities and probabilities are positive numerical values, the density function must lie entirely above the  $x$  axis.

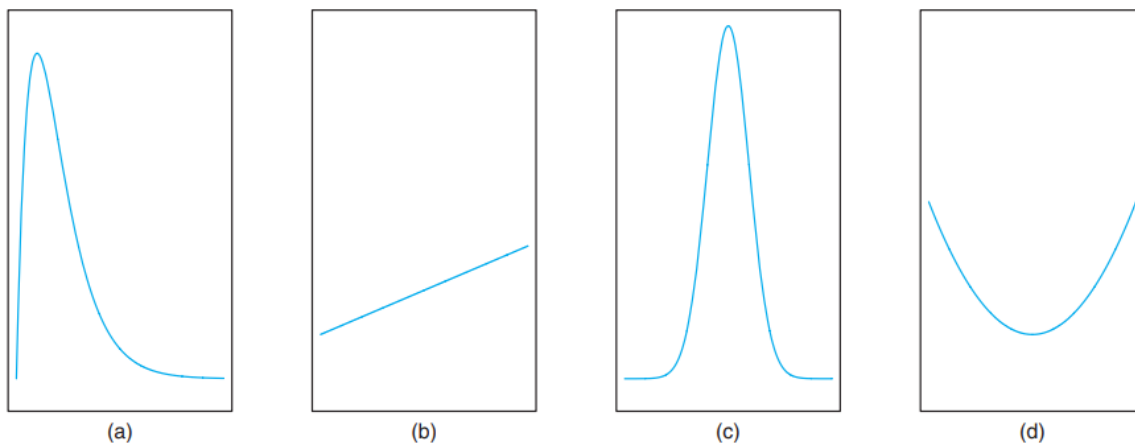


Figure 3.4: Typical density functions.

A probability density function is constructed so that the area under its curve bounded by the  $x$  axis is equal to 1 when computed over the range of  $X$  for which  $f(x)$  is defined. Should this range of  $X$  be a finite interval, it is always possible to extend the interval to include the entire set of real numbers by defining  $f(x)$  to be zero at all points in the extended portions of the interval. In Figure 3.5, the probability that  $X$  assumes a value between  $a$  and  $b$  is equal to the shaded area under the density function between the ordinates at  $x = a$  and  $x = b$ , and from integral calculus is given by

$$P(a < X < b) = \int_a^b f(x) dx.$$

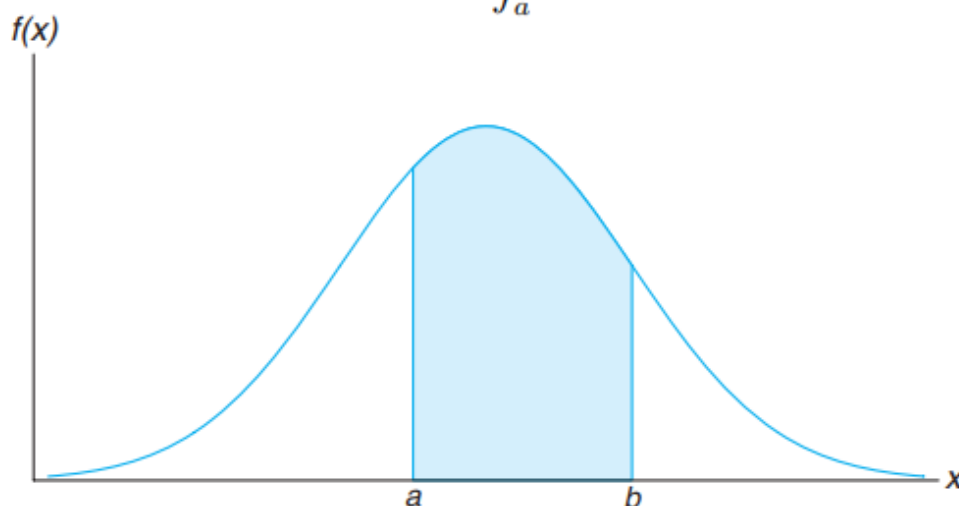


Figure 3.5:  $P(a < X < b)$ .

The function  $f(x)$  is a **probability density function** (pdf) for the continuous random variable  $X$ , defined over the set of real numbers, if

1.  $f(x) \geq 0$ , for all  $x \in R$ .
2.  $\int_{-\infty}^{\infty} f(x) dx = 1$ .
3.  $P(a < X < b) = \int_a^b f(x) dx$ .

The **cumulative distribution function**  $F(x)$  of a continuous random variable  $X$  with density function  $f(x)$  is

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt, \quad \text{for } -\infty < x < \infty.$$



### Example 3.11:

Suppose that the error in the reaction temperature, in  $^{\circ}\text{C}$ , for a controlled laboratory experiment is a continuous random variable  $X$  having the probability density function

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2, \\ 0, & \text{elsewhere.} \end{cases}$$

- (a) Verify that  $f(x)$  is a density function.  
(b) Find  $P(0 < X \leq 1)$ .

We use Definition 3.6.

- (a) Obviously,  $f(x) \geq 0$ . To verify condition 2 in Definition 3.6, we have

$$\int_{-\infty}^{\infty} f(x) dx = \int_{-1}^2 \frac{x^2}{3} dx = \frac{x^3}{9} \Big|_{-1}^2 = \frac{8}{9} + \frac{1}{9} = 1.$$

- (b) Using formula 3 in Definition 3.6, we obtain

$$P(0 < X \leq 1) = \int_0^1 \frac{x^2}{3} dx = \frac{x^3}{9} \Big|_0^1 = \frac{1}{9}.$$

For the density function of Example 3.11, find  $F(x)$ , and use it to evaluate  $P(0 < X \leq 1)$ .

For  $-1 < x < 2$ ,

$$F(x) = \int_{-\infty}^x f(t) dt = \int_{-1}^x \frac{t^2}{3} dt = \frac{t^3}{9} \Big|_{-1}^x = \frac{x^3 + 1}{9}.$$

Therefore,

$$F(x) = \begin{cases} 0, & x < -1, \\ \frac{x^3+1}{9}, & -1 \leq x < 2, \\ 1, & x \geq 2. \end{cases}$$

The cumulative distribution function  $F(x)$  is expressed in Figure 3.6. Now

$$P(0 < X \leq 1) = F(1) - F(0) = \frac{2}{9} - \frac{1}{9} = \frac{1}{9},$$

## Practice Problems

**Question:** According to a survey, 30% of adults are against using animals for research. Assume that this result holds true for the current population of all adults. Let  $x$  be the number of adults who are against using animals for research in a random sample of two adults. Obtain the probability distribution of  $x$ . Draw a tree diagram for this problem.

**3.5** Determine the value  $c$  so that each of the following functions can serve as a probability distribution of the discrete random variable  $X$ :

(a)  $f(x) = c(x^2 + 4)$ , for  $x = 0, 1, 2, 3$ ;

(b)  $f(x) = c \binom{2}{x} \binom{3}{3-x}$ , for  $x = 0, 1, 2$ .

**3.13** The probability distribution of  $X$ , the number of imperfections per 10 meters of a synthetic fabric in continuous rolls of uniform width, is given by

$x$	0	1	2	3	4
$f(x)$	0.41	0.37	0.16	0.05	0.01

Construct the cumulative distribution function of  $X$ .

**3.6** The shelf life, in days, for bottles of a certain prescribed medicine is a random variable having the density function

$$f(x) = \begin{cases} \frac{20,000}{(x+100)^3}, & x > 0, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the probability that a bottle of this medicine will have a shelf life of

- (a) at least 200 days;
- (b) anywhere from 80 to 120 days.

**3.9** The proportion of people who respond to a certain mail-order solicitation is a continuous random variable  $X$  that has the density function

$$f(x) = \begin{cases} \frac{2(x+2)}{5}, & 0 < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

- (a) Show that  $P(0 < X < 1) = 1$ .
- (b) Find the probability that more than  $1/4$  but fewer than  $1/2$  of the people contacted will respond to this type of solicitation.

**3.18** A continuous random variable  $X$  that can assume values between  $x = 2$  and  $x = 5$  has a density function given by  $f(x) = 2(1 + x)/27$ . Find

- (a)  $P(X < 4)$ ;
- (b)  $P(3 \leq X < 4)$ .

**3.21** Consider the density function

$$f(x) = \begin{cases} k\sqrt{x}, & 0 < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

- (a) Evaluate  $k$ .
- (b) Find  $F(x)$  and use it to evaluate

$$P(0.3 < X < 0.6).$$

**3.26** From a box containing 4 black balls and 2 green balls, 3 balls are drawn in succession, each ball being replaced in the box before the next draw is made. Find the probability distribution for the number of green balls.

**3.28** A cereal manufacturer is aware that the weight of the product in the box varies slightly from box to box. In fact, considerable historical data have allowed the determination of the density function that describes the probability structure for the weight (in ounces). Letting  $X$  be the random variable weight, in ounces, the density function can be described as

$$f(x) = \begin{cases} \frac{2}{5}, & 23.75 \leq x \leq 26.25, \\ 0, & \text{elsewhere.} \end{cases}$$

- (a) Verify that this is a valid density function.
- (b) Determine the probability that the weight is smaller than 24 ounces.
- (c) The company desires that the weight exceeding 26 ounces be an extremely rare occurrence. What is the probability that this rare occurrence does actually occur?