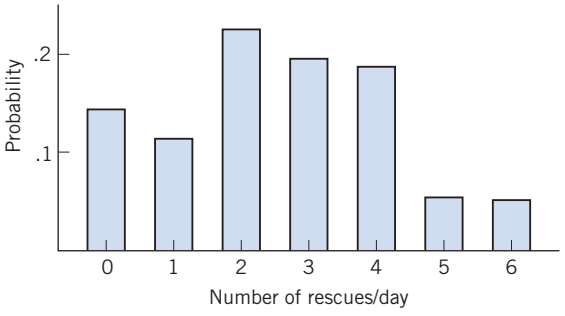
5

Probability Distributions

Rescue Service on a Lake

Student sailors and other boaters on Lake Mendota are protected by a boating rescue service. The relative frequencies from a long record for summer days lead to an approximate distribution of the number of rescues per day.



© John Terrence Turner/FPG International/ Getty Images

The distribution describes the randomness of daily rescue activity. For instance, on any given day, the most probable number of rescues is 2. The distribution can be the basis for decisions concerning manpower and the need for additional rescue boats.

1. INTRODUCTION

A prescription for the probability model of an experiment contains two basic ingredients: the sample space and the assignment of probability to each elementary outcome. In Chapter 4, we encountered several examples where the elementary outcomes had only qualitative descriptions rather than numerical values. For instance, with two tosses of a coin, the outcomes HH, HT, TH, and TT are pairs of letters that identify the occurrences of heads or tails. If a new vaccine is studied for the possible side effects of nausea, the response of each subject may be severe, moderate, or no feeling of nausea. These are qualitative outcomes rather than measurements on a numerical scale.

Often, the outcomes of an experiment are numerical values: for example, the daily number of burglaries in a city, the hourly wages of students on summer jobs, and scores on a college placement examination. Even in the former situation where the elementary outcomes are only qualitatively described, interest frequently centers on some related numerical aspects. If a new vaccine is tested on 100 individuals, the information relevant for an evaluation of the vaccine may be the numbers of responses in the categories—severe, moderate, or no nausea. The detailed record of 100 responses can be dispensed with once we have extracted this summary. Likewise, for an opinion poll conducted on 500 residents to determine support for a proposed city ordinance, the information of particular interest is how many residents are in favor of the ordinance, and how many are opposed. In these examples, the individual observations are not numerical, yet a numerical summary of a collection of observations forms the natural basis for drawing inferences. In this chapter, we concentrate on the numerical aspects of experimental outcomes.

2. RANDOM VARIABLES

Focusing our attention on the numerical features of the outcomes, we introduce the idea of a random variable.

> A **random variable** X associates a numerical value with each outcome of an experiment.

Corresponding to every elementary outcome of an experiment, a random variable assumes a numerical value, determined from some characteristic pertaining to the outcome. (In mathematical language, we say that a random variable X is a real-valued function defined on a sample space.) The word "random" serves as a reminder of the fact that, beforehand, we do not know the outcome of an experiment or its associated value of X.

Example 1 The Number of Heads as a Random Variable

Consider X to be the number of heads obtained in three tosses of a coin. List the numerical values of X and the corresponding elementary outcomes.

SOLUTION First, X is a variable since the number of heads in three tosses of a coin can have any of the values 0, 1, 2, or 3. Second, this variable is random in the sense that the value that would occur in a given instance cannot be predicted with certainty. We can, though, make a list of the elementary outcomes and the associated values of X.

Outcome	Value of X
HHH	3
HHT	2
HTH	2
HTT	1
THH	2
THT	1
TTH	1
TTT	0

Note that, for each elementary outcome there is only one value of X. However, several elementary outcomes may yield the same value. Scanning our list, we now identify the events (i.e., the collections of the elementary outcomes) that correspond to the distinct values of X.

Numerical Value of X as an Event	Composition of the Event
$[X = 0]$	$= \{TTT\}$
$[X = 1]$	$= \{HTT, THT, TTH\}$
$[X = 2]$	$= \{HHT, HTH, THH\}$
$[X = 3]$	$= \{HHH\}$

Guided by this example, we observe the following general facts.

The events corresponding to the distinct values of *X* are incompatible.

The union of these events is the entire sample space.

Typically, the possible values of a random variable X can be determined directly from the description of the random variable without listing the sample

space. However, to assign probabilities to these values, treated as events, it is sometimes helpful to refer to the sample space.

Example 2 A Random Variable That Is a Count with a Finite Maximum Value

Fifty cars are entered in a 100-mile road race. Let X be the number of cars that actually finish the race. Here X could conceivably take any of the values $0, 1, \ldots, 50$.

Example 3 A Random Variable That Is a Count with No Upper Limit

Once a week, a student buys a single lottery ticket. Let X be the number of tickets she purchases before she wins at least \$1000 on a ticket. The possible values of X are then $1, 2, 3, \ldots$, where the list never terminates.

A random variable is said to be **discrete** if it has either a finite number of values or infinitely many values that can be arranged in a sequence. All the preceding examples are of this type. On the other hand, if a random variable represents some measurement on a continuous scale and is therefore capable of assuming all values in an interval, it is called a **continuous** random variable. Of course, any measuring device has a limited accuracy and, therefore, a continuous scale must be interpreted as an abstraction. Some examples of continuous random variables are the height of an adult male, the daily milk yield of a holstein, and the survival time of a patient following a heart attack.

Probability distributions of discrete random variables are explored in this chapter. As we shall see, the developments stem directly from the concepts of probability introduced in Chapter 4. A somewhat different outlook is involved in the process of conceptualizing the distribution of a continuous random variable. Details for the continuous case are postponed until Chapter 6.

Exercises

5.1 Identify each of the following as a discrete or continuous random variable.

(a) Number of empty seats on a flight from Atlanta to London.

(b) Yearly low temperature in your city.

(c) Yearly maximum daily amount of ozone in Los Angeles.

(d) Time it takes for a plumber to fix a bathroom faucet.

(e) Number of cars ticketed for illegal parking on campus today.

5.2 Identify the variable as a discrete or a continuous random variable in parts (a) – (e).

(a) The loss of weight following a diet program.

(b) The magnitude of an earthquake as measured on the open-ended Richter scale.

(c) The seating capacity of an airplane.

(d) The number of cars sold at a dealership on one day.

(e) The percentage of fruit juice in a drink mix.

5.3 Two of the integers {1, 3, 5, 6, 7} are chosen at random without replacement. Let X denote the difference larger minus smaller number.

(a) List all choices and the corresponding values of X.

(b) List the distinct values of X and determine their probabilities.

5.4 The three finalists for an award are A, B, and C. They will be rated by two judges. Each judge assigns the ratings 1 for best, 2 for intermediate, and 3 for worst. Let X denote the total score for finalist A (the sum of the ratings received from the two judges).

(a) List all pairs of ratings that finalist A can receive.

(b) List the distinct values of X.

5.5 Refer to Exercise 5.4. Suppose instead there are two finalists A and B and four judges. Each judge assigns the ratings 1 for the best and 2 for the worst finalists.

(a) List all possible assignments of ratings to finalist A by the four judges.

(b) List the distinct values of X, the total score of A.

5.6 Two brands of beverages, B and M, are popular with students. The owner of one campus establishment will observe sales and, for each of three weekends, record which brand has the highest sales. List the possible outcomes, and for each outcome record the number of weekends X that the sales of B are highest. (Assume there are no ties.)

5.7 Each week a grocery shopper buys either canned (C) or bottled (B) soft drinks. The type of soft drink purchased in 3 consecutive weeks is to be recorded.

(a) List the sample space.

(b) If a different type of soft drink is purchased than in the previous week, we say that there is a switch. Let X denote the number of switches. Determine the value of X for each elementary outcome. (*Example:* For BBB, $X = 0$; for BCB, $X = 2$.)

5.8 A child psychologist interested in how friends are selected studies groups of three children. For one group, Ann, Barb, and Carol, each is asked which of the other two she likes best.

(a) Make a list of the outcomes. (Use A, B, and C to denote the three children.)

(b) Let X be the number of times Carol is chosen. List the values of X.

3. PROBABILITY DISTRIBUTION OF A DISCRETE RANDOM VARIABLE

The list of possible values of a random variable X makes us aware of all the eventualities of an experiment as far as the realization of X is concerned. By employing the concepts of probability, we can ascertain the chances of observing the various values. To this end, we introduce the notion of a probability distribution.

The **probability distribution** or, simply the **distribution,** of a discrete random variable X is a list of the distinct numerical values of X along with their associated probabilities.

Often, a formula can be used in place of a detailed list.

Example 4 The Probability Distribution for Tossing a Fair Coin

If X represents the number of heads obtained in three tosses of a fair coin, find the probability distribution of X.

SOLUTION In Example 1, we have already listed the eight elementary outcomes and the associated values of X. The distinct values of X are 0, 1, 2, and 3. We now calculate their probabilities.

The model of a fair coin entails that the eight elementary outcomes are equally likely, so each is assigned the probability $\frac{1}{8}$. The event $[X = 0]$ has the single outcome TTT, so its probability is $\frac{1}{8}$. Similarly, the probabilities of $[X = 1]$, $[X = 2]$, and $[X = 3]$ are found to be $\frac{3}{8}$, $\frac{3}{8}$, and $\frac{1}{8}$, respectively. Collecting these results, we obtain the probability distribution of X displayed in Table 1.

TABLE 1 The Probability Distribution of X, the Number of Heads in Three Tosses of a Coin

Value of X	Probability
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$
Total	1

For general discussion, we will use the notation x_1, x_2, and so on, to designate the distinct values of a random variable X. The probability that a particular value x_i occurs will be denoted by $f(x_i)$. As in Example 4, if X can take k possible values $x_1, \ldots, x_k$ with the corresponding probabilities $f(x_1), \ldots, f(x_k)$, the probability distribution of X can be displayed in the format of Table 2. Since the quantities $f(x_i)$ represent probabilities, they must all be numbers between 0 and 1. Furthermore, when summed over all possible values of X, these probabilities must add up to 1.

TABLE 2 Form of a Discrete Probability Distribution

Value of x	Probability $f(x)$
x_1	$f(x_1)$
x_2	$f(x_2)$
.	.
.	.
.	.
x_k	$f(x_k)$
Total	1

The **probability distribution** of a discrete random variable X is described as the function

$$f(x_i) = P[X = x_i]$$

which gives the probability for each value and satisfies:

1. $0 \le f(x_i) \le 1$ for each value x_i of X
2. $\sum_{i=1}^{k} f(x_i) = 1$

A probability distribution or the probability function describes the manner in which the total probability 1 gets apportioned to the individual values of the random variable.

A graphical presentation of a probability distribution helps reveal any pattern in the distribution of probabilities. Is there symmetry about some value or a long tail to one side? Is the distribution peaked with a few values having high probabilities or is it uniform?

We consider a display similar in form to a relative frequency histogram, discussed in Chapter 2. It will also facilitate the building of the concept of a continuous distribution. To draw a **probability histogram,** we first mark the values of X on the horizontal axis. With each value x_i as center, a vertical rectangle is drawn whose area equals the probability $f(x_i)$. The probability histogram for the distribution of Example 4 is shown in Figure 1.

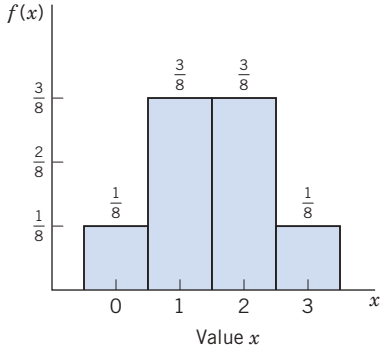


Figure 1 The probability histogram of X, the number of heads in three tosses of a coin.

Example 5 Probability Distribution for News Source Preference

Suppose 60% of the students at a large university prefer getting their daily news from the Internet as opposed to television. These are the only two choices. Four students are randomly selected. Let X be the number of students sampled who prefer news from the Internet. Obtain the probability distribution of X and plot the probability histogram.

SOLUTION Because each student will prefer either Internet (I) news or television (T), the number of elementary outcomes concerning a sample of four students is $2 \times 2 \times 2 \times 2 = 16$. These can be conveniently enumerated in the scheme of Example 8, Chapter 4, called a tree diagram. However, we list them here according to the count X.

$X = 0$	$X = 1$	$X = 2$	$X = 3$	$X = 4$
TTTT	TTTI	TTII	TIII	IIII
	TTIT	TITI	ITII	
	TITT	TIIT	IITI	
	ITTT	ITTI	IIIT	
		ITIT		
		IITT		

Our objective here is to calculate the probability of each value of X. To this end, we first reflect on the assignment of probabilities to the elementary outcomes. For one student selected at random, we obviously have $P(I) = .6$ and $P(T) = .4$ because 60% of the students prefer Internet news. Moreover, as the population is vast while the sample size is very small, the observations on four students can, for all practical purposes, be treated as independent. That is, knowledge that the first student selected prefers Internet news does not change the probability that the second will prefer Internet news and so on.

Invoking independence and the multiplication law of probability, we calculate $P(\text{TTTT}) = .4 \times .4 \times .4 \times .4 = .0256$ so $P(X = 0) = .0256$. The event $[X = 1]$ has four elementary outcomes, each containing three T's and one I. Since $P(\text{TTTI}) = (.4)^3 \times (.6) = .0384$ and the same result holds for each of these 4 elementary outcomes, we get $P[X = 1] = 4 \times .0384 = .1536$. In the same manner,

$$P[X = 2] = 6 \times (.4)^2 \times (.6)^2 = .3456$$

$$P[X = 3] = 4 \times (.4) \times (.6)^3 = .3456$$

$$P[X = 4] = (.6)^4 = .1296$$

Collecting these results, we obtain the probability distribution of X presented in Table 3 and the probability histogram plotted in Figure 2.

TABLE 3 The Probability Distribution of X in Example 5

x	$f(x)$
0	.0256
1	.1536
2	.3456
3	.3456
4	.1296
Total	1.0000

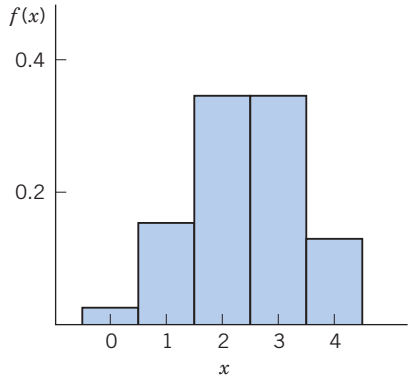


Figure 2 Probability histogram.

At this point, we digress briefly for an explanation of the role of probability distributions in statistical inference. To calculate the probabilities associated with the values of a random variable, we require a full knowledge of the uncertainties of the experimental outcomes. For instance, when X represents some numerical characteristic of a random sample from a population, we assume a known composition of the population in order that the distribution of X can be calculated numerically. In Example 5, the chances of observing the various values of X were calculated under the assumption that the proportion of all students who prefer Internet news was .6. Ordinarily, in practical applications, this population quantity would be unknown to us. Suppose the letter p stands for this unknown proportion of students who prefer Internet news. Statistical

inference attempts to determine the values of p that are deemed plausible in light of the value of X actually observed in a sample. To fix ideas, suppose all four of the students sampled prefer Internet news. Based on this observation, is .6 a plausible value of p? Table 3 shows that if p were indeed .6, the chance of observing the extreme value $X = 0$ is only .0256. This very low probability casts doubt on the hypothesis that $p = .6$. This kind of statistical reasoning will be explored further in later chapters.

The probability distributions in Examples 4 and 5 were obtained by first assigning probabilities to the elementary outcomes using a process of logical deduction. When this cannot be done, one must turn to an empirical determination of the distribution. This involves repeating the experiment a large number of times and using the relative frequencies of the various values of X as approximations of the corresponding probabilities.

Example 6 A Probability Distribution Based on an Empirical Study

Let X denote the number of magazines to which a college senior subscribes. From a survey of 400 college seniors, suppose the frequency distribution of Table 4 was observed. Approximate the probability distribution of X.

TABLE 4 Frequency Distribution of the Number X of Magazine Subscriptions

Magazine Subscriptions (x)	Frequency	Relative Frequency[a]
0	61	.15
1	153	.38
2	106	.27
3	56	.14
4	24	.06
Total	400	1.00

[a]Rounded to second decimal.

SOLUTION Viewing the relative frequencies as empirical estimates of the probabilities, we have essentially obtained an approximate determination of the probability distribution of X. The true probability distribution would emerge if a vast number (ideally, the entire population) of seniors were surveyed.

The reader should bear in mind an important distinction between a relative frequency distribution and the probability distribution. The former is a sample-based entity and is therefore susceptible to variation on different occasions of sampling. By contrast, the probability distribution is a stable entity that refers to

the entire population. It is a theoretical construct that serves as a model for describing the variation in the population.

The probability distribution of X can be used to calculate the probabilities of events defined in terms of X.

Example 7 Determining the Probability of an Event Determined from a Probability Distribution

Table 5 describes the number of homework assignments due next week for a randomly selected set of students taking at least 14 credits. Determine the probability that (a) X is equal to or larger than 2 and (b) X is less than or equal to 2.

TABLE 5 A Probability Distribution for Number of Homework Assignments Due Next Week

Value x	Probability $f(x)$
0	.02
1	.23
2	.40
3	.25
4	.10

SOLUTION (a) The event $[X \geq 2]$ is composed of $[X = 2]$, $[X = 3]$, and $[X = 4]$. Thus,

$$P[X \geq 2] = f(2) + f(3) + f(4)$$
$$= .40 + .25 + .10 = .75$$

(b) Similarly, we also calculate

$$P[X \leq 2] = f(0) + f(1) + f(2)$$
$$= .02 + .23 + .40 = .65$$

Exercises

5.9 Faced with a tight deadline on two major projects, you decide to hire two of the five available persons to help complete the work. They have 1, 2, 4, 2 and 1 years experience, respectively. Since their references are very similar, you decide to select two of these workers at random. Let X denote the sum of their years experience. Obtain the probability distribution of X.

5.10 Refer to Exercise 5.9 but let X denote the maximum years experience among the two persons selected.

(a) List all choices and the corresponding values of X.

(b) List the distinct values of X.

(c) Obtain the probability distribution of X.

5.11 Let the random variable X represent the sum of the points in two tosses of a die.

(a) List the possible values of X.

(b) For each value of X, list the corresponding elementary outcomes.

(c) Obtain the probability distribution of X.

5.12 Examine if the following are legitimate probability distributions.

(a)

x	$f(x)$
-1	.3
2	.5
7	.2
9	.1

(b)

x	$f(x)$
1	.2
3	.4
4	.3
6	.1

(c)

x	$f(x)$
-2	.25
0	.50
2	.25
4	0

(d)

x	$f(x)$
0	.3
1	$-.1$
2	.8

5.13 For each case, list the values of x and $f(x)$ and examine if the specification represents a probability distribution. If it does not, state what properties are violated.

(a) $f(x) = \frac{1}{6}(x - 1)$ for $x = 1, 2, 3, 4$

(b) $f(x) = \frac{1}{3}(x - 2)$ for $x = 1, 2, 3, 4$

(c) $f(x) = \frac{1}{20}(2x + 4)$ for $x = -2, -1, 0, 1, 2$

(d) $f(x) = \frac{8}{15}\frac{1}{2^x}$ for $x = 0, 1, 2, 3$

5.14 The probability distribution of X is given by the function

$$f(x) = \frac{1}{30}\binom{5}{x} \quad \text{for} \quad x = 1, 2, 3, 4$$

Find (a) $P[X = 3]$ (b) $P[X \text{ is even}]$.

5.15 Refer to Exercise 5.7. Suppose that for each purchase $P(B) = \frac{1}{2}$ and the decisions in different weeks are independent. Assign probabilities to the elementary outcomes and obtain the distribution of X.

5.16 Refer to Exercise 5.8. Assuming each choice is equally likely, determine the probability distribution of X.

5.17 Market researchers are concerned if people who view a commercial remember the product. They often make phone surveys two hours after a commercial is shown. Suppose that 20% of the people who watch one commercial will remember the product two hours later. Four persons are randomly selected from those who viewed the commercial. Let X denote the number in the sample who remember the product. Obtain the probability distribution of X.

5.18 New video games are rated, by editors, at various Web sites (e.g., www.gamespot.com). You are equally interested in five games that received editors' ratings of

10 10 10 9 9

on a ten point scale. Suppose you decide to randomly choose two games to purchase at this time. Let X denote the sum of the ratings on the two games purchased. List the possible values of X and determine the probability distribution.

5.19 Suppose, for a loaded die, the probabilities of the faces

1 2 3 4 5 6

are in the ratios 3 : 1 : 1 : 1 : 1 : 3. Let X denote the number appearing on a single roll of the die.

(a) Determine the probability distribution of X.

(b) What is the probability of getting an even number?

5.20 A surprise quiz contains three multiple-choice questions: Question 1 has four suggested answers, Question 2 has three, and Question 3 has two. A completely unprepared student decides to choose the answers at random. Let X denote the number of questions the student answers correctly.

(a) List the possible values of X.

(b) Find the probability distribution of X.

(*continued*)

(c) Find $P[\text{At least 1 correct}] = P[X \geq 1]$.

(d) Plot the probability histogram.

5.21 A probability distribution is partially given in the following table with the additional information that the even values of X are equally likely. Determine the missing entries in the table.

x	$f(x)$
1	.1
2	
3	0
4	
5	.3
6	

5.22 Consider the following setting of a random selection: A box contains 100 cards, of which 25 are numbered 1, 35 are numbered 2, 30 are numbered 3, 10 are numbered 4. One card will be drawn from the box and its number X observed. Give the probability distribution of X.

5.23 Two probability distributions are shown in the following tables. For each case, describe a specific setting of random selection (like the one given in Exercise 5.22) that yields the given probability distribution.

(a)

x	$f(x)$
2	.32
4	.44
6	.24

(b)

x	$f(x)$
−2	3/4
0	4/14
4	5/14
5	2/14

5.24 In a study of the life length of a species of mice, 120 newborn mice are observed. The numbers staying alive past the first, second, third, and fourth years are 106, 72, 25, and 0, respectively. Let X denote the life length (in discrete units of whole years) of this species of mice. Using these data, make an empirical determination of the probability distribution of X.

5.25 Use the approximate probability distribution in Example 6 to calculate

(a) $P[X \leq 3]$

(b) $P(X \geq 2]$

(c) $P[2 \leq X \leq 3]$

5.26 Of eight candidates seeking three positions at a counseling center, five have degrees in social science and three do not. If three candidates are selected at random, find the probability distribution of X, the number having social science degrees among the selected persons.

5.27 Based on recent records, the manager of a car painting center has determined the following probability distribution for the number of customers per day.

x	$f(x)$
0	.05
1	.20
2	.30
3	.25
4	.15
5	.05

(a) If the center has the capacity to serve two customers per day, what is the probability that one or more customers will be turned away on a given day?

(b) What is the probability that the center's capacity will not be fully utilized on a day?

(c) By how much must the capacity be increased so the probability of turning a customer away is no more than .10?

5.28 Among cable TV customers, let X denote the number of television sets in a single-family residential dwelling. From an examination of the subscription records of 361 residences in a city, the frequency distribution at the top of page 185 is obtained.

(a) Based on these data, obtain an approximate determination of the probability distribution of X.

(b) Why is this regarded as an approximation?
(c) Plot the probability histogram.

No. of Television Sets (x)	No. of Residences (Frequency)
1	64
2	161
3	89
4	47
Total	361

4. EXPECTATION (MEAN) AND STANDARD DEVIATION OF A PROBABILITY DISTRIBUTION

We will now introduce a numerical measure for the center of a probability distribution and another for its spread. In Chapter 2, we discussed the concepts of mean, as a measure of the center of a data set, and standard deviation, as a measure of spread. Because probability distributions are theoretical models in which the probabilities can be viewed as long-run relative frequencies, the sample measures of center and spread have their population counterparts.

To motivate their definitions, we first refer to the calculation of the mean of a data set. Suppose a die is tossed 20 times and the following data obtained.

4, 3, 4, 2, 5, 1, 6, 6, 5, 2
2, 6, 5, 4, 6, 2, 1, 6, 2, 4

The mean of these observations, called the sample mean, is calculated as

$$\bar{x} = \frac{\text{Sum of the observations}}{\text{Sample size}} = \frac{76}{20} = 3.8$$

Alternatively, we can first count the frequency of each point and use the relative frequencies to calculate the mean as

$$\bar{x} = 1\left(\frac{2}{20}\right) + 2\left(\frac{5}{20}\right) + 3\left(\frac{1}{20}\right) + 4\left(\frac{4}{20}\right) + 5\left(\frac{3}{20}\right) + 6\left(\frac{5}{20}\right) = 3.8$$

This second calculation illustrates the formula

$$\text{Sample mean } \bar{x} = \sum (\text{Value} \times \text{Relative frequency})$$

Rather than stopping with 20 tosses, if we imagine a very large number of tosses of a die, the relative frequencies will approach the probabilities, each of which is $\frac{1}{6}$ for a fair die. The mean of the (infinite) collection of tosses of a fair die should then be calculated as

$$1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + \cdots + 6\left(\frac{1}{6}\right) = \sum (\text{Value} \times \text{Probability}) = 3.5$$

Motivated by this example and the stability of long-run relative frequency, it is then natural to define the mean of a random variable X or its probability distribution as

$$\sum (\text{Value} \times \text{Probability}) \qquad \text{or} \qquad \sum x_i f(x_i)$$

where x_i's denote the distinct values of X. The mean of a probability distribution is also called the population mean for the variable X and is denoted by the Greek letter μ.

The mean of a random variable X is also called its **expected value** and, alternatively, denoted by $E(X)$. That is, the mean μ and expected value $E(X)$ are the same quantity and will be used interchangeably.

The **mean** of X or **population mean**

$$E(X) = \mu$$
$$= \sum (\text{Value} \times \text{Probability}) = \sum x_i f(x_i)$$

Here the sum extends over all the distinct values x_i of X.

Example 8 Calculating the Population Mean Number of Heads

With X denoting the number of heads in three tosses of a fair coin, calculate the mean of X.

SOLUTION The probability distribution of X was recorded in Table 1. From the calculations exhibited in Table 6 we find that the mean is 1.5.

The mean of a probability distribution has a physical interpretation. If a metal sheet is cut in the shape of the probability histogram, then μ represents the point on the base at which the sheet will balance. For instance, the mean $\mu = 1.5$ calculated in Example 8 is exactly at the center of mass for the distribution depicted in Figure 1. Because the amount of probability corresponds to the amount of mass in a bar, we interpret the balance point μ as the center of the probability distribution.

Like many concepts of probability, the idea of the mean or expectation originated from studies of gambling. When X refers to the financial gain in a game of chance, such as playing poker or participating in a state lottery, the name "expected gain" is more appealing than "mean gain." In the realm of statistics, both the names "mean" and "expected value" are widely used.

TABLE 6 Mean of the Distribution of Table 1

x	$f(x)$	$xf(x)$
0	$\frac{1}{8}$	0
1	$\frac{3}{8}$	$\frac{3}{8}$
2	$\frac{3}{8}$	$\frac{6}{8}$
3	$\frac{1}{8}$	$\frac{3}{8}$
Total	1	$\frac{12}{8} = 1.5 = \mu$

Example 9 Expected Value—Setting a Premium

A trip insurance policy pays \$2000 to the customer in case of a loss due to theft or damage on a five-day trip. If the risk of such a loss is assessed to be 1 in 200, what is the expected cost, per customer, to cover?

SOLUTION The probability that the company will be liable to pay \$2000 to a customer is $\frac{1}{200} = .005$. Therefore, the probability distribution of X, the payment per customer, is as follows.

Payment x	Probability $f(x)$
\$0	.995
\$2000	.005

We calculate

$$E(X) = 0 \times .995 + 2000 \times .005$$
$$= \$10.00$$

The company's expected cost per customer is \$10.00 and, therefore, a premium equal to this amount is viewed as the fair premium. If this premium is charged and no other costs are involved, then the company will neither make a profit nor lose money in the long run. In practice, the premium is set at a higher price because it must include administrative costs and intended profit.

No Casino Game Has a Positive Expected Profit

Each year, thousands of visitors come to casinos to gamble. Although all count on being lucky and a few indeed return with a smiling face, most leave the casino with a light purse. But, what should a gambler's expectation be?

Consider a simple bet on the red of a roulette wheel that has 18 red, 18 black, and 2 green slots. This bet is at even money so a \$10 wager on red has an expected profit of

$$E(\text{Profit}) = (10)\left(\frac{18}{38}\right) + (-10)\left(\frac{20}{38}\right) = -.526$$

The negative expected profit says we expect to lose an average of 52.6¢ on every \$10 bet. Over a long series of bets, the relative frequency of winning will approach the probability $\frac{18}{38}$ and that of losing will approach $\frac{20}{38}$, so a player will lose a substantial amount of money.

Other bets against the house have a similar negative expected profit. How else could a casino stay in business?

The concept of expected value also leads to a numerical measure for the spread of a probability distribution—namely, the standard deviation. When we define the standard deviation of a probability distribution, the reasoning parallels that for the standard deviation discussed in Chapter 2.

Because the mean μ is the center of the distribution of X, we express variation of X in terms of the deviation $X - \mu$. We define the variance of X as the expected value of the squared deviation $(X - \mu)^2$. To calculate this expected value, we note that

$(X - \mu)^2$ Takes Value	With Probability
$(x_1 - \mu)^2$	$f(x_1)$
$(x_2 - \mu)^2$	$f(x_2)$
.	.
.	.
.	.
$(x_k - \mu)^2$	$f(x_k)$

The expected value of $(X - \mu)^2$ is obtained by multiplying each value $(x_i - \mu)^2$ by the probability $f(x_i)$ and then summing these products. This motivates the definition:

$$\begin{aligned}\text{Variance of } X &= \sum (\text{Deviation})^2 \times (\text{Probability}) \\ &= \sum (x_i - \mu)^2 f(x_i)\end{aligned}$$

The variance of X is abbreviated as $\text{Var}(X)$ and is also denoted by σ^2. The **standard deviation** of X is the positive square root of the variance and is denoted by $\text{sd}(X)$ or σ (a Greek lower-case sigma.)

The variance of X is also called the **population variance** and σ denotes the **population standard deviation.**

Variance and Standard Deviation of *X*

$$\sigma^2 = \text{Var}(X) = \sum (x_i - \mu)^2 f(x_i)$$

$$\sigma = \text{sd}(X) = +\sqrt{\text{Var}(X)}$$

Example 10 Calculating a Population Variance and Standard Deviation

Calculate the variance and the standard deviation of the distribution of X that appears in the left two columns of Table 7.

SOLUTION We calculate the mean μ, the deviations $x - \mu$, $(x - \mu)^2$, and finally $(x - \mu)^2 f(x)$. The details are shown in Table 7.

TABLE 7 Calculation of Variance and Standard Deviation

x	$f(x)$	$xf(x)$	$(x - \mu)$	$(x - \mu)^2$	$(x - \mu)^2 f(x)$
0	.1	0	−2	4	.4
1	.2	.2	−1	1	.2
2	.4	.8	0	0	0
3	.2	.6	1	1	.2
4	.1	.4	2	4	.4
Total	1.0	$2.0 = \mu$			$1.2 = \sigma^2$

$$\text{Var}(X) = \sigma^2 = 1.2$$
$$\text{sd}(X) = \sigma = \sqrt{1.2} = 1.095$$

An alternative formula for σ^2 often simplifies the numerical work (see Appendix A2.2).

Alternative Formula for Hand Calculation

$$\sigma^2 = \sum x_i^2 f(x_i) - \mu^2$$

Example 11 Alternative Calculation of Variance

We illustrate the alternative formula for σ^2 using the probability distribution in Example 10. See Table 8.

TABLE 8 Calculation of Variance by the Alternative Formula

x	$f(x)$	$xf(x)$	$x^2 f(x)$
0	.1	.0	.0
1	.2	.2	.2
2	.4	.8	1.6
3	.2	.6	1.8
4	.1	.4	1.6
Total	1.0	$2.0 = \mu$	$5.2 = \sum x^2 f(x)$

$$\begin{aligned}\sigma^2 &= 5.2 - (2.0)^2 \\ &= 1.2 \\ \sigma &= \sqrt{1.2} = 1.095\end{aligned}$$

The standard deviation σ, rather than σ^2, is the appropriate measure of spread. Its unit is the same as that of X. For instance, if X refers to income in dollars, σ will have the unit dollar, whereas σ^2 has the rather artificial unit (dollar)^2.

Exercises

5.29 Given the following probability distribution concerning Web sites visited almost every day:

(a) Construct the probability histogram.

(b) Find $E(X)$, σ^2, and σ.

x	$f(x)$
1	.1
2	.2
3	.3
4	.4

5.30 A wait person proposed a distribution for the number of meals served on a two-for-one deal.

x	$f(x)$
2	.3
4	.5
6	.1
8	.1

Find the mean and standard deviation.

5.31 In bidding for a remodeling project, a carpenter determines that he will have a net profit of $5000 if he gets the contract and a net loss of $86 if his bid fails. If the probability of his getting the contract is .2, calculate his expected return.

5.32 A book club announces a sweepstakes in order to attract new subscribers. The prizes and the corresponding chances are listed here (typically, the prizes are listed in bold print in an advertisement flyer while the chances are entered in fine print or not mentioned at all).

Prize	Chance
$50,000	1 in one million
$ 5,000	1 in 250,000
$ 100	1 in 5,000
$ 20	1 in 500

Suppose you have just mailed in a sweepstakes ticket and X stands for your winnings.

(a) List the probability distribution of X. (*Caution:* What is not listed is the chance of winning nothing, but you can figure that out from the given information.)

(b) Calculate your expected winnings.

5.33 Calculate the mean and standard deviation for the probability distribution of Example 5.

5.34 Referring to Exercise 5.27, find the mean and standard deviation of the number of customers.

5.35 A construction company submits bids for two projects. Listed here are the profit and the probability of winning each project.

Assume that the outcomes of the two bids are independent.

	Profit	Chance of Winning Bid
Project *A*	$ 75,000	.50
Project *B*	$120,000	.65

(a) List the possible outcomes (win/not win) for the two projects and find their probabilities.

(b) Let X denote the company's total profit out of the two contracts. Determine the probability distribution of X.

(c) If it costs the company $2000 for preparatory surveys and paperwork for the two bids, what is the expected net profit?

5.36 Refer to Exercise 5.35, but suppose that the projects are scheduled consecutively with *A* in the first year and *B* in the second year. The company's chance of winning project *A* is still .50. Instead of the assumption of independence, now assume that if the company wins project *A*, its chance of winning *B* becomes .80 due to a boost of its image, whereas its chance drops to .40 in case it fails to win *A*. Under this premise, do parts (a–c).

5.37 Upon examination of the claims records of 280 policy holders over a period of five years, an insurance company makes an empirical determination of the probability distribution of X = number of claims in five years.

(a) Calculate the expected value of X.

(b) Calculate the standard deviation of X.

x	$f(x)$
0	.315
1	.289
2	.201
3	.114
4	.063
5	.012
6	.006

5.38 Suppose the probability distribution of a random variable X is given by the function

$$f(x) = \frac{12}{25} \cdot \frac{1}{x} \quad \text{for} \quad x = 1, 2, 3, 4$$

Calculate the mean and standard deviation of this distribution.

5.39 The probability distribution of a random variable X is given by the function

$$f(x) = \frac{1}{84}\binom{5}{x}\binom{4}{3-x} \quad \text{for} \quad x = 0, 1, 2, 3$$

(a) Calculate the numerical probabilities and list the distribution.

(b) Calculate the mean and standard deviation of X.

*5.40 Given here are the probability distributions of two random variables X and Y.

x	$f(x)$
1	.1
2	.3
3	.4
4	.2

y	$f(y)$
0	.2
2	.4
4	.3
6	.1

(a) From the X distribution, determine the distribution of the random variable $8 - 2X$ and verify that it coincides with the Y distribution. (Hence, identify $Y = 8 - 2X$.)

(b) Calculate the mean and standard deviation of X (call these μ_X and σ_X, respectively).

(c) From the Y distribution, calculate the mean and standard deviation of Y (call these μ_Y and σ_Y, respectively).

(d) If $Y = a + bX$, then according to theory, we must have the relations $\mu_Y = a + b\mu_X$ and $\sigma_Y = |b|\sigma_X$. Verify these relations from your results in parts (b) and (c).

5.41 A salesman of small-business computer systems will contact four customers during a week. Each contact can result in either a sale, with probability .3, or no sale, with probability .7.

Assume that customer contacts are independent.

(a) List the elementary outcomes and assign probabilities.

(b) If X denotes the number of computer systems sold during the week, obtain the probability distribution of X.

(c) Calculate the expected value of X.

5.42 Refer to Exercise 5.41. Suppose the computer systems are priced at \$2000, and let Y denote the salesman's total sales (in dollars) during a week.

(a) Give the probability distribution of Y.

(b) Calculate $E(Y)$ and see that it is the same as $2000 \times E(X)$.

5.43 Definition: The **median** of a distribution is the value m_0 of the random variable such that $P[X \leq m_0] \geq .5$ and $P[X \geq m_0] \geq .5$. In other words, the probability at or below m_0 is at least .5, and the probability at or above m_0 is at least .5. Find the median of the distribution given in Exercise 5.29.

5.44 Given the two probability distributions

x	$f(x)$
1	.2
2	.6
3	.2

y	$f(y)$
0	.1
1	.2
2	.4
3	.2
4	.1

(a) Construct probability histograms. Which distribution has a larger spread?

(b) Verify that both distributions have the same mean.

(c) Compare the two standard deviations.

5. SUCCESSES AND FAILURES—BERNOULLI TRIALS

Often, an experiment can have only two possible outcomes. Example 5 concerned individual students who either preferred Internet or television news. The proportion of the population that preferred Internet news was .60. Also, only two outcomes are possible for a single trial in the scenarios of Examples 1 and 2. In all these circumstances, a simple probability model can be developed for the chance variation in the outcomes. Moreover, the population proportion need not be known as in the previous examples. Instead, the probability distribution will involve this unknown population proportion as a **parameter.**

Sampling situations where the elements of a population have a dichotomy abound in virtually all walks of life. A few examples are:

Inspect a specified number of items coming off a production line and count the number of defectives.

Survey a sample of voters and observe how many favor an increase of public spending on welfare.

Analyze the blood specimens of a number of rodents and count how many carry a particular viral infection.

Examine the case histories of a number of births and count how many involved delivery by Cesarean section.

Selecting a single element of the population is envisioned as a trial of the (sampling) experiment, so that each trial can result in one of two possible outcomes.

Boy or girl?

A model for the potential sex of a newborn is the assignment of probability to each of the two outcomes. For most applications, .5 is assigned to "male" but extensive official statistics establish that the probability is actually about .52. Blaine Harrington/Photolibrary Group Limited.

Our ultimate goal is to develop a probability model for the number of outcomes in one category when repeated trials are performed.

An organization of the key terminologies, concerning the successive repetitions of an experiment, is now in order. We call each repetition by the simpler name—a **trial.** Furthermore, the two possible outcomes of a trial are now assigned the technical names **success** (S) and **failure** (F) just to emphasize the point that they are the only two possible results. These names bear no connotation of success or failure in real life. Customarily, the outcome of primary interest in a study is labeled success (even if it is a disastrous event). In a study of the rate of unemployment, the status of being unemployed may be attributed the statistical name success!

Further conditions on the repeated trials are necessary in order to arrive at our intended probability distribution. Repeated trials that obey these conditions are called **Bernoulli trials** after the Swiss mathematician Jacob Bernoulli.

Perhaps the simplest example of Bernoulli trials is the prototype model of tossing a coin, where the occurrences *head* and *tail* can be labeled S and F, respectively. For a fair coin, we assign probability $p = \frac{1}{2}$ to success and $q = \frac{1}{2}$ to failure.

Bernoulli Trials

1. Each trial yields one of two outcomes, technically called success (S) and failure (F).
2. For each trial, the probability of success $P(S)$ is the same and is denoted by $p = P(S)$. The probability of failure is then $P(F) = 1 - p$ for each trial and is denoted by q, so that $p + q = 1$.
3. Trials are independent. The probability of success in a trial remains unchanged given the outcomes of all the other trials.

Example 12 Sampling from a Population with Two Categories of Elements

Consider a lot (population) of items in which each item can be classified as either defective or nondefective. Suppose that a lot consists of 15 items, of which 5 are defective and 10 are nondefective.

Do the conditions for Bernoulli trials apply when sampling (1) with replacement and (2) without replacement?

SOLUTION

1. ***Sampling with replacement.*** An item is drawn at random (i.e., in a manner that all items in the lot are equally likely to be selected). The quality of the item is recorded and it is returned to the lot before the next drawing. The conditions for Bernoulli trials are satisfied. If the occurrence of a defective is labeled S, we have $P(S) = \frac{5}{15}$.
2. ***Sampling without replacement.*** In situation (2), suppose that 3 items are drawn one at a time but without replacement. Then the condition concerning the independence of trials is violated. For the first drawing, $P(S) = \frac{5}{15}$. If the first draw produces S, the lot then consists of 14 items, 4 of which are defective. Given this information about the result of the first draw, the conditional probability of obtaining an S on the second draw is then $\frac{4}{14} \neq \frac{5}{15}$, which establishes the lack of independence.

This violation of the condition of independence loses its thrust when the population is vast and only a small fraction of it is sampled. Consider sampling 3 items without replacement from a lot of 1500 items, 500 of which are defective. With S_1 denoting the occurrence of an S in the first draw and S_2 that in the second, we have

$$P(S_1) = \frac{500}{1500} = \frac{5}{15}$$

and

$$P(S_2|S_1) = \frac{499}{1499}$$

For most practical purposes, the latter fraction can be approximated by $\frac{5}{15}$. Strictly speaking, there has been a violation of the independence of trials, but it is to such a negligible extent that the model of Bernoulli trials can be assumed as a good approximation.

Example 12 illustrates the important points:

If elements are sampled from a dichotomous population at random and with replacement, the conditions for Bernoulli trials are satisfied.

When the sampling is made without replacement, the condition of the independence of trials is violated. However, if the population is large and only a small fraction of it (less than 10%, as a rule of thumb) is sampled, the effect of this violation is negligible and the model of the Bernoulli trials can be taken as a good approximation.

Example 13 further illustrates the kinds of approximations that are sometimes employed when using the model of the Bernoulli trials.

Example 13 Testing a New Antibiotic—Bernoulli Trials?

Suppose that a newly developed antibiotic is to be tried on 10 patients who have a certain disease and the possible outcomes in each case are cure (S) or no cure (F).

Comment on the applicability of the Bernoulli trial model.

SOLUTION Each patient has a distinct physical condition and genetic constitution that cannot be perfectly matched by any other patient. Therefore, strictly speaking, it may not be possible to regard the trials made on 10 different patients as 10 repetitions of an experiment under identical conditions, as the definition of Bernoulli trials demands. We must remember that the conditions of a probability model are abstractions that help to realistically simplify the complex mechanism governing the outcomes of an experiment. Identification with Bernoulli trials in such situations is to be viewed as an approximation of the real world, and its merit rests on how successfully the model explains chance variations in the outcomes.

Exercises

5.45 Is the model of Bernoulli trials plausible in each of the following situations? Discuss in what manner (if any) a serious violation of the assumptions can occur.

(a) Seven friends go to a blockbuster movie and each is asked whether the movie was excellent.

(b) A musical aptitude test is given to 10 students and the times to complete the test are recorded.

(c) Items coming off an assembly line are inspected and classified as defective or nondefective.

(d) Going house by house down the block and recording if the newspaper was delivered on time.

5.46 In each case, examine whether or not repetitions of the stated experiment conform to the model of Bernoulli trials. Where the model is appropriate, determine the numerical value of p or indicate how it can be determined.

(a) Roll a fair die and observe the number that shows up.

(b) Roll a fair die and observe whether or not the number 6 shows up.

(c) Roll two fair dice and observe the total of the points that show up.

(d) Roll two fair dice and observe whether or not both show the same number.

(e) Roll a loaded die and observe whether or not the number 6 shows up.

5.47 A jar contains 25 candies of which 6 are brown, 12 are yellow, and 7 are of other colors. Consider 4 successive draws of 1 candy at random from the jar and suppose the appearance of a yellow candy is the event of interest. For each of the following situations, state whether or not the model of Bernoulli trials is reasonable, and if so, determine the numerical value of p.

(a) After each draw, the selected candy is returned to the jar.

(b) After each draw, the selected candy is not returned to the jar.

(c) After each draw, the selected candy is returned to the jar and one new candy of the same color is added in the jar.

5.48 Refer to Exercise 5.47 and suppose instead that the mix consists of 2500 candies, of which 600 are brown, 1200 are yellow, and 700 are of other colors. Repeat parts (a–c) of Exercise 5.47 in this setting.

5.49 From four agricultural plots, two will be selected at random for a pesticide treatment. The other two plots will serve as controls. For each plot, denote by S the event that it is treated with the pesticide. Consider the assignment of treatment or control to a single plot as a trial.

(a) Is $P(\mathrm{S})$ the same for all trials? If so, what is the numerical value of $P(\mathrm{S})$?

(b) Are the trials independent? Why or why not?

5.50 Refer to Exercise 5.49. Now suppose for each plot a fair coin will be tossed. If a head shows up, the plot will be treated; otherwise, it will be a control. With this manner of treatment allocation, answer parts (a) and (b).

5.51 A market researcher intends to study the consumer preference between regular and decaffeinated coffee. Examine the plausibility of the model of Bernoulli trials in the following situations.

(a) One hundred consumers are randomly selected and each is asked to report the types of coffee (regular or decaffeinated) purchased in the five most recent occasions. If we consider each purchase as a trial, this inquiry deals with 500 trials.

(b) Five hundred consumers are randomly selected and each is asked about the most recent purchase of coffee. Here again the inquiry deals with 500 trials.

5.52 A backpacking party carries three emergency signal flares, each of which will light with a

probability of .98. Assuming that the flares operate independently, find:

(a) The probability that at least one flare lights.

(b) The probability that exactly two flares light.

5.53 Consider Bernoulli trials with success probability $p = .3$.

(a) Find the probability that four trials result in all failures.

(b) Given that the first four trials result in all failures, what is the conditional probability that the next four trials are all successes?

(c) Find the probability that the first success occurs in the fourth trial.

5.54 If in three Bernoulli trials $P[\text{All three are successes}] = .064$, what is the probability that all three are failures?

5.55 According to the U. S. Census Bureau, in 2007 about 10% of persons between 25 and 30 years old live alone. Let S be the event a person lives alone. If five persons in that age group are randomly selected,

(a) Find the probability of the sequence *SFFSF*.

(b) Find the probability of exactly 2 S's.

5.56 A graphic designer makes a presentation to clients and this results in sales of her services in one-fourth of the cases. Assuming the results for different clients are independent

(a) Find the probability that exactly 3 of the next 4 presentations will result in sales.

(b) Find the probability that none of the presentations result in a sale.

5.57 An animal either dies (D) or survives (S) in the course of a surgical experiment. The experiment is to be performed first with two animals. If both survive, no further trials are to be made. If exactly one animal survives, one more animal is to undergo the experiment. Finally, if both animals die, two additional animals are to be tried.

(a) List the sample space.

(b) Assume that the trials are independent and the probability of survival in each trial is $\frac{1}{4}$. Assign probabilities to the elementary outcomes.

(c) Let X denote the number of survivors. Obtain the probability distribution of X by referring to part (b).

5.58 The accompanying table shows the percentages of residents in a large community when classified according to gender and presence of a particular allergy.

	Allergy	
	Present	Absent
Male	16	36
Female	9	39

Suppose that the selection of a person is considered a trial and the presence of the allergy is considered a success. For each case, identify the numerical value of p and find the required probability.

(a) Four persons are selected at random. What is the probability that none has the allergy?

(b) Four males are selected at random. What is the probability that none has the allergy?

(c) Two males and two females are selected at random. What is the probability that none has the allergy?

6. THE BINOMIAL DISTRIBUTION

This section deals with a basic distribution that models chance variation in repetitions of an experiment that has only two possible outcomes. The random variable X of interest is the frequency count of one of the categories. Previously, its distribution was calculated under the assumption that the population proportion is known. For instance, the probability distribution of Table 3, from Example 5, resulted from the specification that 60% of the population of students

prefer news from the Internet. In a practical situation, however, the population proportion is usually an unknown quantity. When this is so, the probability distribution of X cannot be numerically determined. However, we will see that it is possible to construct a model for the probability distribution of X that contains the unknown population proportion as a **parameter.** The probability model serves as the major vehicle of drawing inferences about the population from observations of the random variable X.

A **probability model** is an assumed form of the probability distribution that describes the chance behavior for a random variable X.

Probabilities are expressed in terms of relevant population quantities, called the **parameters.**

Consider a **fixed number** n of Bernoulli trials with the success probability p in each trial. The number of successes obtained in n trials is a random variable that we denote by X. The probability distribution of this random variable X is called a binomial distribution.

The binomial distribution depends on the two quantities n and p. For instance, the distribution appearing in Table 1 is precisely the binomial distribution with $n = 3$ and $p = .5$, whereas that in Table 3 is the binomial distribution with $n = 4$ and $p = .6$.

The Binomial Distribution

Denote

n = a fixed number of Bernoulli trials
p = the probability of success in each trial
X = the (random) number of successes in n trials

The random variable X is called a **binomial random variable.** Its distribution is called a **binomial distribution.**

A review of the developments in Example 5 will help motivate a formula for the general binomial distribution.

Example 14 Example 5 Revisited—An Example of the Binomial Distribution

The random variable X represents the number of students who prefer news from the Internet among a random sample of $n = 4$ students from a

large university. Instead of the numerical value .6, we now denote the population proportion of students who prefer Internet news by the symbol p. Furthermore, we relabel the outcome "Internet" as a success (S) and "not Internet" as a failure (F). The elementary outcomes of sampling 4 students, the associated probabilities, and the value of X are listed as follows.

FFFF	SFFF	SSFF	SSSF	SSSS
	FSFF	SFSF	SSFS	
	FFSF	SFFS	SFSS	
	FFFS	FSSF	FSSS	
		FSFS		
		FFSS		

Value of X	0	1	2	3	4
Probability of each outcome	q^4	pq^3	p^2q^2	p^3q	p^4
Number of outcomes	1 $= \binom{4}{0}$	4 $= \binom{4}{1}$	6 $= \binom{4}{2}$	4 $= \binom{4}{3}$	1 $= \binom{4}{4}$

Because the population of students at a large university is vast, the trials can be treated as independent. Also, for an individual trial, $P(\mathrm{S}) = p$ and $P(\mathrm{F}) = q = 1 - p$. The event $[X = 0]$ has one outcome, FFFF, whose probability is

$$P[X = 0] = P(\mathrm{FFFF}) = q \times q \times q \times q = q^4$$

To arrive at an expression for $P[X = 1]$, we consider the outcomes listed in the second column. The probability of SFFF is

$$P(\mathrm{SFFF}) = p \times q \times q \times q = pq^3$$

and the same result holds for every outcome in this column. There are 4 outcomes so we obtain $P[X = 1] = 4pq^3$. The factor 4 is the number of outcomes with one S and three F's. Even without making a complete list of the outcomes, we can obtain this count. Every outcome has 4 places and the 1 place where S occurs can be selected from the total of 4 in $\binom{4}{1} = 4$ ways, while the remaining 3 places must be filled with an F. Continuing in the same line of reasoning, we see that the value $X = 2$ occurs with $\binom{4}{2} = 6$ outcomes, each of which has a probability of p^2q^2. Therefore $P[X = 2] = \binom{4}{2}p^2q^2$. After we work out the remaining terms, the binomial distribution with $n = 4$ trials can be presented as in Table 9.

TABLE 9 Binomial Distribution with $n = 4$ Trials

Value x	0	1	2	3	4
Probability $f(x)$	$\binom{4}{0} p^0 q^4$	$\binom{4}{1} p^1 q^3$	$\binom{4}{2} p^2 q^2$	$\binom{4}{3} p^3 q^1$	$\binom{4}{4} p^4 q^0$

It would be instructive for the reader to verify that the numerical probabilities appearing in Table 3 are obtained by substituting $p = .6$ and $q = .4$ in the entries of Table 9.

Extending the reasoning of Example 14 to the case of a general number n of Bernoulli trials, we observe that there are $\binom{n}{x}$ outcomes that have exactly x successes and $n - x$ failures. The probability of every such outcome is $p^x q^{n-x}$. Therefore,

$$f(x) = P[X = x] = \binom{n}{x} p^x q^{n-x} \qquad \text{for} \qquad x = 0, 1, \ldots, n$$

is the formula for the binomial probability distribution with n trials.

The **binomial distribution** with n trials and success probability p is described by the function

$$f(x) = P[X = x] = \binom{n}{x} p^x (1 - p)^{n-x}$$

for the possible values $x = 0, 1, \ldots, n$.

Example 15 The Binomial Distribution and Genetics

According to the Mendelian theory of inherited characteristics, a cross fertilization of related species of red- and white-flowered plants produces a generation whose offspring contain 25% red-flowered plants. Suppose that a horticulturist wishes to cross 5 pairs of the cross-fertilized species. Of the resulting 5 offspring, what is the probability that:

(a) There will be no red-flowered plants?

(b) There will be 4 or more red-flowered plants?

SOLUTION Because the trials are conducted on different parent plants, it is natural to assume that they are independent. Let the random variable X denote the number

of red-flowered plants among the 5 offspring. If we identify the occurrence of a red as a success S, the Mendelian theory specifies that $P(S) = p = \frac{1}{4}$, and hence X has a binomial distribution with $n = 5$ and $p = .25$. The required probabilities are therefore

$$\text{(a) } P[X = 0] = f(0) = (.75)^5 = .237$$

$$\text{(b) } P[X \geq 4] = f(4) + f(5) = \binom{5}{4}(.25)^4(.75)^1 + \binom{5}{5}(.25)^5(.75)^0$$
$$= .015 + .001 = .016$$

To illustrate the manner in which the values of p influence the shape of the binomial distribution, the probability histograms for three binomial distributions with $n = 6$ and p values of .5, .3, and .7, respectively, are presented in Figure 3. When $p = .5$, the binomial distribution is symmetric with the highest probability occurring at the center (see Figure 3*a*).

For values of p smaller than .5, more probability is shifted toward the smaller values of x and the distribution has a longer tail to the right. Figure 3*b*, where the binomial histogram for $p = .3$ is plotted, illustrates this tendency. On the other hand, Figure 3*c* with $p = .7$ illustrates the opposite tendency: The value of p is higher than .5, more probability mass is shifted toward higher values of x, and the distribution has a longer tail to the left. Considering the histograms in Figures 3*b* and 3*c*, we note that the value of p in one histogram is the same as the value of q in the other. The probabilities in one histogram are exactly the same as those in the other, but their order is reversed. This illustrates a general property of the binomial distribution: When p and q are interchanged, the distribution of probabilities is reversed.

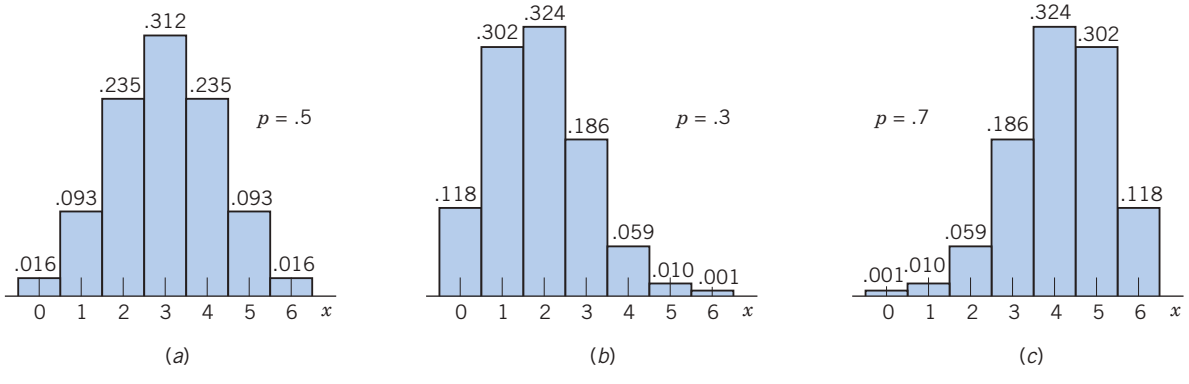


Figure 3 Binomial distributions for $n = 6$.

How to Use the Binomial Table (Appendix B, Table 2)

Although the binomial distribution is easily evaluated on a computer and some hand calculators, we provide a short table in Appendix B, Table 2. It covers selected sample sizes n ranging from 1 to 25 and several values of p. For a given pair (n, p), the table entry corresponding to each c represents the cumulative probability $P[X \le c] = \sum_{x=0}^{c} f(x)$, as is explained in the following scheme.

The Binomial Distribution

Value x	Probability $f(x)$
0	$f(0)$
1	$f(1)$
2	$f(2)$
.	.
.	.
.	.
n	$f(n)$
Total	1

Appendix B, Table 2 provides

c	Table Entry $\sum_{x=0}^{c} f(x) = P[X \le c]$
0	$f(0)$
1	$f(0) + f(1)$
2	$f(0) + f(1) + f(2)$
.	.
.	.
.	.
n	1.000

The probability of an individual value x can be obtained from this table by a subtraction of two consecutive entries. For example,

$$P[X = 2] = f(2) = \begin{pmatrix}\text{table entry at}\\ c = 2\end{pmatrix} - \begin{pmatrix}\text{table entry at}\\ c = 1\end{pmatrix}$$

Example 16 Binomial Distribution for the Number Cured

Suppose it is known that a new treatment is successful in curing a muscular pain in 50% of the cases. If it is tried on 15 patients, find the probability that:

(a) At most 6 will be cured.

(b) The number cured will be no fewer than 6 and no more than 10.

(c) Twelve or more will be cured.

SOLUTION Designating the cure of a patient by S and assuming that the results for individual patients are independent, we note that the binomial distribution with $n = 15$ and $p = .5$ is appropriate for $X =$ number of patients who are cured. To compute the required probabilities, we consult the binomial table for $n = 15$ and $p = .5$.

(a) $P[X \leq 6] = .304$, which is directly obtained by reading from the row $c = 6$.

(b) We are to calculate

$$P[6 \leq X \leq 10] = f(6) + f(7) + f(8) + f(9) + f(10)$$
$$= \sum_{x=6}^{10} f(x)$$

The table entry corresponding to $c = 10$ gives

$$P[X \leq 10] = \sum_{x=0}^{10} f(x) = .941$$

and the entry corresponding to $c = 5$ yields

$$P[X \leq 5] = \sum_{x=0}^{5} f(x) = .151$$

Because their difference represents the sum $\sum_{x=6}^{10} f(x)$, we obtain

$$\begin{aligned} P[6 \leq X \leq 10] &= P[X \leq 10] - P[X \leq 5] \\ &= .941 - .151 \\ &= .790 \end{aligned}$$

(c) To find $P[X \geq 12]$, we use the law of complement:

$$\begin{aligned} P[X \geq 12] &= 1 - P[X \leq 11] \\ &= 1 - .982 \\ &= .018 \end{aligned}$$

Note that $[X < 12]$ is the same event as $[X \leq 11]$.

(*An Aside:* Refer to our "muscular pain" example in Section 1 of Chapter 4. The mystery surrounding the numerical probability .018 is now resolved.)

The Mean and Standard Deviation of the Binomial Distribution

Although we already have a general formula that gives the binomial probabilities for any n and p, in later chapters we will need to know the mean and the standard deviation of the binomial distribution. The expression np for the mean is apparent from the following intuitive reasoning: If a fair coin is tossed 100 times, the expected number of heads is $100 \times \frac{1}{2} = 50$. Likewise, if the probability of an event is p, then in n trials the event is expected to happen np times. The formula for the standard deviation requires some mathematical derivation, which we omit.

The binomial distribution with n trials and success probability p has

$$\text{Mean} = np$$
$$\text{Variance} = npq \quad (\text{Recall: } q = 1 - p)$$
$$\text{sd} = \sqrt{npq}$$

Example 17 Calculating the Population Mean and Standard Deviation of a Binomial Distribution

For the binomial distribution with $n = 3$ and $p = .5$, calculate the mean and the standard deviation.

SOLUTION Employing the formulas, we obtain

$$\text{Mean} = np = 3 \times .5 = 1.5$$
$$\text{sd} = \sqrt{npq} = \sqrt{3 \times .5 \times .5} = \sqrt{.75} = .866$$

The mean agrees with the results of Example 8. The reader may wish to check the standard deviation by numerical calculations using the definition of σ.

Exercises

5.59 For each situation, state whether or not a binomial distribution holds for the random variable X. Also, identify the numerical values of n and p when a binomial distribution holds.

(a) A fair die is rolled 10 times, and X denotes the number of times 6 shows up.

(b) A fair die is rolled until 6 appears, and X denotes the number of rolls.

(c) In a jar, there are ten marbles, of which four are numbered 1, three are numbered 2, two are numbered 3, and one is numbered 4. Three marbles are drawn at random, one after another and with replacement, and X denotes the count of the selected marbles that are numbered either 1 or 2.

(d) The same experiment as described in part (c), but now X denotes the sum of the numbers on the selected marbles.

5.60 Construct a tree diagram for three Bernoulli trials. Attach probabilities in terms of p and q to each outcome and then table the binomial distribution for $n = 3$.

5.61 In each case, find the probability of x successes in n Bernoulli trials with success probability p for each trial.

(a) $x = 2 \quad n = 3 \quad p = .35$

(b) $x = 3 \quad n = 6 \quad p = .25$

(c) $x = 2 \quad n = 6 \quad p = .75$

5.62 (a) Plot the probability histograms for the binomial distributions for $n = 5$ and p equal to .2, .5, and .8.

(b) Locate the means.

(c) Find $P[X \geq 4]$ for each of the three cases.

5.63 An interior designer makes a presentation to potential clients and this results in sales of her services in 35% of the cases. Let X denote the

number of sales in the next four presentations. Assuming the results for different clients are independent, calculate the probabilities $f(x) = P[X = x]$ for $x = 0, 1, \ldots, 5$ and find

(a) $P[X \leq 3]$

(b) $P[X \geq 3]$

(c) $P[X = 2 \text{ or } 4]$

5.64 Refer to Exercise 5.63. What is the most probable value of X (called the **mode** of a distribution)?

5.65 About 75% of dog owners buy holiday presents for their dogs.[1] Suppose $n = 4$ dog owners are randomly selected. Find the probability that

(a) three or more buy their dog holiday presents.

(b) at most three buy their dog holiday presents

(c) Find the expected number of persons, in the sample, who buy their dog holiday presents.

5.66 According to a recent survey, outside of their own family members, 26% of adult Americans have no close friend to confide in. If this is the prevailing probability today, find the probability that in a random sample of $n = 5$ adults

(a) two or more have no close friend.

(b) at most two have no close friend.

(c) Find the expected number of persons who have no close friend.

5.67 Suppose 15% of the trees in a forest have severe leaf damage from air pollution. If 5 trees are selected at random, find the probability that:

(a) Three of the selected trees have severe leaf damage.

(b) No more than two have severe leaf damage.

5.68 Rh-positive blood appears in 85% of the white population in the United States. If 8 people are sampled at random from that population, find the probability that:

(a) At least 6 of them have Rh-positive blood.

(b) At most 3 of them have Rh-negative blood, that is, an absence of Rh positive.

5.69 Using the binomial table, find the probability of:

(a) Four successes in 13 trials when $p = .3$.

(b) Eight failures in 13 trials when $p = .7$.

(c) Eight successes in 13 trials when $p = .3$.

Explain why you get identical answers in parts (b) and (c).

5.70 According to the U.S. Census Bureau, in 2007 about 10% of persons between 25 and 30 years old live alone. For a random sample of size n, use the binomial table to find the probability of

(a) 1 or fewer persons living alone when $n = 12$.

(b) 2 or more persons living alone when $n = 12$.

(c) Find the expected number when $n = 12$.

(d) 1 or fewer persons living alone when $n = 20$.

5.71 About 30% of adults say that reading is a favorite leisure activity. Let success be the outcome that reading is a favorite leisure activity. Find the probability that

(a) More than 5 trials are needed in order to obtain 3 successes. (*Hint:* In other words, the event is: at most 2 successes in 5 trials.)

(b) More than 9 trials are needed in order to obtain 5 successes.

5.72 A survey report states that 70% of adult women visit their doctors for a physical examination at least once in two years. If 20 adult women are randomly selected, find the probability that

(a) Fewer than 14 of them have had a physical examination in the past two years.

(b) At least 17 of them have had a physical examination in the past two years.

[1] 75% is between the two 2008 survey results obtained by the American Kennel Club and Harris Interactive Poll.

5.73 Calculate the mean and standard deviation of the binomial distribution using the formulas in mean = np sd = $\sqrt{np(1-p)}$.

(a) Exercise 5.65 if n is changed to 20.

(b) Exercise 5.70 when $n = 20$.

(c) Exercise 5.71 when $n = 40$.

5.74 (a) For the binomial distribution with $n = 3$ and $p = .6$, list the probability distribution $(x, f(x))$ in a table.

(b) From this table, calculate the mean and standard deviation by using the methods of Section 4.

(c) Check your results with the formulas mean $= np$, sd $= \sqrt{npq}$.

5.75 Suppose that 20% of the college seniors support an increase in federal funding for care of the elderly. If 20 college seniors are randomly selected, what is the probability that at most 3 of them support the increased funding?

5.76 Referring to Exercise 5.75, find:

(a) The expected number of college seniors, in a random sample of 20, supporting the increased funding.

(b) The probability that the number of sampled college seniors supporting the increased funding equals the expected number.

5.77 According to a recent report of the American Medical Association, 9.0% of practicing physicians are in the specialty area of family practice. Assuming that the same rate prevails, find the mean and standard deviation of the number of physicians specializing in family practice out of a current random selection of 545 medical graduates.

5.78 According to the Mendelian theory of inheritance of genes, offspring of a dihybrid cross of peas could be any of the four types: round-yellow (RY), wrinkled-yellow (WY), round-green (RG) and wrinkled-green (WG), and their probabilities are in the ratio 9:3:3:1.

(a) If X denotes the number of RY offspring from 130 such crosses, find the mean and standard deviation of X.

(b) If Y denotes the number of WG offspring from 85 such crosses, find the mean and standard deviation of Y.

5.79 The following table (see Exercise 5.58) shows the percentages of residents in a large community when classified according to gender and presence of a particular allergy. For each part below, find the mean and standard deviation of the specified random variable.

	Allergy	
	Present	Absent
Male	16	36
Female	9	39

(a) X stands for the number of persons having the allergy in a random sample of 40 persons.

(b) Y stands for the number of males having the allergy in a random sample of 40 males.

(c) Z stands for the number of females not having the allergy in a random sample of 40 females.

The Following Exercise Requires a Computer

5.80 Many computer packages produce binomial probabilities. We illustrate the MINITAB commands for obtaining the binomial probabilities with $n = 5$ and $p = .33$. The probabilities $P[X = x]$ are obtained by first setting 0, 1, 2, 3, 4, 5 in *C1* and then selecting:

Calc > Probability distributions.

Type 5 in **Number of trials**

and .33 in **Probability.**

Enter *C1* in **Input constant.** Click **OK.**

which produces the output

Probability Density Function

```
Binomial with n = 5 and p = 0.33

x    P(X=x)
0      0.135013
1      0.332493
2      0.327531
3      0.161321
4      0.039728
5      0.003914
```

To obtain the cumulative probabilities $P(X \leq x)$, click **Cumulative probability** instead of **Probability.** The resulting output is

Cumulative Distribution Function

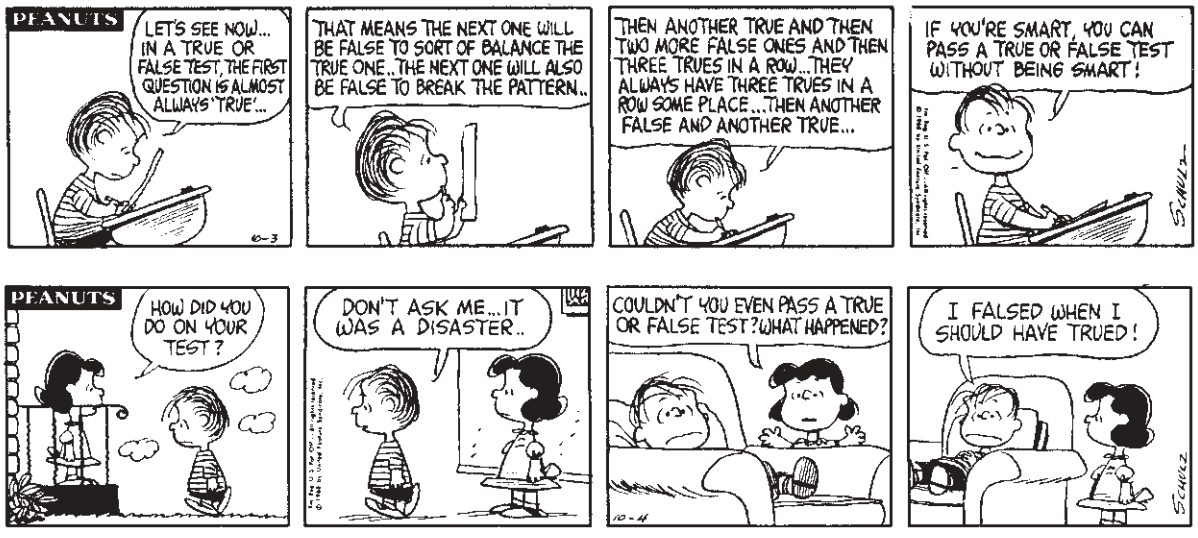
```
Binomial with n = 5 and p = 0.33

x    P(X<=x)
0      0.13501
1      0.46751
2      0.79504
3      0.95636
4      0.99609
5      1.00000
```

Using the computer, calculate

(a) $P[X \leq 8]$ and $P[X = 8]$ when $p = .67$ and $n = 12$

(b) $P[10 \leq X \leq 15]$ when $p = .43$ and $n = 35$



Poor Linus. Chance did not even favor him with half correct. Reprinted by permission of United Features Syndicate, Inc. © 1968.

7. THE BINOMIAL DISTRIBUTION IN CONTEXT

Requests for credit cards must be processed to determine if the applicant meets certain financial standards. In many instances, such as when the applicant already has a long-term good credit record, only a short review is required.

Usually this consists of a credit score assigned on the basis of the answers to questions on the application and then a computerized check of credit records. Many other cases require a full review with manual checks of information to determine the credit worthiness of the applicant.

Each week, a large financial institution selects a sample of 20 incoming applications and counts the number requiring full review. From data collected over several weeks, it is observed that about 40% of the applications require full review. If we take this long-run relative frequency as the probability, what is an unusually large number of full reviews and what is an unusually small number?

Let X be the number in the sample that require a full review. From the binomial table, with $n = 20$ and $p = .4$, we get

$$P[X \leq 3] = .016$$
$$P[X \geq 13] = 1 - P[X \leq 12] = 1 - .979 = .021$$

Taken together, the probability of X being 3 or less or 13 or more is .037, so those values should occur less than four times in 100 samples. That is, they could be considered unusual. In Exercise 5.81, you will be asked to show that either including 4 or including 12 will lead to a combined probability greater than .05. That is, the large and small values should then occur more than 1 in 20 times. For many people, this would be too frequent to be considered rare or unusual.

For the count X, we expect $np = 20 \times .4 = 8$ applications in the sample to require a full review. The standard deviation of this count is $\sqrt{20(.4)(1 - .4)} = 2.191$. Alternatively, when n is moderate or large, we could describe as unusual two or more standard deviations from the mean. A value at least two standard deviations, or $2(2.191) = 4.382$, above the mean of 8 must be 13 or more. A value 2 or more standard deviations below the mean must be 3 or less. These values correspond exactly to the values above that which we called unusual. In other cases, the two standard deviations approach provides a reasonable and widely used approximation.

p Charts for Trend

A series of sample proportions should be visually inspected for trend. A graph called a ***p* chart** helps identify times when the population proportion has changed from its long-time stable value. Because many sample proportions will be graphed, it is customary to set control limits at 3 rather than 2 standard deviations.

When p_0 is the expected or long-run proportion, we obtain a lower control limit by dividing the lower limit for X, $np_0 - 3\sqrt{np_0(1 - p_0)}$, by the sample size. Doing the same with the upper bound, we obtain an upper control limit.

Lower control limit

$$p_0 - 3\sqrt{\frac{p_0(1 - p_0)}{n}}$$

Upper control limit

$$p_0 + 3\sqrt{\frac{p_0(1 - p_0)}{n}}$$

In the context of credit applications that require a full review, $.4 = p_0$, so the control limits are

$$p_0 - 3\sqrt{\frac{p_0(1 - p_0)}{n}} = .4 - 3\sqrt{\frac{.4(1 - .4)}{20}} = .4 - 3(.110) = .07$$

$$p_0 + 3\sqrt{\frac{p_0(1 - p_0)}{n}} = .4 + 3\sqrt{\frac{.4(1 - .4)}{20}} = .4 + 3(.110) = .73$$

The **centerline** is drawn as a solid line at the expected or long-run proportion $p_0 = .4$, and the two control limits each at a distance of three standard deviations of the sample proportion from the centerline are also drawn as horizontal lines as in Figure 4. Sample proportions that fall outside of the control limits are considered unusual and should result in a search for a cause that may include a change in the mix of type of persons requesting credit cards.

The number of applications requiring full review out of the 20 in the sample were recorded for 19 weeks:

11 7 8 4 9 10 4 8 8 7 10 6 9 10 7 7 6 9 10

After converting to sample proportions by dividing by 20, we can graph the points in a p chart as in Figure 4. All the points are in control, and the financial institution does not appear to have reached the point where the mix of applicants includes more marginal cases.

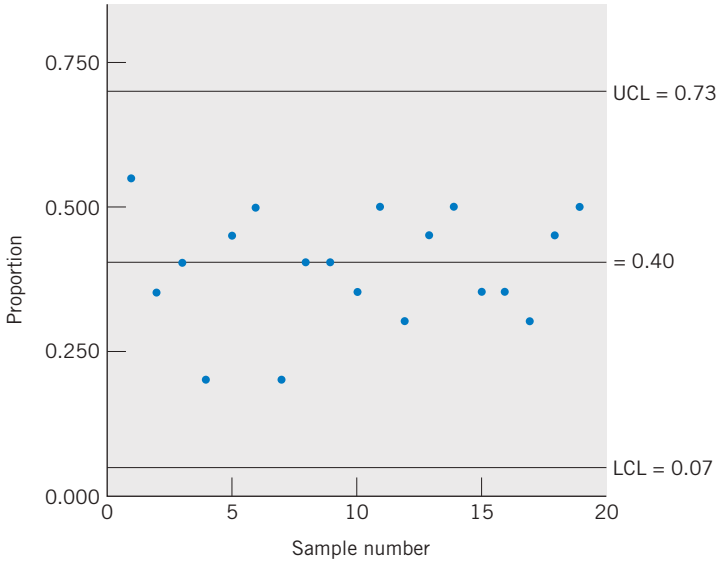


Figure 4 A p chart for the proportion of applications requiring a full review.

Exercises

5.81 Refer to the credit card application approval process on page 209 where unusual values are defined.

(a) Show that if 4 is included as an unusual value, then the probability $P[X \leq 4 \text{ or } X \geq 13]$ is greater than .05.

(b) Show that if 12 is included as an unusual value, then the probability $P[X \leq 3 \text{ or } X \geq 12]$ is greater than .05.

5.82 Refer to the credit card application approval process on page 209.

(a) Make a p chart using the centerline and control limits calculated for $p_0 = .4$.

(b) Suppose the next five weeks bring 12, 10, 15, 11, and 16 applications requiring full review. Graph the corresponding proportions on your p chart.

(c) Identify any weeks where the chart signals "out of control."

5.83 Several fast food restaurants advertise quarter-pound hamburgers. This could be interpreted as meaning half the hamburgers made have an uncooked weight of at least a quarter-pound and half have a weight that is less. An inspector checks 20 uncooked hamburgers at each restaurant.

(a) Make a p chart using the centerline and control limits calculated for $p_0 = .5$.

(b) Suppose that five restaurants have 8, 11, 7, 15, and 10 underweight hamburgers in samples of size 20. Graph the corresponding proportions on your p chart.

(c) Identify any restaurants where the chart signals "out of control."

5.84 Refer to Exercise 5.83.

(a) What are the unusual values for the number of underweight hamburgers in the sample if they correspond to proportions outside of the control limits of the p chart?

(b) Use the binomial table to find the probability of observing one of these unusual values.

5.85 Refer to Exercise 5.83. A syndicated newspaper story reported that inspectors found 22 of 24 hamburgers underweight at restaurant W and fined that restaurant. Draw new control limits on your chart, from Exercise 5.83, for one new sample of size 24. Plot the new proportion and determine if this point is "out of control."

USING STATISTICS WISELY

1. The assignment of a value to each possible outcome, which creates a random variable, should quantify an important feature of the outcomes.
2. Describe the chance behavior of a discrete random variable X by its probability distribution

$$f(x) = P[X = x] \qquad \text{for each possible value } x$$

3. Summarize a probability distribution, or the random variable, by its

$$\text{Mean:} \quad \mu = \sum_{\text{all } x} x \cdot f(x)$$

$$\text{Variance:} \quad \sigma^2 = \sum_{\text{all } x} (x - \mu)^2 \cdot f(x)$$

4. If the use of Bernoullie trials is reasonable, probabilities concerning the number of successes in n Bernoulli trials can be calculated using the formula for the binomial distribution

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

having mean np and variance $np(1 - p)$.

5. Never use the formula $np(1 - p)$ for the variance of a count of successes in n trials without first checking that the conditions for Bernoulli trials, independent trials with the same probability of success for each trial, hold. If the conditions are satisfied, then the binomial distribution is appropriate.

KEY IDEAS AND FORMULAS

The outcomes of an experiment are quantified by assigning each of them a numerical value related to a characteristic of interest. The rule for assigning the numerical value is called a **random variable** X.

A random variable having a finite number of values, or a sequence of values like a count, is called **discrete.** If a random variable can take any value in an interval, it is called a **continuous** random variable.

The **probability distribution** of X, or simply **distribution,** describes the manner in which probability is distributed over the possible values of X. Specifically, it is a list or formula giving the pairs x and $f(x) = P[X = x]$.

A probability distribution serves as a model for explaining variation in a population.

A **probability histogram** graphically displays a discrete distribution using bars whose area equals the probability.

A probability distribution has a

Mean $\mu = \sum (\text{Value} \times \text{Probability}) = \sum x f(x)$

which is interpreted as the **population mean.** This quantity is also called the **expected value** $E(X)$. Although X is a variable, $E(X)$ is a constant.

The **population variance** is

$$\sigma^2 = E(X - \mu)^2 = \sum (x - \mu)^2 f(x)$$

The **standard deviation** σ is the positive square root of variance. The standard deviation is a measure of the spread or variation of the population.

Bernoulli trials are defined by the characteristics: (1) two possible outcomes, **success** (S) or **failure** (F) for each trial; (2) a constant probability of success; and (3) independence of trials.

Sampling from a finite population without replacement violates the requirement of independence. If the population is large and the sample size small, the trials can be treated as independent for all practical purposes.

The number of successes X in a fixed number of Bernoulli trials is called a **binomial random variable.** Its probability distribution, called the **binomial distribution,** is given by

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

where n = number of trials, p = probability of success in each trial, and $q = 1 - p$.

The **binomial distribution** has

$$\text{Mean} = np$$
$$\text{Standard deviation} = \sqrt{npq}$$

A **probability model** is an assumed model for the probability distribution of a random variable. Probabilities are expressed in terms of **parameters** which are features of the population. For example, the binomial distribution is a probability model and the proportion p is a parameter.

A ***p* chart** displays sample proportions to reveal trends or changes in the population proportion over time.

TECHNOLOGY

Calculating the binomial probabilities $P[X = x]$ and $P[X \leq x]$

MINITAB

Calculating the binomial probability $P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$

The following commands illustrate the calculation of $P[X = 14]$ under the binomial distribution having $n = 21$ and $p = .57$:

Dialog box:

Calc > Probability distributions > Binomial.
Select **Probability.**
Type *21* in **Number of Trials** and *.57* in **Probability of success.**
Select **Input constant**, and enter *14*. Click **OK.**

Alternatively to get all of the binomial probabilities for $n = 21$, you can first

Enter 0, 1, 2, . . ., 21 in *C1*.

Calc > Probability distributions > Binomial.
Select **Probability.**
Type *21* in **Number of Trials** and *.57* in **Probability of success.**
Then, instead of clicking **Input constant**,
Enter *C1* **in Input column and** *C2* in **Optional storage.** Click **OK.**

Calculating the cumulative binomial probability

$$P[X \le c] = \sum_{x=0}^{c} \binom{n}{x} p^x (1-p)^{n-x}$$

Follow the same steps as in the calculation of the probability of single terms except:

Select **Cumulative probability** instead of **Probability.**
Calc > Probability distributions > Binomial.
Select **Cumulative Probability.**
Type *21* in **Number of trials** and *.57* in **Probability of success.**
Select **Input constant** and enter *14*. Click **OK.**

EXCEL

Calculating the binomial probability $P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$

We illustrate the calculation of $P[X = 14]$ under the binomial distribution having $n = 21$ and $p = .57$.

With the cursor in a blank cell, select the f_x icon, or select **Insert** and then **Function.** Choose **Statistical** and then **BINOMDIST.** Click **OK.**
Type *14* in **Number_s,** *21* in **Trials,** *.57* in **Probability_s,** and *False* in **Cumulative.** Click **OK.**

Calculating the cumulative binomial probability

$$P[X \le c] = \sum_{x=0}^{c} \binom{n}{x} p^x (1-p)^{n-x}$$

Follow the same steps as in the calculation of $P[X = x]$ except:

Type *True* in **Cumulative.**

TI-84/-83 PLUS

Calculating the binomial probability $P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$

The following commands illustrate the calculation of $P[X = 14]$ under the binomial distribution having $n = 21$ and $p = .57$:

Press **2ndVARS** to read the *probability distribution* menu.
Select[2] **0:binompdf(** and press **ENTER.**
With **binompdf**(on the home screen, type *21*, *.57*, *14*) to give:

binompdf(21, .57, 14)

Press **ENTER.**

[2]The options **O** and **A** are different for some TI-84 calculators.

Calculating the cumulative binomial probability

$$P[X \le c] = \sum_{x=0}^{c} \binom{n}{x} p^x (1-p)^{n-x}$$

Follow the same steps as in the calculation of $P[X = x]$ except replace the second step by

Select **A:binompdf(** and press **ENTER.**

8. REVIEW EXERCISES

5.86 Let X denote the difference (no. of heads − no. of tails) in three tosses of a coin.

(a) List the possible values of X.

(b) List the elementary outcomes associated with each value of X.

5.87 A large science department at the university made a list of its top 20 juniors and top 30 seniors. The first list contains 8 females and the second 10 females. One person is randomly selected from each list and the two selections are independent. Let X denote the number of females selected.

(a) For each possible value of X, identify the elementary outcomes associated with that value.

(b) Determine the probability distribution of X.

*5.88 Refer to Exercise 8.87 but now suppose the sampling is done in two stages: First a list is selected at random and then, from that list, two persons are selected at random without replacement. Let Y denote the number of females in the sample.

(a) List the elementary outcomes concerning the possible selections of this and the possible compositions of the sample (use a tree diagram). Find their probabilities. (*Hint:* Use conditional probability and the multiplication rule.)

(b) Determine the probability distribution of Y.

5.89 Refer to Exercise 5.87, but now suppose that the contents of the two lists are pooled together into a single larger list. Then two persons are drawn at random and without replacement. Let W denote the number of females in the sample. Obtain the probability distribution of W.

*5.90 In a tennis championship, player A competes against player B in consecutive sets, and the game continues until one player wins three sets. Assume that, for each set, $P(\text{A wins}) = .4$, $P(\text{B wins}) = .6$, and the outcomes of different sets are independent. Let X stand for the number of sets played.

(a) List the possible values of X and identify the elementary outcomes associated with each value.

(b) Obtain the probability distribution of X.

5.91 A list of the world's 10 largest companies, in terms of value, contains 5 from the United States, 3 from China, and 1 each from Germany and The Netherlands. A potential investor randomly selects 3 of the companies to research further.

(a) Find the probability distribution of X, the number of United States companies selected.

(b) Determine the mean and variance of X.

5.92 Refer to the monthly intersection accident data in Exercise 2.4. Considering an even longer record leads to a distribution for X = number of accidents in a month.

Value x	Probability $f(x)$
0	.08
1	.20
2	.19
3	.24
4	.14
5	.13
6	.02

(a) Calculate $E(X)$.
(b) Calculate $\text{sd}(X)$.
(c) Draw the probability histogram and locate the mean.

5.93 The following distribution has been proposed for the number of times a student will eat a gourmet restaurant dinner next week.

x	$f(x)$
0	.3
1	.4
2	.3

(a) Calculate the mean and variance.
(b) Plot the probability histogram and locate μ.

5.94 Refer to Exercise 5.92.
(a) List the x values that lie in the interval $\mu - \sigma$ to $\mu + \sigma$ and calculate $P[\mu - \sigma \le X \le \mu + \sigma]$.
(b) List the x values that lie in the interval $\mu - 2\sigma$ to $\mu + 2\sigma$ and calculate $P[\mu - 2\sigma \le X \le \mu + 2\sigma]$.

5.95 A student buys a lottery ticket for \$1. For every 1000 tickets sold, two bicycles are to be given away in a drawing.
(a) What is the probability that the student will win a bicycle?
(b) If each bicycle is worth \$200, determine the student's expected gain.

5.96 In the finals of a tennis match, the winner will get \$60,000 and the loser \$15,000. Find the expected winnings of player B if (a) the two finalists are evenly matched and (b) player B has probability .8 of winning.

5.97 The number of overnight emergency calls X to the answering service of a heating and air conditioning firm have the probabilities .05, .1, .15, .35, .20, and .15 for 0, 1, 2, 3, 4, and 5 calls, respectively.
(a) Find the probability of fewer than 3 calls.
(b) Determine $E(X)$ and $\text{sd}(X)$.

5.98 Suppose the number of parking tickets X issued during a police officer's shift has the probability distribution

x	0	1	2	3	4
$f(x)$	0.13	0.14	0.43	0.20	0.10

(a) Find the mean and standard deviation of the number of parking tickets issued.
(b) Let $A = [X \le 2]$ and $B = [X \ge 1]$. Find $P(A|B) = P(X \le 2 \mid X \ge 1)$.
(c) Suppose the numbers of tickets issued on different days are independent. What is the probability that, over the next five days, no parking tickets will be issued on exactly one of the days?

5.99 A botany student is asked to match the popular names of three house plants with their obscure botanical names. Suppose the student never heard of these names and is trying to match by sheer guess. Let X denote the number of correct matches.
(a) Obtain the probability distribution of X.
(b) What is the expected number of matches?

5.100 The number of days, X, that it takes the post office to deliver a letter between City A and City B has the probability distribution

x	$f(x)$
3	.5
4	.3
5	.2

Find:

(a) The expected number of days.

(b) The standard deviation of the number of days.

5.101 A roulette wheel has 38 slots, of which 18 are red, 18 black, and 2 green. A gambler will play three times, each time betting \$5 on red. The gambler gets \$10 if red occurs and loses the bet otherwise. Let X denote the net gain of the gambler in 3 plays (for instance, if he loses all three times, then $X = -15$).

(a) Obtain the probability distribution of X.

(b) Calculate the expected value of X.

(c) Will the expected net gain be different if the gambler alternates his bets between red and black? Why or why not?

5.102 Suppose that X can take the values 0, 1, 2, 3, and 4, and the probability distribution of X is incompletely specified by the function

$$f(x) = \frac{1}{4}\left(\frac{3}{4}\right)^x \qquad \text{for} \qquad x = 0, 1, 2, 3$$

Find (a) $f(4)$ (b) $P[X \geq 2]$ (c) $E(X)$ and (d) $\text{sd}(X)$.

5.103 ***The cumulative probabilities for a distribution.*** A probability distribution can also be described by a function that gives the accumulated probability at or below each value of X. Specifically,

$$F(c) = P[X \leq c] = \sum_{x \leq c} f(x)$$

Cumulative distribution function at c = Sum of probabilities of all values $x \leq c$

For the probability distribution given here, we calculate

x	$f(x)$	$F(x)$
1	.07	.07
2	.12	.19
3	.25	
4	.28	
5	.18	
6	.10	

$$F(1) = P[X \leq 1] = f(1) = .07$$
$$F(2) = P[X \leq 2] = f(1) + f(2) = .19$$

(a) Complete the $F(x)$ column in this table.

(b) Now cover the $f(x)$ column with a strip of paper. From the $F(x)$ values, reconstruct the probability function $f(x)$. [*Hint:* $f(x) = F(x) - F(x-1)$.]

*5.104 ***Runs.*** In a row of six plants two are infected with a leaf disease and four are healthy. If we restrict attention to the portion of the sample space for exactly two infected plants, the model of randomness (or lack of contagion) assumes that any two positions, for the infected plants in the row are as likely as any other.

(a) Using the symbols I for infected and H for healthy, list all possible occurrences of two I's and four H's in a row of 6.

(*Note*: There are $\binom{6}{2} = 15$ elementary outcomes.)

(b) A random variable of interest is the number of runs X that is defined as the number of unbroken sequences of letters of the same kind. For example, the arrangement IHHHIH has four runs, IIHHHH has two. Find the value of X associated with each outcome you listed in part (a).

(c) Obtain the probability distribution of X under the model of randomness.

*5.105 Refer to part (c) of Exercise 5.104. Calculate the mean and standard deviation of X.

5.106 Let the random variable Y denote the proportion of times a head occurs in three tosses of a coin, that is, $Y = (\text{No. of heads in 3 tosses})/3$.

(a) Obtain the probability distribution Y.

(b) Draw the probability histogram.

(c) Calculate the $E(Y)$ and $\text{sd}(Y)$.

5.107 Is the model of Bernoulli trials plausible in each of the following situations? Identify any serious violations of the conditions.

(a) A dentist records if each tooth in the lower jaw has a cavity or has none.

(b) Persons applying for a driver's license will be recorded as writing left- or right-handed.

(c) For each person taking a seat at a lunch counter, observe the time it takes to be served.

(d) Each day of the first week in April is recorded as being either clear or cloudy.

(e) Cars selected at random will or will not pass state safety inspection.

5.108 Give an example (different from those appearing in Exercise 5.107) of repeated trials with two possible outcomes where:

(a) The model of Bernoulli trials is reasonable.

(b) The condition of independence is violated.

(c) The condition of equal $P(S)$ is violated.

5.109 If the probability of having a male child is .5, find the probability that the third child is the first son.

5.110 A basketball team scores 35% of the times it gets the ball. Find the probability that the first basket occurs on its third possession. (Assume independence.)

5.111 If in three Bernoulli trials the probability that the first two trials are both failures is 4/49, what is the probability that the first two are successes and the third is a failure?

5.112 The proportion of people having the blood type O in a large southern city is .4. For two randomly selected donors:

(a) Find the probability of at least one type O.

(b) Find the expected number of type O.

(c) Repeat parts (a) and (b) if there are three donors.

5.113 A viral infection is spread by contact with an infected person. Let the probability that a healthy person gets the infection in one contact be $p = .4$.

(a) An infected person has contact with six healthy persons. Specify the distribution of X = No. of persons who contract the infection.

(b) Find $P[X \leq 3]$, $P[X = 0]$, and $E(X)$.

5.114 The probability that a voter will believe a rumor about a politician is .3. If 20 voters are told individually:

(a) Find the probability that none of the 20 believes the rumor.

(b) Find the probability that seven or more believe the rumor.

(c) Determine the mean and standard deviation of the number who believe the rumor.

5.115 National safety statistics suggest that about 33% of the persons treated in an emergency room because of moped accidents are under 16 years of age. Suppose you count the number of persons under 16 among the next 14 moped accident victims to come to the emergency room.

(a) Find the mean of X.

(b) Find the standard deviation of X.

(c) Find the probability that the first injured person is under 16 years old and the second is at least 16 years old.

5.116 For each situation, state if a binomial distribution is reasonable for the random variable X. Justify your answer.

(a) A multiple-choice examination consists of 10 problems, each of which has 5 suggested answers. A student marks answers by pure guesses (i.e., one answer is chosen at random out of the 5), and X denotes the number of marked answers that are wrong.

(b) A multiple-choice examination has two parts: Part 1 has 8 problems, each with 5 suggested answers, and Part 2 has 10 problems, each with 4 suggested answers. A student marks answers by pure guesses, and X denotes the total number of problems that the student correctly answers.

(c) Twenty-five married couples are interviewed about exercise, and X denotes the number of persons (out of the 50 people interviewed) who exercise regularly.

5.117 A school newspaper claims that 70% of the students support its view on a campus issue. A random sample of 20 students is taken, and 10 students agree with the newspaper. Find $P[10 \text{ or less agree}]$ if 70% support the view and comment on the plausibility of the claim.

5.118 At one large midwest university, about 40% of the college seniors have a social science major. Fourteen seniors will be selected at random. Let X denote the number that have a social science major. Determine

(a) $P[3 \leq X \leq 9]$

(b) $P[3 < X \leq 9]$

(c) $P[3 < X < 9]$

(d) $E(X)$ (e) $\text{sd}(X)$

5.119 Refer to the population of social science majors in Exercise 5.118 but change the sample size to $n = 5$. Using the binomal table,

(a) List the probability distribution.

(b) Plot the probability histogram.

(c) Calculate $E(X)$ and $\text{Var}(X)$ from the entries in the list from part (a).

(d) Calculate $E(X) = np$ and $\text{Var}(X) = npq$ and compare your answer with part (c).

*5.120 For a binomial distribution with $p = .15$, find the smallest number n such that 1 success is more probable than no successes in n trials.

5.121 Only 30% of the people in a large city feel that its mass transit system is adequate. If 20 persons are selected at random, find the probability that 10 or more will feel that the system is adequate. Find the probability that exactly 10 will feel that the system is adequate.

5.122 A sociologist feels that only half of the high school seniors capable of graduating from college go to college. Of 17 high school seniors who have the ability to graduate from college, find the probability that 12 or more will go to college if the sociologist is correct. Assume that the seniors will make their decisions independently. Also find the expected number.

5.123 Jones claims to have extrasensory perception (ESP). In order to test the claim, a psychologist shows Jones five cards that carry different pictures. Then Jones is blindfolded and the psychologist selects one card and asks Jones to identify the picture. This process is repeated 16 times. Suppose, in reality, that Jones has no ESP but responds by sheer guesses.

(a) What is the probability that the identifications are correct at most 8 times?

(b) What is the probability that the identifications are wrong at least 10 times?

(c) Find the expected value and standard deviation of the number of correct identifications.

*5.124 ***Geometric distribution.*** Instead of performing a fixed number of Bernoulli trials, an experimenter performs trials until the first success occurs. The number of successes is now fixed at 1, but the number of trials Y is now random. It can assume any of the values 1, 2, 3, and so on with no upper limit.

(a) Show that $f(y) = q^{y-1}p$ for $y = 1, 2, \ldots$

(b) Find the probability of 3 or fewer trials when $p = .5$.

*5.125 ***Poisson distribution for rare events.*** The Poisson distribution is often appropriate when the probability of an event (success) is small. It has served as a probability model for the number of plankton in a liter of water, calls per hour to an answering service, and earthquakes in a year. The Poisson distribution also approximates the binomial when the expected value np is small but n is large. The Poisson distribution with mean m has the form

$$f(x) = e^{-m}\frac{m^x}{x!} \quad \text{for} \quad x = 0, 1, 2, \ldots$$

where e is the exponential number or 2.718 (rounded) and $x!$ is the number $x(x-1)(x-2)\cdots 1$ with $0! = 1$. Given $m = 3$ and $e^{-3} = .05$, find: (a) $P[X = 0]$, (b) $P[X = 1]$.

5.126 An inspector will sample bags of potato chips to see if they fall short of the weight, 14 ounces, printed on the bag. Samples of 20 bags will be selected and the number with weight less than 14 ounces will be recorded.

(a) Make a p chart using the centerline and control limits calculated for $p_0 = .5$.

(b) Suppose that samples from ten different days have

11 8 14 10 13 12 7 14 10 13

underweight bags. Graph the corresponding proportions on your p chart.

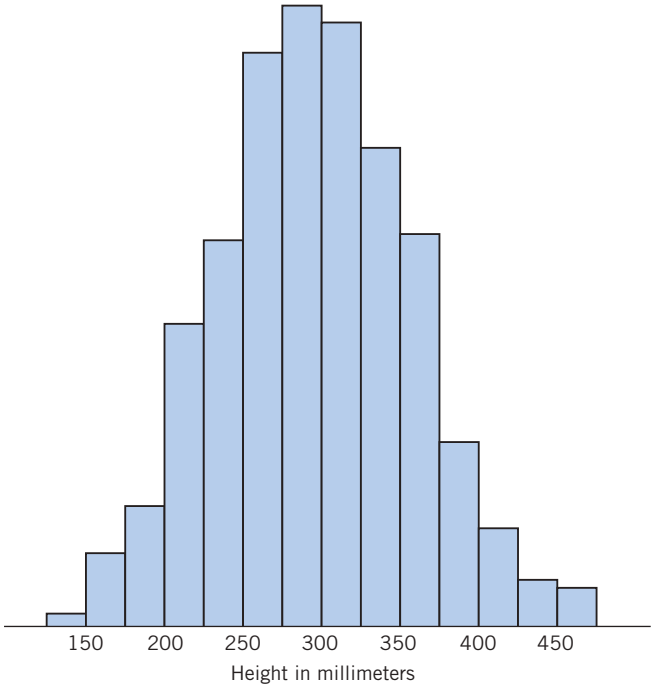
(c) Identify any days where the chart signals "out of control."

6

The Normal Distribution

Bell-shaped Distribution of Heights of Red Pine Seedlings

Trees are a renewable resource that is continually studied to both monitor the current status and improve this valuable natural resource. One researcher followed the growth of red pine seedlings. The heights (mm) of 1456 three-year-old seedlings are summarized in the histogram. This histogram suggests a distribution with a single peak and which falls off in a symmetric manner. The histogram of the heights of adult males, or of adult females, has a similar pattern. A bell-shaped distribution is likely to be appropriate for the size of many things in nature.



1. PROBABILITY MODEL FOR A CONTINUOUS RANDOM VARIABLE

Up to this point, we have limited our discussion to probability distributions of discrete random variables. Recall that a discrete random variable takes on only some isolated values, usually integers representing a count. We now turn our attention to the probability distribution of a continuous random variable—one that can ideally assume any value in an interval. Variables measured on an underlying continuous scale, such as weight, strength, life length, and temperature, have this feature.

Just as probability is conceived as the long-run relative frequency, the idea of a continuous probability distribution draws from the relative frequency histogram for a large number of measurements. The reader may wish to review Section 3.3 of Chapter 2 where grouping of data in class intervals and construction of a relative frequency histogram were discussed. We have remarked that with an increasing number of observations in a data set, histograms can be constructed with class intervals having smaller widths. We will now pursue this point in order to motivate the idea of a continuous probability distribution. To focus the discussion let us consider that the weight X of a newborn baby is the continuous random variable of our interest. How do we conceptualize the probability distribution of X? Initially, suppose that the birth weights of 100 babies are recorded, the data grouped in class intervals of 1 pound, and the relative frequency histogram in Figure 1*a* on page 224, is obtained. Recall that a relative frequency histogram has the following properties:

1. The total area under the histogram is 1.
2. For two points a and b such that each is a boundary point of some class, the relative frequency of measurements in the interval a to b is the **area** under the histogram above this interval.

For example, Figure 1*a* shows that the interval 7.5 to 9.5 pounds contains a proportion $.28 + .25 = .53$ of the 100 measurements.

Next, we suppose that the number of measurements is increased to 5000 and they are grouped in class intervals of .25 pound. The resulting relative frequency histogram appears in Figure 1*b*. This is a refinement of the histogram in Figure 1*a* in that it is constructed from a larger set of observations and exhibits relative frequencies for finer class intervals. (Narrowing the class interval without increasing the number of observations would obscure the overall shape of the distribution.) The refined histogram in Figure 1*b* again has the properties 1 and 2 stated above.

By proceeding in this manner, even further refinements of relative frequency histograms can be imagined with larger numbers of observations and smaller class intervals. In pursuing this conceptual argument, we ignore the difficulty that the accuracy of the measuring device is limited. In the course of refining the histograms, the jumps between consecutive rectangles tend to dampen out, and the top of the histogram approximates the shape of a smooth curve, as illustrated in Figure 1*c*. Because probability is interpreted as long-run relative

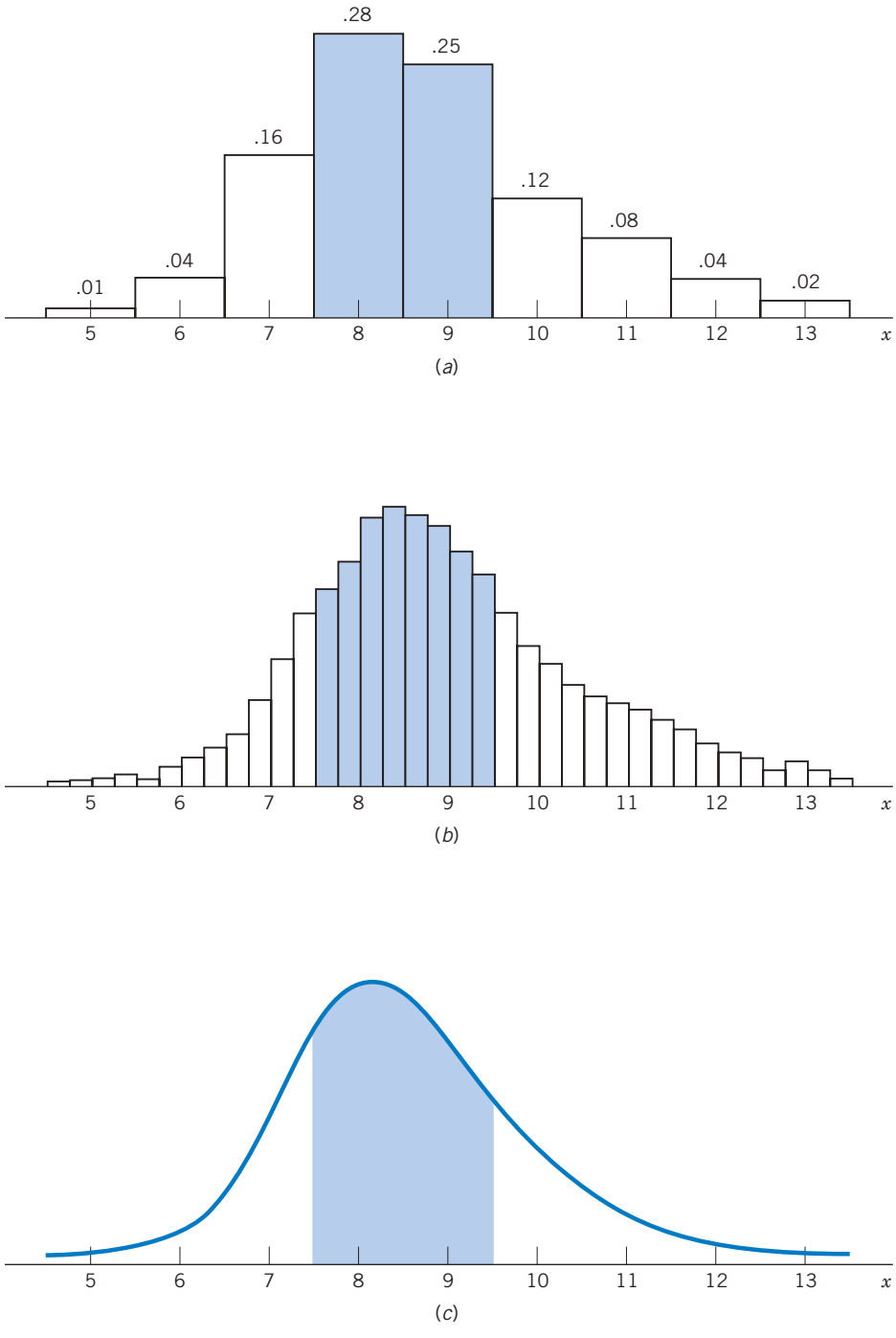


Figure 1 Probability density curve viewed as a limiting form of relative frequency histograms.

frequency, the curve obtained as the limiting form of the relative frequency histograms represents the manner in which the total probability 1 is distributed over the interval of possible values of the random variable X. This curve is called the **probability density curve** of the continuous random variable X. The mathematical function $f(x)$ whose graph produces this curve is called the **probability density function** of the continuous random variable X.

The properties 1 and 2 that we stated earlier for a relative frequency histogram are shared by a probability density curve that is, after all, conceived as a limiting smoothed form of a histogram. Also, since a histogram can never protrude below the x axis, we have the further fact that $f(x)$ is nonnegative for all x.

The **probability density function** $f(x)$ describes the distribution of probability for a continuous random variable. It has the properties:

1. The total area under the probability density curve is 1.
2. $P[a \leq X \leq b]$ = area under the probability density curve between a and b.
3. $f(x) \geq 0$ for all x.

Unlike the description of a discrete probability distribution, the probability density $f(x)$ for a continuous random variable does not represent the probability that the random variable will exactly equal the value x, or the event $[X = x]$ Instead, a probability density function relates the probability of an interval $[a, b]$ to the area under the curve in a strip over this interval. A single point x, being an interval with a width of 0, supports 0 area, so $P[X = x] = 0$.

With a continuous random variable, the probability that $X = x$ is **always** 0. It is only meaningful to speak about the probability that X lies in an interval.

The deduction that the probability at every single point is zero needs some clarification. In the birth-weight example, the statement $P[X = 8.5 \text{ pounds}] = 0$ probably seems shocking. Does this statement mean that no child can have a birth weight of 8.5 pounds? To resolve this paradox, we need to recognize that the accuracy of every measuring device is limited, so that here the number 8.5 is actually indistinguishable from all numbers in an interval surrounding it, say, [8.495, 8.505]. Thus, the question really concerns the probability of an interval surrounding 8.5, and the area under the curve is no longer 0.

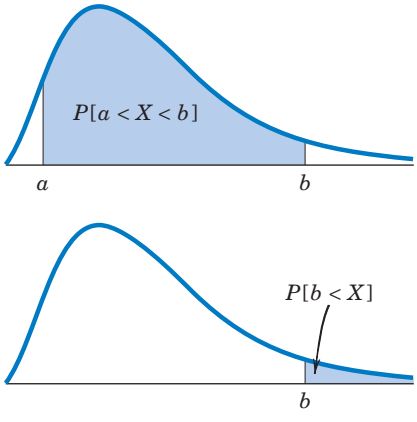
When determining the probability of an interval a to b, we need not be concerned if either or both endpoints are included in the interval. Since the probabilities of $X = a$ and $X = b$ are both equal to 0,

$$P[a \le X \le b] = P[a < X \le b] = P[a \le X < b] = P[a < X < b]$$

In contrast, these probabilities may not be equal for a discrete distribution.

Fortunately, for important distributions, areas have been extensively tabulated. In most tables, the entire area to the left of each point is tabulated. To obtain the probabilities of other intervals, we must apply the following rules.

$$P[a < X < b] = (\text{Area to left of } b) - (\text{Area to the left of } a)$$



$$P[b < X] = 1 - (\text{Area to left of } b)$$

SPECIFICATION OF A PROBABILITY MODEL

A probability model for a continuous random variable is specified by giving the mathematical form of the probability density function. If a fairly large number of observations of a continuous random variable are available, we may try to approximate the top of the staircase silhouette of the relative frequency histogram by a mathematical curve.

In the absence of a large data set, we may tentatively assume a reasonable model that may have been suggested by data from a similar source. Of course, any model obtained in this way must be closely scrutinized to verify that it conforms to the data at hand. Section 6 addresses this issue.

FEATURES OF A CONTINUOUS DISTRIBUTION

As is true for relative frequency histograms, the probability density curves of continuous random variables could possess a wide variety of shapes. A few of these are illustrated in Figure 2. Many statisticians use the term **skewed** for a long tail in one direction.

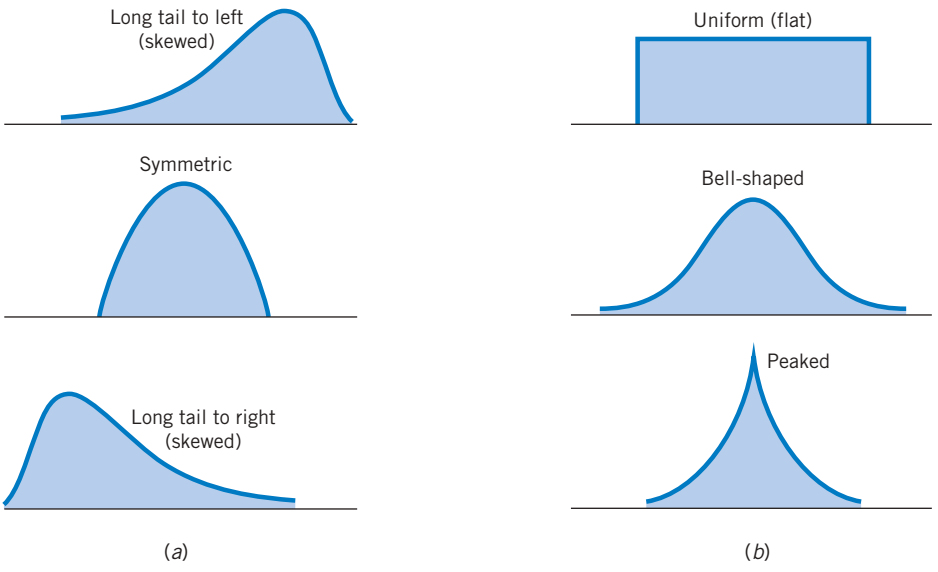


Figure 2 Different shapes of probability density curves. (*a*) Symmetry and deviations from symmetry. (*b*) Different peakedness.

A continuous random variable X also has a mean, or expected value $E(X)$, as well as a variance and a standard deviation. Their interpretations are the same as in the case of discrete random variables, but their formal definitions involve integral calculus and are therefore not pursued here. However, it is instructive to see in Figure 3 that the mean $\mu = E(X)$ marks the balance point of the probability mass. The median, another measure of center, is the value of X that divides the area under the curve into halves.

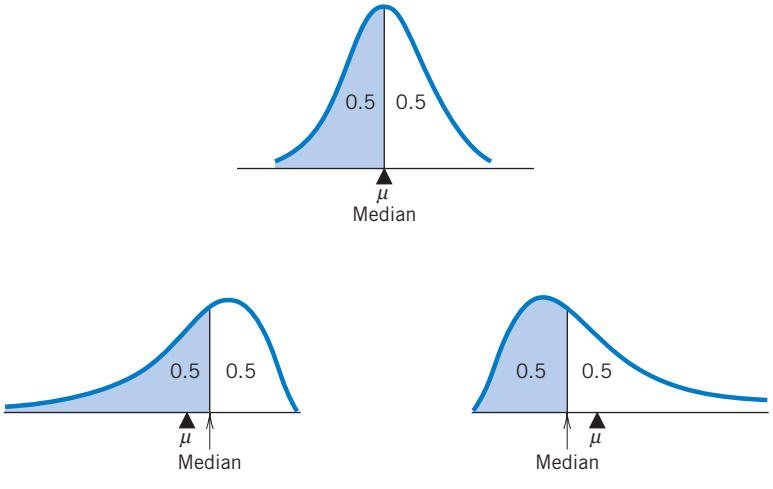


Figure 3 Mean as the balance point and median as the point of equal division of the probability mass.

Besides the median, we can also define the quartiles and other percentiles of a probability distribution.

The population **100*p*-th percentile** is an x value that supports area p to its left and $1 - p$ to its right.

Lower (first) quartile = 25th percentile
Second quartile (or median) = 50th percentile
Upper (third) quartile = 75th percentile

The quartiles for two distributions are shown in Figure 4.

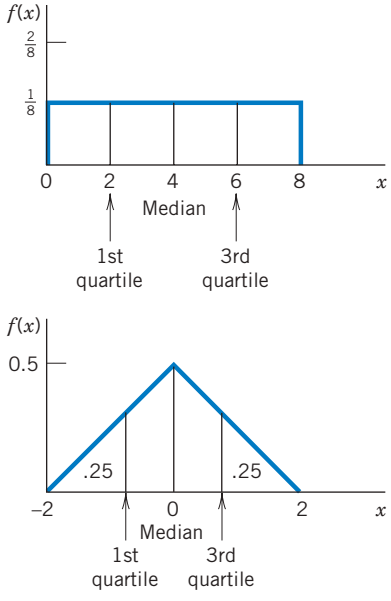


Figure 4 Quartiles of two continuous distributions.

Statisticians often find it convenient to convert random variables to a dimensionless scale. Suppose X, a real estate salesperson's commission for a month, has mean \$4000 and standard deviation \$500. Subtracting the mean produces the deviation $X - 4000$ measured in dollars. Then, dividing by the standard deviation, expressed in dollars, yields the dimensionless variable $Z = (X - 4000)/500$. Moreover, the standardized variable Z can be shown to have mean 0 and standard deviation 1. (See Appendix A2.1 for details.)

The observed values of standardized variables provide a convenient way to compare SAT and ACT scores or compare heights of male and female partners.

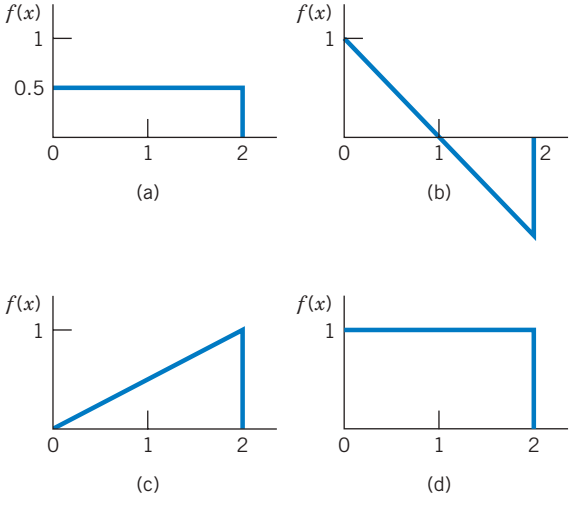
The **standardized variable**

$$Z = \frac{X - \mu}{\sigma} = \frac{\text{Variable} - \text{Mean}}{\text{Standard deviation}}$$

has mean 0 and sd 1.

Exercises

6.1 Which of the functions sketched below could be a probability density function for a continuous random variable? Why or why not?



6.2 Determine the following probabilities from the curve $f(x)$ diagrammed in Exercise 6.1(a).

(a) $P[\,0 < X < .5\,]$

(b) $P[\,.5 < X < 1\,]$

(c) $P[\,1.5 < X < 2\,]$

(d) $P[\,X = 1\,]$

6.3 For the curve $f(x)$ graphed in Exercise 6.1(c), which of the two intervals $[\,0 < X < .5\,]$ or $[\,1.5 < X < 2\,]$ is assigned a higher probability?

6.4 Determine the median and the quartiles for the probability distribution depicted in Exercise 6.1(a).

6.5 Determine the median and the quartiles for the curve depicted in Exercise 6.1(c).

6.6 Determine the 15th percentile of the curve in Exercise 6.1(a).

6.7 If a student is more likely to be late than on time for the 1:20 PM history class:

(a) Determine if the median of the student's arrival time distribution is earlier than, equal to, or later than 1:20 PM.

(b) On the basis of the given information, can you determine if the mean of the student's arrival time distribution is earlier than, equal to, or later than 1:20 PM? Comment.

6.8 Which of the distributions in Figure 3 are compatible with the following statements?

(a) The first test was too easy because over half the class scored above the mean.

(b) In spite of recent large increases in salary, half of the professional football players still make less than the average salary.

6.9 Find the standardized variable Z if X has

(a) Mean 15 and standard deviation 4.

(b) Mean 61 and standard deviation 9.

(c) Mean 161 and variance 25.

6.10 Males 20 to 29 years old have a mean height of 70.0 inches with a standard deviation of 3.0 inches. Females 20 to 29 years old have a mean height of 64.2 inches with a standard deviation of 2.6 inches. (Based on *Statistical Abstract of the U.S. 2009*, Table 201.)

(a) Find the standardized variable for the heights of males.

(b) Find the standardized variable for the heights of females.

(c) For a 68-inch-tall person, find the value of the standardized variable for males.

(d) For a 68-inch-tall person, find the value of the standardized variable for females. Compare your answer with part (c) and comment.

6.11 Find the standardized variable Z if X has

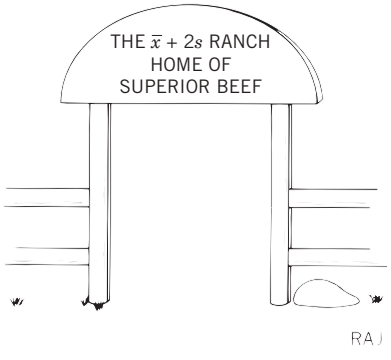
(a) Mean 7 and standard deviation 2.

(b) Mean 250 and standard deviation 6.

(c) Mean 444 and variance 81.

2. THE NORMAL DISTRIBUTION—ITS GENERAL FEATURES

The normal distribution, which may already be familiar to some readers as the curve with the bell shape, is sometimes associated with the names of Pierre Laplace and Carl Gauss, who figured prominently in its historical development. Gauss derived the normal distribution mathematically as the probability distribution of the error of measurements, which he called the "normal law of errors." Subsequently, astronomers, physicists, and, somewhat later, data collectors in a wide variety of fields found that their histograms exhibited the common feature of first rising gradually in height to a maximum and then decreasing in a symmetric manner. Although the normal curve is not unique in exhibiting this form, it has been found to provide a reasonable approximation in a great many situations. Unfortunately, at one time during the early stages of the development of statistics, it had many overzealous admirers. Apparently, they felt that all real-life data must conform to the bell-shaped normal curve, or otherwise, the process of data collection should be suspect. It is in this context that the distribution became known as the **normal distribution.** However, scrutiny of data has often revealed inadequacies of the normal distribution. In fact, the universality of the normal distribution is only a myth, and examples of quite nonnormal distributions abound in virtually every field of study. Still, the normal distribution plays a central role in statistics, and inference procedures derived from it have wide applicability and form the backbone of current methods of statistical analysis.



Does the $\bar{x}$ ranch have average beef?

A **normal distribution** has a bell-shaped density[1] as shown in Figure 5. It has

$$\text{Mean} = \mu$$
$$\text{Standard deviation} = \sigma$$

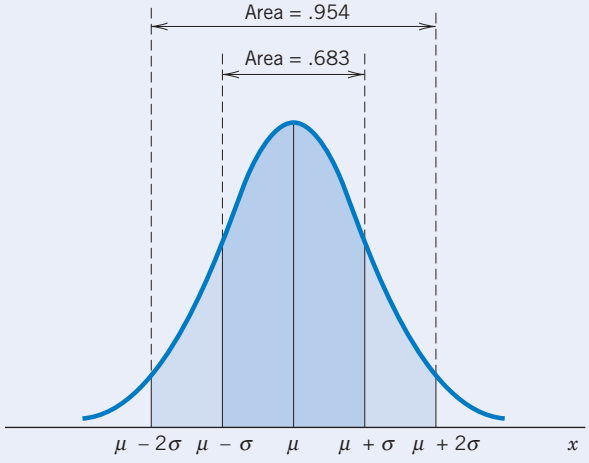


Figure 5 Normal distribution.

The probability of the interval extending

One sd on each side of the mean: $P[\mu - \sigma < X < \mu + \sigma] = .683$

Two sd on each side of the mean: $P[\mu - 2\sigma < X < \mu + 2\sigma] = .954$

Three sd on each side of the mean: $P[\mu - 3\sigma < X < \mu + 3\sigma] = .997$

Notation

The normal distribution with a mean of μ and a standard deviation of σ is denoted by $N(\mu, \sigma)$.

[1]The formula, which need not concern us, is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for} \quad -\infty < x < \infty$$

where π is the area of a circle having unit radius, or approximately 3.1416, and e is approximately 2.7183.

Although we are speaking of the importance of the normal distribution, our remarks really apply to a whole class of distributions having bell-shaped densities. There is a normal distribution for each value of its mean μ and its standard deviation σ.

A few details of the normal curve merit special attention. The curve is symmetric about its mean μ, which locates the peak of the bell (see Figure 5). The interval running one standard deviation in each direction from μ has a probability of .683, the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$ has a probability of .954, and the interval from $\mu - 3\sigma$ to $\mu + 3\sigma$ has a probability of .997. It is these probabilities that give rise to the empirical rule stated in Chapter 2. The curve never reaches 0 for any value of x, but because the tail areas outside $(\mu - 3\sigma, \mu + 3\sigma)$ are very small, we usually terminate the graph at these points.

Interpreting the parameters, we can see in Figure 6 that a change of mean from μ_1 to a larger value μ_2 merely slides the bell-shaped curve along the axis until a new center is established at μ_2. There is no change in the shape of the curve.

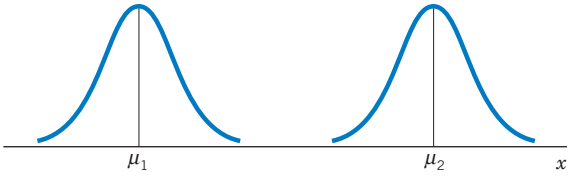


Figure 6 Two normal distributions with different means, with μ_1 less than μ_2, but the same standard deviation.

A different value for the standard deviation results in a different maximum height of the curve and changes the amount of the area in any fixed interval about μ (see Figure 7). The position of the center does not change if only σ is changed.

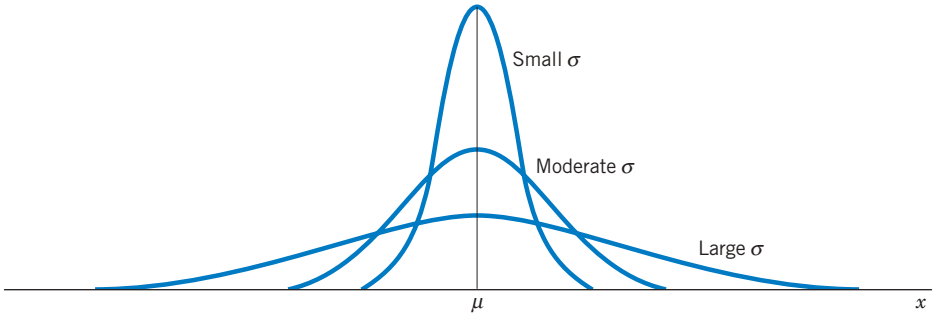


Figure 7 Decreasing σ increases the maximum height and the concentration of probability about μ.

3. THE STANDARD NORMAL DISTRIBUTION

The particular normal distribution that has a mean of 0 and a standard deviation of 1 is called the **standard normal distribution.** It is customary to denote the standard normal variable by Z. The standard normal curve is illustrated in Figure 8.

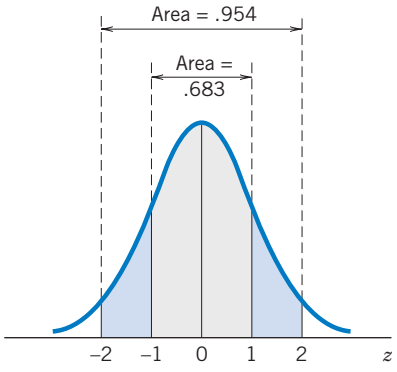


Figure 8 The standard normal curve.

The **standard normal distribution** has a bell-shaped density with

$$\text{Mean } \mu = 0$$
$$\text{Standard deviation } \sigma = 1$$

The standard normal distribution is denoted by $N(0, 1)$.

USE OF THE STANDARD NORMAL TABLE (APPENDIX B, TABLE 3)

The standard normal table in the appendix gives the area to the left of a specified value of z as

$$P[Z \leq z] = \text{Area under curve to the left of } z$$

For the probability of an interval $[a, b]$,

$$P[a \leq Z \leq b] = [\text{Area to left of } b] - [\text{Area to left of } a]$$

The following properties can be observed from the symmetry of the standard normal curve about 0 as exhibited in Figure 9.

1. $P[Z \leq 0] = .5$
2. $P[Z \leq -z] = 1 - P[Z \leq z] = P[Z \geq z]$

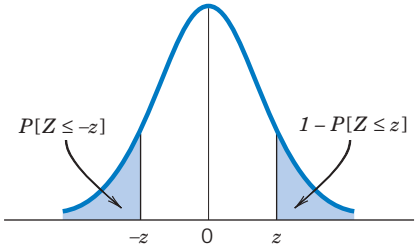


Figure 9 Equal normal tail probabilities.

Example 1 Determining Standard Normal Probabilities for Tail Events

Find $P[Z \leq 1.37]$ and $P[Z > 1.37]$.

SOLUTION From the normal table, we see that the probability or area to the left of 1.37 is .9147. (See Table 1.) Consequently, $P[Z \leq 1.37] = .9147$. Moreover, because $[Z > 1.37]$ is the complement of $[Z \leq 1.37]$,

$$P[Z > 1.37] = 1 - P[Z \leq 1.37] = 1 - .9147 = .0853$$

as we can see in Figure 10. An alternative method is to use symmetry to show that $P[Z > 1.37] = P[Z < -1.37]$, which can be obtained directly from the normal table.

TABLE 1 How to Read from Appendix B, Table 3 for $z = 1.37 = 1.3 + .07$

z	.00	⋯	.07	⋯
.0			↓	
⋮			⋮	
1.3			.9147	
⋮				

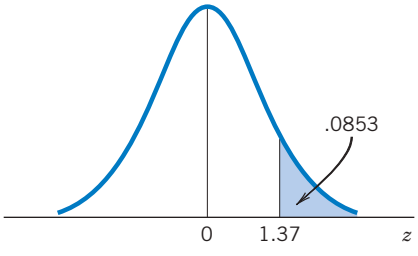


Figure 10 An upper tail normal probability.

Example 2 Determining the Standard Normal Probability of an Interval

Calculate $P[-.155 < Z < 1.60]$.

SOLUTION From Appendix B, Table 3, we see that

$$P[Z \leq 1.60] = \text{Area to left of } 1.60 = .9452$$

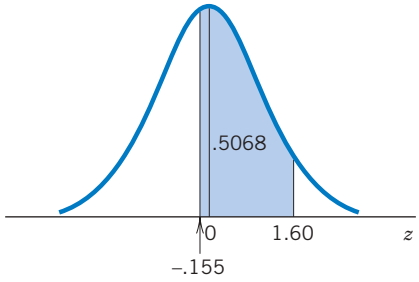


Figure 11 Normal probability of an interval.

We interpolate[2] between the entries for $-.15$ and $-.16$ to obtain

$$P[Z \leq -.155] = \text{Area to left of } -.155 = .4384$$

Therefore,

$$P[-.155 < Z < 1.60] = .9452 - .4384 = .5068$$

which is the shaded area in Figure 11.

Example 3 Determining the Standard Normal Probability Outside of an Interval

Find $P[Z < -1.9 \text{ or } Z > 2.1]$.

SOLUTION The two events $[Z < -1.9]$ and $[Z > 2.1]$ are incompatible, so we add their probabilities:

$$P[Z < -1.9 \text{ or } Z > 2.1] = P[Z < -1.9] + P[Z > 2.1]$$

As indicated in Figure 12,

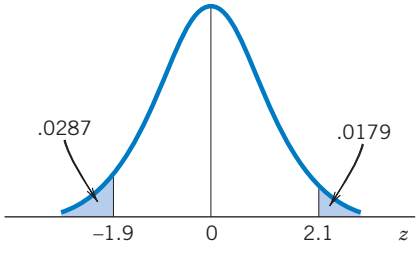


Figure 12 Normal probabilities for Example 3.

[2]Since $z = -.155$ is halfway between $-.15$ and $-.16$, the interpolated value is halfway between the table entries .4404 and .4364. The result is .4384. We actually used computer software (see Technology section) to get .4384. You may want to just eyeball a value between two entries in the table.

$P[Z > 2.1]$ is the area to the right of 2.1, which is $1 - [\text{Area to left of } 2.1] = 1 - .9821 = .0179$. The normal table gives $P[Z < -1.9] = .0287$ directly. Adding these two quantities, we get

$$P[Z < -1.9 \text{ or } Z > 2.1] = .0287 + .0179 = .0466$$

Example 4 Determining an Upper Percentile of the Standard Normal Distribution

Locate the value of z that satisfies $P[Z > z] = .025$.

SOLUTION If we use the property that the total area is 1, the area to the left of z must be $1 - .0250 = .9750$. The marginal value with the tabular entry .9750 is $z = 1.96$ (diagrammed in Figure 13).

z	.00	⋯	.06	⋯
.0			↑	
⋮			⋮	
1.9			←———	.9750
⋮				

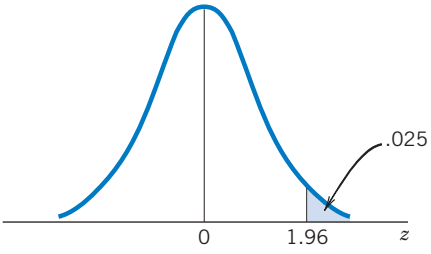


Figure 13 $P[Z > 1.96] = .025$.

Example 5 Determining z for Given Equal Tail Areas

Obtain the value of z for which $P[-z \le Z \le z] = .90$.

SOLUTION We observe from the symmetry of the curve that

$$P[Z < -z] = P[Z > z] = .05$$

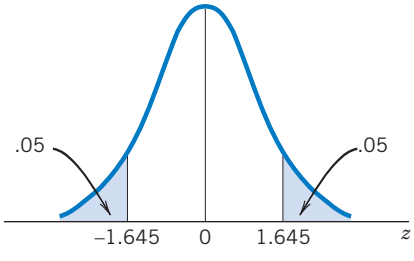


Figure 14 $P[Z < -1.645 \text{ or } Z > 1.645] = .10$.

From the normal table, we see that $z = 1.65$ gives $P[Z < -1.65] = .0495$ and $z = 1.64$ gives $P[Z < -1.64] = .0505$. Because .05 is halfway between these two probabilities, we interpolate between the two z values to obtain $z = 1.645$ (see Figure 14).

Suggestion: The preceding examples illustrate the usefulness of a sketch to depict an area under the standard normal curve. A correct diagram shows how to combine the left-side areas given in the normal table.

Exercises

6.12 Find the area under the standard normal curve to the left of

(a) $z = 1.16$ (b) $z = .16$

(c) $z = -1.71$ (d) $z = -2.43$

6.13 Find the area under the standard normal curve to the left of

(a) $z = .83$ (b) $z = 1.03$

(c) $z = -1.03$ (d) $z = -1.35$

6.14 Find the area under the standard normal curve to the right of

(a) $z = 1.16$

(b) $z = .64$

(c) $z = -1.71$

(d) $z = -1.525$ (interpolate)

6.15 Find the area under the standard normal curve to the right of

(a) $z = .83$

(b) $z = 2.83$

(c) $z = -1.23$

(d) $z = 1.635$ (interpolate)

6.16 Find the area under the standard normal curve over the interval

(a) $z = -.75$ to $z = .75$

(b) $z = -1.09$ to $z = 1.09$

(c) $z = .32$ to $z = 2.65$

(d) $z = -.745$ to $z = 1.244$ (interpolate)

6.17 Find the area under the standard normal curve over the interval

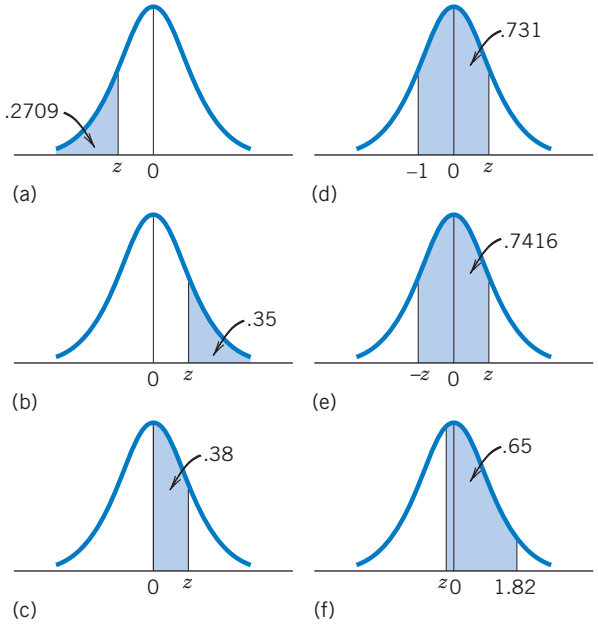
(a) $z = -.44$ to $z = .44$

(b) $z = -1.33$ to $z = 1.33$

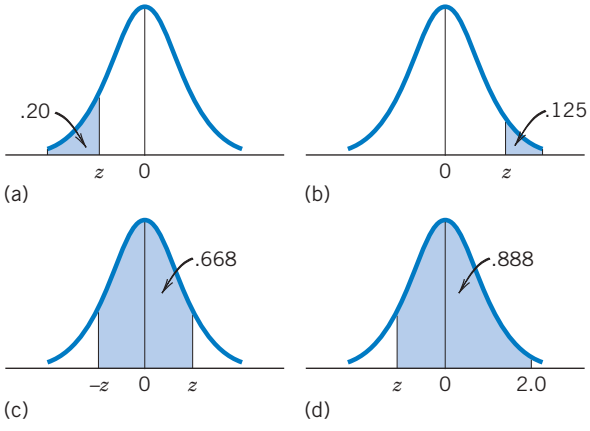
(c) $z = .40$ to $z = 2.03$

(d) $z = 1.405$ to $z = 2.306$ (interpolate)

6.18 Identify the z values in the following diagrams of the standard normal distribution (interpolate, as needed).



6.19 Identify the z values in the following diagrams of the standard normal distribution (interpolate, as needed).



6.20 For a standard normal random variable Z, find

(a) $P[Z < .62]$
(b) $P[Z < -.62]$
(c) $P[Z > 1.49]$
(d) $P[Z > -1.49]$
(e) $P[-1.3 < Z < 2.61]$
(f) $P[.08 < Z < .8]$
(g) $P[-1.62 < Z < -.34]$
(h) $P[|Z| < 1.65]$

6.21 Find the z value in each of the following cases.

(a) $P[Z < z] = .1762$
(b) $P[Z > z] = .10$
(c) $P[-z < Z < z] = .954$
(d) $P[-.6 < Z < z] = .50$

6.22 Find the quartiles of the standard normal distribution.

6.23 Find

(a) $P[Z < .33]$.
(b) The 33rd percentile of the standard normal distribution.
(c) $P[Z < .70]$.
(d) The 70th percentile of the standard normal distribution.

6.24 Find

(a) $P[Z < .46]$.
(b) The 46th percentile of the standard normal distribution.
(c) $P[Z < .85]$.
(d) The 85th percentile of the standard normal distribution.

4. PROBABILITY CALCULATIONS WITH NORMAL DISTRIBUTIONS

Fortunately, no new tables are required for probability calculations regarding the general normal distribution. Any normal distribution can be set in correspondence to the standard normal by the following relation.

If X is distributed as $N(\mu, \sigma)$, then the standardized variable

$$Z = \frac{X - \mu}{\sigma}$$

has the standard normal distribution.

This property of the normal distribution allows us to cast a probability problem concerning X into one concerning Z. To find the probability that X lies in a given interval, convert the interval to the z scale and then calculate the probability by using the standard normal table (Appendix B, Table 3).

Example 6 Converting a Normal Probability to a Standard Normal Probability

Given that X has the normal distribution $N(60, 4)$, find $P[55 \leq X \leq 63]$.

SOLUTION Here, the standardized variable is $Z = \dfrac{X - 60}{4}$. The distribution of X is shown in Figure 15, where the distribution of Z and the z scale are also displayed below the x scale. In particular,

$$x = 55 \text{ gives } z = \frac{55 - 60}{4} = -1.25$$

$$x = 63 \text{ gives } z = \frac{63 - 60}{4} = .75$$

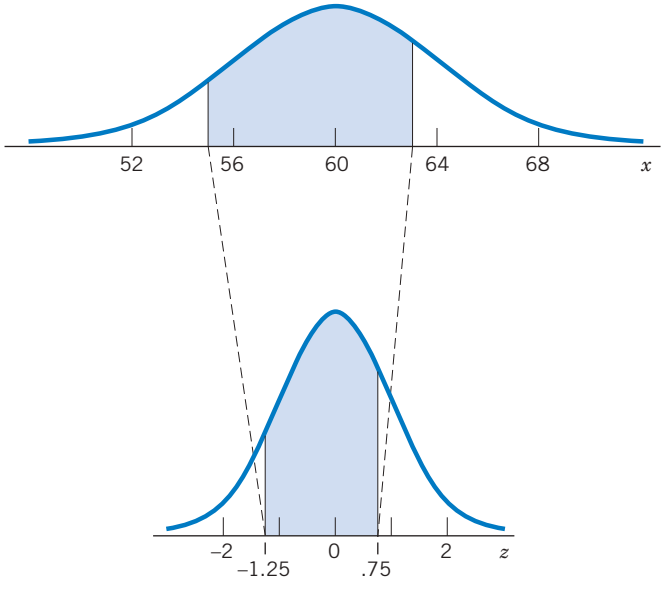


Figure 15 Converting to the z scale.

Therefore,

$$P[55 \leq X \leq 63] = P[-1.25 \leq Z \leq .75]$$

Using the normal table, we find $P[Z \leq .75] = .7734$ and $P[Z \leq -1.25] = .1056$, so the required probability is $.7734 - .1056 = .6678$.

The working steps employed in Example 6 can be formalized into the rule:

If X is distributed as $N(\mu, \sigma)$, then

$$P[a \leq X \leq b] = P\left[\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right]$$

where Z has the standard normal distribution.

Example 7 Probabilities Concerning Calories in a Lunch Salad

The number of calories in a salad on the lunch menu is normally distributed with mean = 200 and sd = 5. Find the probability that the salad you select will contain:

(a) More than 208 calories.

(b) Between 190 and 200 calories.

SOLUTION Letting X denote the number of calories in the salad, we have the standardized variable

$$Z = \frac{X - 200}{5}$$

(a) The z value corresponding to $x = 208$ is

$$z = \frac{208 - 200}{5} = 1.6$$

Therefore,

$$\begin{aligned} P[X > 208] &= P[Z > 1.6] \\ &= 1 - P[Z \leq 1.6] \\ &= 1 - .9452 = .0548 \end{aligned}$$

(b) The z values corresponding to $x = 190$ and $x = 200$ are

$$\frac{190 - 200}{5} = -2.0 \quad \text{and} \quad \frac{200 - 200}{5} = 0$$

respectively. We calculate

$$\begin{aligned} P[190 \leq X \leq 200] &= P[-2.0 \leq Z \leq 0] \\ &= .5 - .0228 = .4772 \end{aligned}$$

Example 8 Determining a Percentile of a Normal Population

The hours of sleep data in Example 5, Chapter 2, suggest that the population of hours of sleep can be modeled as a normal distribution with mean = 7.2 hours and sd = 1.3 hours.

(a) Determine the probability assigned to sleeping less than 6.5 hours.

(b) Find the 70th percentile of the distribution for hours of sleep.

SOLUTION If we denote the hours of sleep by X, the standardized score

$$Z = \frac{X - 7.2}{1.3}$$

is distributed as $N(0, 1)$.

(a) The z score corresponding to 6.5 is

$$z = \frac{6.5 - 7.2}{1.3} = -.538$$

So, interpolating,

$$P[X < 6.5] = P[Z < -.538] = .295$$

Thus, 29.5%, or about 30%, of the students sleep less than 6.5 hours. In other words, 6.5 hours nearly locates the 30th percentile.

(b) We first find the 70th percentile in the z scale and then convert it to the x scale. From the standard normal table, we interpolate to find

$$P[Z \leq .524] = .70$$

The standardized score $z = .524$ corresponds to

$$\begin{aligned} x &= 7.2 + 1.3(.524) \\ &= 7.88 \end{aligned}$$

Therefore, the 70th percentile score is about 7.88 or nearly eight hours.

Exercises

6.25 Records suggest that the normal distribution with mean 50 and standard deviation 9 is a plausible model for a measurement of the amount of suspended solids (ppm) in river water. Find

(a) $P[X < 46.4]$
(b) $P[X \leq 57.2]$
(c) $P[X > 57.2]$
(d) $P[X > 60.8]$
(e) $P[33.8 \leq X \leq 64.4]$
(f) $P[52.5 \leq X \leq 60.9]$

6.26 Data suggests that the normal distribution with mean 13.0 and standard deviation 2.4 is a plausible model for the length (feet) of adult anaconda snakes. Find

(a) $P[X < 10.4]$
(b) $P[X \leq 17.8]$
(c) $P[X > 17.8]$
(d) $P[X > 16.72]$
(e) $P[10.24 \leq X \leq 18.4]$
(f) $P[14.8 \leq X \leq 17.2]$

6.27 Referring to Exercise 6.25, find b such that

(a) $P[X < b] = .975$
(b) $P[X > b] = .025$
(c) $P[X < b] = .305$

6.28 Referring to Exercise 6.26, find b such that

(a) $P[X < b] = .7995$
(b) $P[X > b] = .002$
(c) $P[X < b] = .015$

6.29 Scores on a certain nationwide college entrance examination follow a normal distribution with a mean of 500 and a standard deviation of 100. Find the probability that a student will score:

(a) Over 650.
(b) Less than 250.
(c) Between 325 and 675.

6.30 Refer to Exercise 6.29.

(a) If a school only admits students who score over 680, what proportion of the student pool would be eligible for admission?
(b) What limit would you set that makes 50% of the students eligible?
(c) What should be the limit if only the top 15% are to be eligible?

6.31 According to the children's growth chart that doctors use as a reference, the heights of two-year-old boys are nearly normally distributed with a mean of 34.5 inches and a standard deviation of 1.4 inches. If a two-year-old boy is selected at random, what is the probability that he will be between 32.5 and 36.5 inches tall?

6.32 The time it takes a symphony orchestra to play Beethoven's *Ninth Symphony* has a normal distribution with a mean of 64.3 minutes and a standard deviation of 1.15 minutes. The next time it is played, what is the probability that it will take between 62.5 and 67.7 minutes?

6.33 The weights of apples served at a restaurant are normally distributed with a mean of 5 ounces and standard deviation of 1.2 ounces. What is the probability that the next person served will be given an apple that weighs less than 4 ounces?

6.34 The diameter of hail hitting the ground during a storm is normally distributed with a mean of .5 inch and a standard deviation of .1 inch. What is the probability that:

(a) A hailstone picked up at random will have a diameter greater than .71 inch?

(b) Two hailstones picked up in a row will have diameters greater than .6 inch? (Assume independence of the two diameters.)

(c) By the end of the storm, what proportion of the hailstones would have had diameters greater than .71 inch?

6.35 Refer to Exercise 6.10 where, according to current U.S. Census Bureau data, the heights of 20- to 29-year-old women can be well approximated by a normal distribution with mean 64.2 inches and standard deviation 2.6 inches.

(a) What is the probability that the height of a randomly selected woman 20 to 29 years old exceeds 70 inches?

(b) What is the probability that the height of a randomly selected woman 20 to 29 years old is less than or equal to 60 inches?

6.36 Suppose the contents of bottles of water coming off a production line have a normal distribution with mean 9.1 ounces and standard deviation .1 ounce.

(a) If every bottle is labeled 9 ounces, what proportion of the bottles contain less than the labeled amount?

(b) If only 2.5% of the bottles exceed weight w, what is the value of w?

6.37 The time for an emergency medical squad to arrive at the sports center at the edge of town is distributed as a normal variable with $\mu = 17$ minutes and $\sigma = 3$ minutes.

(a) Determine the probability that the time to arrive is:

(i) More than 22 minutes.

(ii) Between 13 and 21 minutes.

(iii) Between 15.5 and 18.5 minutes.

(b) Which arrival period of duration 1 minute is assigned the highest probability by the normal distribution?

6.38 The force required to puncture a cardboard mailing tube with a sharp object is normally distributed with mean 32 pounds and standard deviation 4 pounds. What is the probability that a tube will puncture if it is struck by

(a) A 25-pound blow with the object?

(b) A 38-pound blow with the object?

5. THE NORMAL APPROXIMATION TO THE BINOMIAL

The binomial distribution, introduced in Chapter 5, pertains to the number of successes X in n independent trials of an experiment. When the success probability p is not too near 0 or 1 and the number of trials is large, the normal distribution serves as a good approximation to the binomial probabilities. Bypassing the mathematical proof, we concentrate on illustrating the manner in which this approximation works.

Figure 16 presents the binomial distribution for the number of trials n being 5, 12, and 25 when $p = .4$. Notice how the distribution begins to assume the

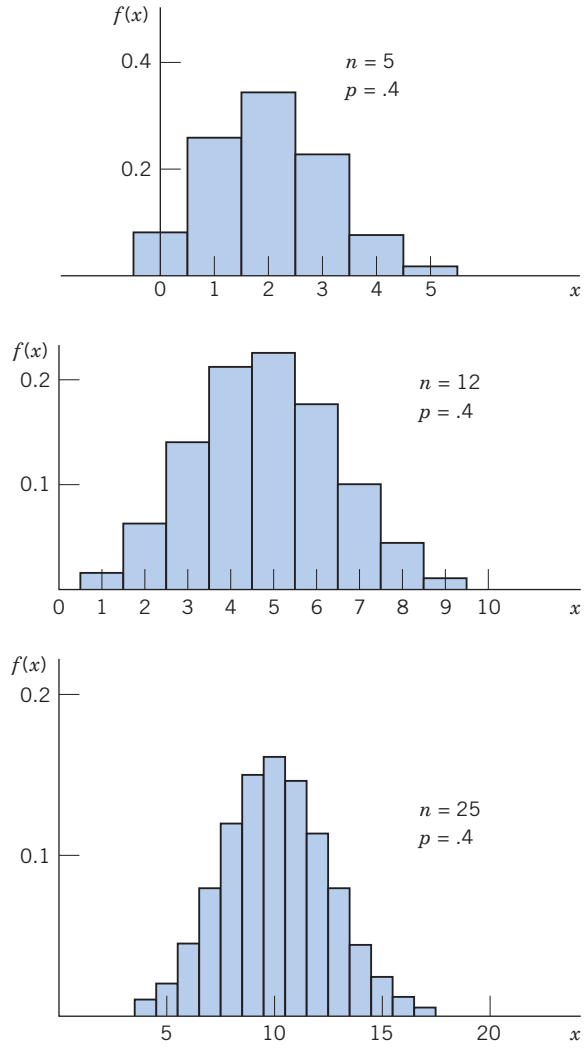


Figure 16 The binomial distributions for $p = .4$ and $n = 5, 12, 25$.

distinctive bell shape for increasing n. Even though the binomial distributions with $p = .4$ are not symmetric, the lack of symmetry becomes negligible for large n.

Figure 17 presents the binomial distribution with $p = .4$ but with n increased to 40. The normal distribution having the same mean $\mu = np = 40 \times .4 = 16$ and variance $\sigma^2 = np(1-p) = 40 \times .4 \times .6 = 9.6$ is also shown. The approximation is quite good. The approximation illustrated in Figure 17 provides the clue on how to approximate the binomial probability

$$P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

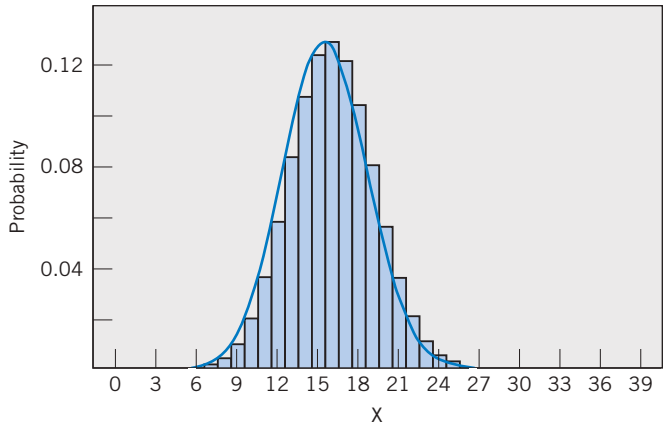


Figure 17 The binomial distribution for $p = .4$ and $n = 40$ along with the normal density having the same mean 16 and standard deviation 3.10.

by a normal probability. The normal probability assigned to a single value x is zero. However, as shown in Figure 18, the probability assigned to the interval $x - \frac{1}{2}$ to $x + \frac{1}{2}$ is the appropriate comparison. The addition and subtraction of $\frac{1}{2}$ is called the **continuity correction.**

The idea of the continuity correction is to approximate the rectangle with area $\binom{n}{x} p^x(1-p)^{n-x}$ by the area under a normal curve. For $n = 15$ and $p = .4$, the binomial distribution assigns

$$P[X = 7] = .787 - .610 = .177$$

Recall from Chapter 5 that the binomial distribution has

$$\text{Mean} = np = 15(.4) = 6$$
$$\text{sd} = \sqrt{np(1-p)} = \sqrt{15(.4)(.6)} = 1.897$$

To obtain an approximation, we select the normal distribution with the same mean, $\mu = 6$, and same $\sigma = 1.897$. The normal approximation is then the probability assigned to the interval $7 - \frac{1}{2}$ to $7 + \frac{1}{2}$.

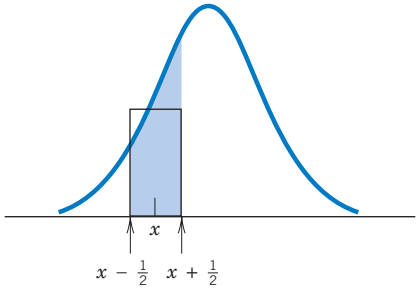


Figure 18 Idea of continuity correction.

$$P[6.5 < X < 7.5] = P\left[\frac{6.5 - 6}{1.897} < \frac{X - 6}{1.897} < \frac{7.5 - 6}{1.897}\right]$$
$$\approx P[.264 < Z < .791] = .7855 - .6041 = .1814$$

Of course $n = 15$ is small, so the approximation .1814 differs somewhat from the exact value .177. However, the accuracy of the approximation increases with the number of trials n.

The normal approximation to the binomial applies when n is large and the success probability p is not too close to 0 or 1. The binomial probability of $[a \le X \le b]$ is approximated by the normal probability of $[a - \frac{1}{2} \le X \le b + \frac{1}{2}]$.

The Normal Approximation to the Binomial

When np and $n(1 - p)$ are both large,[3] say, greater than 15, the binomial distribution is well approximated by the normal distribution having mean $= np$ and sd $= \sqrt{np(1 - p)}$. That is,

$$Z = \frac{X - np}{\sqrt{np(1 - p)}} \quad \text{is approximately } N(0, 1)$$

Example 9 Normal Approximation to the Binomial

In a large scale statewide survey concerning television viewing by children, about 40% of the babies a few months old were reported to watch TV regularly. In a future random sample of 150 babies in this age group, let X be the number who regularly watch TV. Approximate the probability that

(a) X is between 52 and 71 both inclusive.

(b) X is greater than 67.

SOLUTION

(a) Because the population is large, X has a binomial distribution with $p = .4$. To obtain a normal approximation, we first calculate the mean and standard deviation of X. Since $n = 150$,

$$\text{Mean} = np = 150(.4) = 60$$

$$\text{sd} = \sqrt{np(1 - p)} = \sqrt{150(.4)(.6)} = \sqrt{36} = 6$$

The standardized variable is

$$Z = \frac{X - 60}{6}$$

The event $[52 \le X \le 71]$ includes both endpoints. The appropriate continuity correction is to *subtract* $\frac{1}{2}$ from the lower end and *add* $\frac{1}{2}$ to the upper end. We then approximate

[3]To be consistent with the rule in Chapter 13 some suggest using the normal approximation when np is greater than 5. The exact calculation of binomial probabilities using statistical software is always preferable.

$$P[51.5 \le X \le 71.5] = P\left[\frac{51.5 - 60}{6} \le \frac{X - 60}{6} \le \frac{71.5 - 60}{6}\right]$$
$$\approx P[-1.417 \le Z \le 1.917]$$

where Z is standard normal. From the normal table, we interpolate

$$P[-1.417 \le Z \le 1.917] = .9724 - .0782 = .8942$$

and approximate $P[52 \le X \le 71]$ by the normal probability .8942.

(b) For $[X > 67]$, we reason that 67 is not included so that $[X \ge 67 + .5]$ or $[X \ge 67.5]$ is the event of interest:

$$P[X \ge 67.5] = P\left[\frac{X - 60}{6} \ge \frac{67.5 - 60}{6}\right]$$
$$\approx P[Z \ge 1.25] = 1 - .8944$$

The normal approximation to the binomial gives $P[X > 67] \approx .1056$.

Example 10 A Normal Probability Approximation for a Survey

A recent study reported that 54% of the adults in the United States drink at least one cup of coffee a day. If this is still the current rate, what is the probability that in a random sample of 1000 adults the number that drink at least one cup of coffee a day will be (a) less than 519 and (b) 556 or more?

SOLUTION Let X be the number of adults in the sample of 1000 adults who drink at least one cup of coffee a day. Under the assumption that the proportion remains at .54, the distribution of X is well modeled by the binomial distribution with $n = 1000$ and $p = .54$. Since n is large and

$$np = 540, \qquad \sqrt{np(1-p)} = \sqrt{284.4} = 15.76$$

the binomial distribution of X is approximately $N(540, 15.76)$.

(a) Because X is a count, the event $[X < 519]$ is the same as $[X \le 518]$. Using the continuity correction, we have

$$P[X \le 518] \approx P\left[Z \le \frac{518.5 - 540}{15.76}\right]$$
$$= P[Z \le -1.364]$$
$$= .0863$$

(b)

$$P[X \ge 556] \approx P\left[Z \ge \frac{555.5 - 540}{15.76}\right]$$
$$= P[Z \ge .9835]$$
$$= 1 - .8373 = .1627$$

Remark: If the object is to calculate binomial probabilities, today the best practice is to evaluate them directly using an established statistical

computing package. The numerical details need not concern us. However, the fact that

$$\frac{X - np}{\sqrt{np(1 - p)}} \quad \text{is approximately normal}$$

when np and $n(1 - p)$ are both large remains important. We will use it in later chapters when discussing inferences about proportions. Because the continuity correction will not be crucial, we will drop it for the sake of simplicity. Beyond this chapter, we will employ the normal approximation but *without* the continuity correction.

Exercises

6.39 Let the number of successes X have a binomial distribution with $n = 25$ and $p = .6$.

(a) Find the exact probabilities of each of the following:

$X = 17 \qquad 11 \le X \le 18 \qquad 11 < X < 18$

(b) Apply the normal approximation to each situation in part (a).

6.40 Let the number of successes X have a binomial distribution with $n = 25$ and $p = .4$.

(a) Find the exact probability of each of the following:

$X = 11 \qquad 6 \le X \le 12 \qquad 6 < X < 12$

(b) Apply the normal approximation to each situation in part (a).

6.41 A survey by the National Endowment of the Arts concerned participation in music, plays, or dance performance. About 17% of persons 18–24 years old participated in the last 12 months. Suppose you will randomly select $n = 300$ persons in this age group. Let success correspond to participation and let X denote the number of successes. Approximate the probability of (a) $X = 60$ (b) $X \le 45$ and (c) $48 \le X \le 69$

6.42 A National Newspaper Association survey showed that 66% of adults would prefer to get their local news and information from a local paper. Suppose you will randomly select $n = 200$ adults. Let success correspond to prefer local paper and let X denote the number of successes. Use the normal distribution to approximate the probability of

(a) $X = 130$

(b) $X \le 150$

(c) $137 \le X \le 152$

6.43 State whether or not the normal approximation to the binomial is appropriate in each of the following situations.

(a) $n = 90, \quad p = .24$

(b) $n = 100, p = .03$

(c) $n = 120, p = .98$

(d) $n = 61, \quad p = .40$

6.44 State whether or not the normal approximation to the binomial is appropriate in each of the following situations.

(a) $n = 500, p = .23$

(b) $n = 10, \quad p = .40$

(c) $n = 300, p = .02$

(d) $n = 150, p = .97$

(e) $n = 100, p = .71$

6.45 Copy Figure 16 and add the standard score scale $z = (x - np)/\sqrt{np(1 - p)}$ underneath the x-axis for $n = 5, 12, 25$. Notice how the distributions center on zero and most of the probability lies between $z = -2$ and $z = 2$.

6.46 The median age of residents of the United States is 35.6 years. If a survey of 200 residents is taken, approximate the probability that at least 110 will be under 35.6 years of age.

6.47 The unemployment rate in a city is 7.9%. A sample of 300 persons is selected from the labor force. Approximate the probability that

(a) Less than 18 unemployed persons are in the sample

(b) More than 30 unemployed persons are in the sample

6.48 A survey reports that 96% of the people think that violence has increased in the past five years. Out of a random sample of 50 persons, 48 expressed the opinion that citizens have become more violent in the past five years. Does the normal approximation seem appropriate for X = the number of persons who expressed the opinion that citizens have become more violent in the past five years? Explain.

6.49 According to the U. S. Statistical Abstract 2009, about 25% of persons age 18–24 participated in charity work in the past year. Among a sample of 64 persons in this age group, find the probability that 20 or more participated in charity work.

6.50 The weekly amount spent by a small company for in-state travel has approximately a normal distribution with mean \$1450 and standard deviation \$220. What is the probability that the actual expenses will exceed \$1560 in 20 or more weeks during the next year?

6.51 With reference to Exercise 6.50, calculate the probability that the actual expenses would exceed \$1500 for between 18 and 24 weeks, inclusive during the next year.

6.52 In a large midwestern university, 30% of the students live in apartments. If 200 students are randomly selected, find the probability that the number of them living in apartments will be between 55 and 70 inclusive.

6.53 According to a study of mobility, 33% of U.S. residents in the age group 20 to 24 years moved to different housing in 2002 from where they lived in 2001. (Based on *Statistical Abstract of the U.S. 2003*, Table 34.) If the same percentage holds today, give the approximate probability that in a random sample of 100 residents 20 to 24 years old, there will be 39 or more persons who have moved in the past year.

6.54 Suppose that 20% of the trees in a forest are infested with a certain type of parasite.

(a) What is the probability that, in a random sample of 300 trees, the number of trees having the parasite will be between 49 and 71 inclusive?

*(b) After sampling 300 trees, suppose that 72 trees are found to have the parasite. Does this provide strong evidence that the population proportion is higher than 20%? Base your answer on $P[X \geq 72]$ when 20% are infested.

*6. CHECKING THE PLAUSIBILITY OF A NORMAL MODEL

Does a normal distribution serve as a reasonable model for the population that produced the sample? One reason for our interest in this question is that many commonly used statistical procedures require the population to be nearly normal. If a normal distribution is tentatively assumed to be a plausible model, the investigator must still check this assumption once the sample data are obtained.

Although they involve subjective judgment, graphical procedures prove most helpful in detecting serious departures from normality. Histograms can be inspected for lack of symmetry. The thickness of the tails can be checked for conformance with the normal by comparing the proportions of observations in the intervals $(\bar{x} - s, \bar{x} + s)$, $(\bar{x} - 2s, \bar{x} + 2s)$, and $(\bar{x} - 3s, \bar{x} + 3s)$ with those suggested by the empirical guidelines for the bell-shaped (normal) distribution.

A more effective way to check the plausibility of a normal model is to construct a special graph, called a **normal-scores plot,** of the sample data. In order to describe this method, we will first explain the meaning of normal scores, indicate how the plot is constructed, and then explain how to interpret the plot. For an easy explanation of the ideas, we work with a small sample size. In practical applications, at least 15 or 20 observations are needed to detect a meaningful pattern in the plot.

The term **normal scores** refers to an idealized sample from the standard normal distribution—namely, the z values that divide the standard normal distribution into

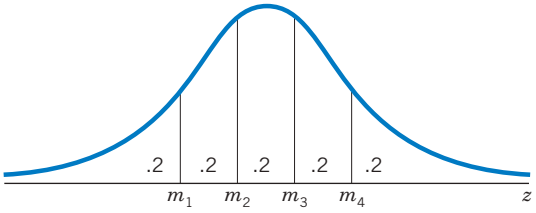


Figure 19 The $N(0, 1)$ distribution and the normal scores for $n = 4$.

equal-probability intervals. For purposes of discussion, suppose the sample size is $n = 4$. Figure 19 shows the standard normal distribution where four points are located on the z axis so the distribution is divided into five segments of equal probability $\frac{1}{5} = .2$. These four points, denoted by m_1, m_2, m_3, and m_4, are precisely the normal scores for a sample of size $n = 4$. Using Appendix B, Table 3, we find that

$$\begin{aligned} m_1 &= -.84 \\ m_2 &= -.25 \\ m_3 &= .25 \\ m_4 &= .84 \end{aligned}$$

A normal-scores plot allows us to visually assess how well a sample mimics the idealized normal sample. To construct a normal-scores plot:

1. Order the sample data from smallest to largest.
2. Obtain the normal scores.
3. Pair the ith largest observation with the ith largest normal score and plot the pairs in a graph.

Example 11 Making a Normal-Scores Plot for Sample Size 4

Suppose a random sample of size 4 has produced the observations 68, 82, 44, and 75. Construct a normal-scores plot.

SOLUTION The ordered observations and the normal scores are shown in Table 2, and the normal-scores plot of the data is given in Figure 20.

TABLE 2 Normal Scores

Normal Scores	Ordered Sample
$m_1 = -.84$	44
$m_2 = -.25$	68
$m_3 = .25$	75
$m_4 = .84$	82

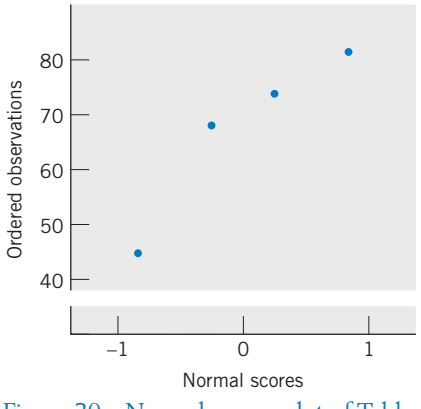


Figure 20 Normal-scores plot of Table 2 data.

INTERPRETATION OF THE PLOT

How does the normal-scores plot of a data set help in checking normality? To explain the main idea, we continue our discussion with the data of Example 11. Let μ and σ denote the mean and standard deviation of the population from which the sample was obtained. The normal scores that are the idealized z observations can then be converted to the x scale by the usual relation $x = \mu + \sigma z$. The actual x observations and the corresponding idealized observations are given in Table 3. If the population were indeed normal, we would expect the two columns of Table 3 to be close. In other words, a plot of the observed x values versus the normal scores would produce a straight-line pattern, where the intercept of the line would indicate the value of μ and the slope of the line would indicate σ.

TABLE 3 Idealized Sample

Observed x Values	Idealized x Values
44	$\mu + \sigma\, m_1$
68	$\mu + \sigma\, m_2$
75	$\mu + \sigma\, m_3$
82	$\mu + \sigma\, m_4$

A straight line pattern in a normal-scores plot supports the plausibility of a normal model. A curve appearance indicates a departure from normality.

A normal-scores plot is easily obtained using software packages. The next example illustrates a straight-line pattern consistent with the assumption that the population is normal.

Example 12 A Normal Scores Plot That Looks Normal

Consider the data on the growth of salmon in Table D.7 of the Data Bank. We plot the freshwater growth of female salmon measured in terms of width of growth rings in hundredths of an inch. Create a normal-scores plot of these data and comment on the pattern.

SOLUTION We use MINITAB software as described in Exercise 6.80 to make the normal scores plot in Figure 21.

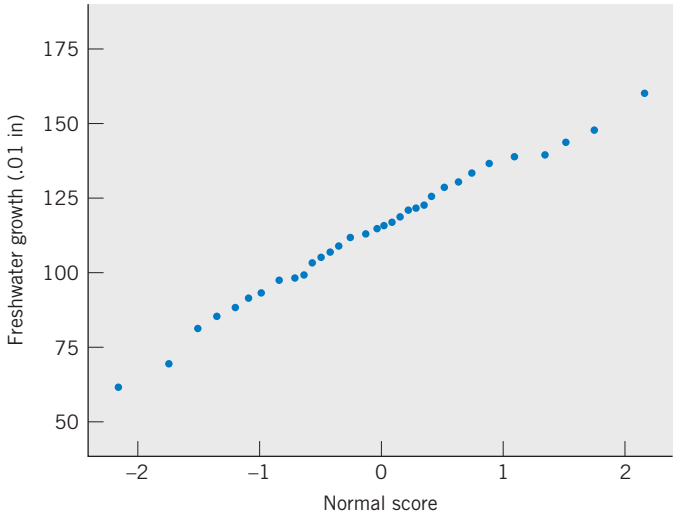


Figure 21 A normal-scores plot of the female salmon growth in freshwater.

Notice that the plot conforms quite well to the straight-line pattern expected for normal observations. The assumption of a normal distribution seems quite reasonable.

The MINITAB package that produced Figure 21 uses one of the many slight variants of the normal scores above but the plots are similar if the sample size is greater than 20.

*7. TRANSFORMING OBSERVATIONS TO ATTAIN NEAR NORMALITY

A valid application of many powerful techniques of statistical inference, especially those suited to small or moderate samples, requires that the population distribution be reasonably close to normal. When the sample measurements appear to have been taken from a population that departs drastically from normality, an appropriate conversion to a new variable may bring the distribution close to normal. Efficient techniques can then be safely applied to the converted data, whereas their application to the original data would have been questionable. Inferential methods requiring the assumption of normality are discussed in later chapters. The goal of our discussion here is to show how a transformation can improve the approximation to a normal distribution.

There is no rule for determining the best transformation in a given situation. For any data set that does not have a symmetric histogram, we consider a variety of transformations.

Some Useful Transformations

Make large values larger:	Make large values smaller:
$x^3, \quad x^2$	$\sqrt{x}, \quad \sqrt[4]{x}, \quad \log_e x, \quad \dfrac{1}{x}$

You may recall that $\log_e x$ is the natural logarithm. Fortunately, computers easily calculate and order the transformed values, so that several transformations in a list can be quickly tested. Note, however, that the observations must be positive if we intend to use $\sqrt{x}$, $\sqrt[4]{x}$, and $\log_e x$.

The selection of a good transformation is largely a matter of trial and error. If the data set contains a few numbers that appear to be detached far to the right, $\sqrt{x}$, $\sqrt[4]{x}$, and $\log_e x$, or negative powers that would pull these stragglers closer to the other data points should be considered.

Example 13 A Transformation to Improve Normality

A forester records the volume of timber, measured in cords, for 49 plots selected in a large forest. The data are given in Table 4 and the corresponding histogram appears in Figure 22*a*. The histogram exhibits a long tail to the right, so it is reasonable to consider the transformations $\sqrt{x}$, $\sqrt[4]{x}$, $\log_e x$, and $1/x$. Transform the data to near normality.

SOLUTION The most satisfactory result, obtained with

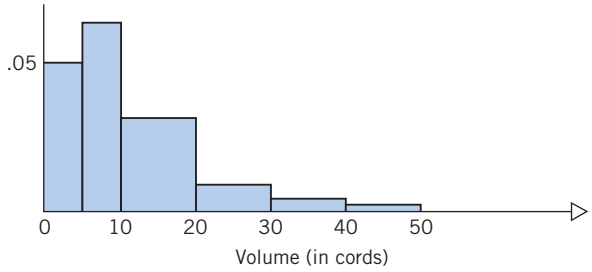
$$\text{Transformed data} = \sqrt[4]{\text{Volume}}$$

is illustrated in Table 5 and Figure 22*b*. The latter histogram more nearly resembles a symmetric bell-shaped pattern expected for normal populations.

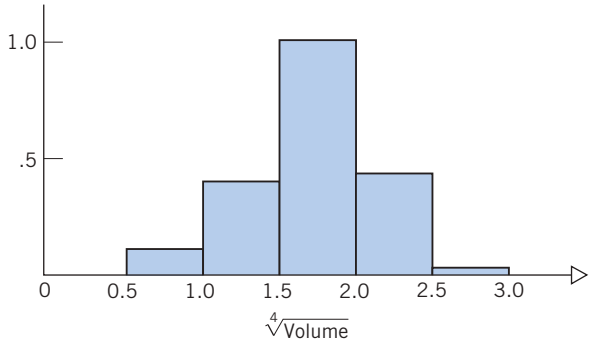
TABLE 4 Volume of Timber in Cords

39.3	14.8	6.3	.9	6.5
3.5	8.3	10.0	1.3	7.1
6.0	17.1	16.8	.7	7.9
2.7	26.2	24.3	17.7	3.2
7.4	6.6	5.2	8.3	5.9
3.5	8.3	44.8	8.3	13.4
19.4	19.0	14.1	1.9	12.0
19.7	10.3	3.4	16.7	4.3
1.0	7.6	28.3	26.2	31.7
8.7	18.9	3.4	10.0	

Courtesy of Professor Alan Ek.



(*a*)



(*b*)

Figure 22 An illustration of the transformation technique. (*a*) Histogram of timber volume. (*b*) Histogram of $\sqrt[4]{\text{Volume}}$.

TABLE 5 The Transformed Data $\sqrt[4]{\text{Volume}}$

2.50	1.96	1.58	.97	1.60
1.37	1.70	1.78	1.07	1.63
1.57	2.03	2.02	.91	1.68
1.28	2.26	2.22	2.05	1.34
1.64	1.60	1.51	1.70	1.56
1.37	1.70	2.59	1.70	1.91
2.10	2.09	1.94	1.17	1.86
2.11	1.79	1.36	2.02	1.44
1.00	1.66	2.31	2.26	2.37
1.72	2.09	1.36	1.78	

HOW MUCH TIMBER IS IN THIS FOREST?
The volume of timber available for making lumber can only be estimated by sampling the number of trees in randomly selected plots within the forest. The distribution of tree size must also be taken into account. © Anthony Baggett/Stockphoto.

USING STATISTICS WISELY

1. A sketch of the bell-shaped normal curve and the area of interest can prevent blunders when determining probabilities and percentiles.
2. Never apply the normal approximation to the binomial, treating

$$Z = \frac{X - np}{\sqrt{np(1 - p)}}$$

as standard normal, when the expected number of successes (or failures) is too small. That is, when either

$$np \quad \text{or} \quad n(1 - p) \quad \text{is 15 or less}$$

3. Do not just assume that data come from a normal distribution. When there are at least 20 to 25 observations, it is good practice to construct a normal-scores plot to check this assumption.

KEY IDEAS AND FORMULAS

The probability distribution for a continuous random variable X is specified by a **probability density curve.** The function that specifies this curve is called a **probability density function.** It can be symmetric about the mean of X or it can be **skewed,** meaning that it has a long tail to either the left or the right.

The probability that X lies in an interval from a to b is determined by the area under the probability density curve between a and b. The total area under the curve is 1, and the curve is never negative.

The population **100 p-th percentile** is an x value that has probability p to its left and probability $1 - p$ to its right.

When X has mean μ and standard deviation σ, the **standardized variable**

$$Z = \frac{X - \mu}{\sigma}$$

has mean 0 and standard deviation 1.

The **normal distribution** has a symmetric bell-shaped curve centered at the mean. The intervals extending one, two, and three standard deviations around the mean contain the probabilities .683, .954, and .997, respectively.

If X is normally distributed with mean μ and standard deviation σ, then

$$Z = \frac{X - \mu}{\sigma}$$

has the **standard normal distribution.**

When the number of trials n is large and the success probability p is not too near 0 or 1, the binomial distribution is well approximated by a normal distribution with mean np and sd $= \sqrt{np(1 - p)}$. Specifically, the probabilities for a binomial variable X can be approximately calculated by treating

$$Z = \frac{X - np}{\sqrt{np(1 - p)}}$$

as standard normal. For a moderate number of trials n, the approximation is improved by appropriately adjusting by $\frac{1}{2}$ called a **continuity correction.**

The **normal scores** are an ideal sample from a standard normal distribution. Plotting each ordered observation versus the corresponding normal score creates a **normal-scores plot,** which provides a visual check for possible departures from a normal distribution.

Transformation of the measurement scale often helps to convert a long-tailed distribution to one that resembles a normal distribution.

TECHNOLOGY

Probability and Percentiles for the Standard Normal and General Normal Distribution

MINITAB

MINITAB uses the same steps for calculations with the standard normal and cases of other means and standard deviations. We illustrate with the calculation of $P[X \leq 8]$ when X is normal with mean 5 and standard deviation 12.5.

Calc > Probability Distributions > Normal.
Select **Cumulative Probability.** Type *5* in **Mean.**
Type *12.5* in **Standard deviation.**
Select **Input Constant** and type *8*. Click **OK.**

The default settings Mean 0 and Standard deviation 1 simplify the steps for obtaining standard normal probabilities.

The inverse problem of finding b so that $P[X \leq b] = a$, where a is a specified probability, is illustrated with finding b so that $P[X \leq b] = .9700$ when X is normal with mean 5 and standard deviation 12.5.

Calc > Probability Distributions > Normal.
Select **Inverse Probability.** Type *5* in **Mean.**
Type *12.5* in **Standard deviation.**
Select **Input Constant** and type *.9700*. Click **OK.**

EXCEL

EXCEL uses the function *NORMSDIST* for standard normal probabilities and *NORMDIST* for a general normal distribution. We illustrate with the calculation of $P[X \leq 8]$ when X is normal with mean 5 and standard deviation 12.5.

Select the f_x icon, or select **Insert** and then **Function.**
Choose **Statistical** and then **NORMDIST.** Click **OK.**
Type *8* in **X,** *5* in **Mean,** *12.5* in **Standard_dev** and *True* in **Cumulative.** Click **OK.**

The inverse problem of finding b so that $P[X \leq b] = a$, where a is a specified probability, is illustrated with finding b so that $P[X \leq b] = .9700$ when X is normal with mean 5 and standard deviation 12.5.

Select the f_x icon, or select **Insert** and then **Function.**
Choose **Statistical,** and then **NORMINV.** Click **OK.**
Type *.9700* in **Probability,** *5* in **Mean,** *12.5* in **Standard_dev.** Click **OK.**

To solve the standard normal inverse problem replace **NORMINV** by **NORMSINV.**

TI-84/-83 PLUS

We illustrate with the calculation of $P[X \leq 8]$ when X is normal with mean 5 and standard deviation 12.5.

In the Home screen, press **2nd *VARS***

From the *DISTR* menu, select **2: Normalcdf(.**

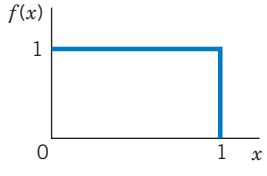
Type entries to obtain **Normalcdf(***—1E99, 8, 5, 12.5***),**
Then press **ENTER.**

The inverse problem of finding b so that $P[X \leq b] = a$, where a is a specified probability, is illustrated with finding b so that $P[X \leq b] = .9700$ when X is normal with mean 5 and standard deviation 12.5.

In the Home screen, press **2nd *VARS.***
From the *DISTR* mean, select **3: InvNorm(.**
Type entries to obtain **InvNorm(***.9700, 5, 12.5***).**
Then press **ENTER.**

8. REVIEW EXERCISES

6.55 Determine (a) the median and (b) the quartiles for the distribution shown in the following illustration.



6.56 For X having the density in Exercise 6.55, find (a) $P[X > .7]$ (b) $P[.5 \leq X \leq .7]$ and (c) $P[.5 < X < .7]$.

6.57 In the context of the height of red pine seedlings presented at the front of the chapter, describe the reasoning that leads from a histogram to the concept of a probability density curve. (Think of successive histograms based on 100 heights, 500 heights, 1456 heights, and then an unlimited number.)

6.58 For a standard normal random variable Z, find
(a) $P[Z < 1.26]$
(b) $P[Z > 1.245]$
(c) $P[.61 < Z < 1.92]$
(d) $P[-1.47 < Z < 1.055]$

6.59 For the standard normal distribution, find the value z such that
(a) Area to its left is .0838
(b) Area to its left is .047
(c) Area to its right is .2611
(d) Area to its right is .12

6.60 Find the 20th, 40th, 60th, and 80th percentiles of the standard normal distribution.

6.61 If Z is a standard normal random variable, what is the probability that

(a) Z exceeds .62?

(b) Z lies in the interval $(-1.40, 1.40)$?

(c) $|Z|$ exceeds 3.0?

(d) $|Z|$ is less than 2.0?

6.62 According to Example 12, a normal distribution with mean 115 and standard deviation 22 hundredths of an inch describes variation in female salmon growth in fresh water.

(a) If a newly caught female salmon has growth 108, what is the corresponding standardized score?

(b) If a standardized score is $-.8$, what is the growth measurement?

(c) Find the interval of standardized scores corresponding to the growth measurements 105 to 128.

(d) Find the interval of growth measurements corresponding to the standardized scores of -1.5 to 1.5.

6.63 The bell-shaped histogram for the heights of three-year-old red pine seedlings on page 222 is consistent with the assumption of a normal distribution having mean = 280 and sd = 58 millimeters. Let X denote the height, at three years of age, of the next red pine that will be measured. Find

(a) $P[X < 337]$

(b) $P[X < 240]$

(c) $P[X > 230]$

(d) $P[X > 90]$

(e) $P[235 < X < 335]$

(f) $P[305 < X < 405]$

6.64 If X has a normal distribution with $\mu = 100$ and $\sigma = 5$, find b such that

(a) $P[X < b] = .6700$

(b) $P[X > b] = .0110$

(c) $P[|X - 100| < b] = .966$

6.65 Suppose that a student's verbal score X from next year's Graduate Record Exam can be considered an observation from a normal population having mean 499 and standard deviation 120. Find

(a) $P[X > 600]$

(b) 90th percentile of the distribution

(c) Probability that the student scores below 400

6.66 The lifting capacities of a class of industrial workers are normally distributed with mean 65 pounds and standard deviation 8 pounds. What proportion of these workers can lift an 80-pound load?

6.67 The bonding strength of a drop of plastic glue is normally distributed with mean 100 pounds and standard deviation 8 pounds. A broken plastic strip is repaired with a drop of this glue and then subjected to a test load of 90 pounds. What is the probability that the bonding will fail?

6.68 ***Grading on a curve.*** The scores on an examination are normally distributed with mean $\mu = 70$ and standard deviation $\sigma = 8$. Suppose that the instructor decides to assign letter grades according to the following scheme (left endpoint included).

Scores	Grade
Less than 58	F
58 to 66	D
66 to 74	C
74 to 82	B
82 and above	A

Find the percentage of students in each grade category.

6.69 Suppose the duration of trouble-free operation of a new robotic vacuum cleaner is normally distributed with mean 750 days and standard deviation 100 days.

(a) What is the probability that the vacuum cleaner will work for at least two years without trouble?

(b) The company wishes to set the warranty period so that no more than 10% of the vacuum cleaners would need repair services while under warranty. How long a warranty period must be set?

6.70 Suppose the amount of a popular sport drink in bottles leaving the filling machine has a normal

distribution with mean 101.5 milliliters (ml) and standard deviation 1.6 ml.

(a) If the bottles are labeled 100 ml, what proportion of the bottles contain less than the labeled amount.

(b) If only 5% of the bottles have contents that exceed a specified amount v, what is the value of v?

6.71 Suppose the amount of sun block lotion in plastic bottles leaving a filling machine has a normal distribution. The bottles are labeled 300 milliliters (ml) but the actual mean is 302 ml and the standard deviation is 2 ml.

(a) What is the probability that an individual bottle will contain less than 299 ml?

(b) If only 5% of the bottles have contents that exceed a specified amount v, what is the value of v?

*6.72 ***A property of the normal distribution.*** Suppose the random variable X is normally distributed with mean μ and standard deviation σ. If Y is a linear function of X—that is, $Y = a + bX$, where a and b are constants—then Y is also normally distributed with

$$\text{Mean} = a + b\mu$$

$$\text{sd} = |b|\sigma$$

For instance, if X is distributed as $N(25, 2)$ and $Y = 7 - 3X$, then the distribution of Y is normal with Mean $= 7 - 3(25) = -68$ and sd $= |-3| \times 2 = 6$.

(a) At the "low" setting of a water heater, the temperature X of water is normally distributed with Mean $= 102°$F and sd $= 4°$F. If Y refers to the temperature measurement in the centigrade scale, that is, $Y = \frac{5}{9}(X - 32)$, what is the distribution of Y?

(b) Referring to part (a), find the probability of $[35 \le Y \le 42]$.

Remark: The relation between a general normal and the standard normal is only a special case of this property. Specifically, the standardized variable Z is the linear function.

$$Z = \frac{X - \mu}{\sigma} = -\frac{\mu}{\sigma} + \frac{1}{\sigma}X$$

where Z has

$$\text{Mean} = -\frac{\mu}{\sigma} + \frac{1}{\sigma}\mu = 0$$

$$\text{sd} = \frac{1}{\sigma}\sigma = 1$$

6.73 Let X denote the number of successes in n Bernoulli trials with a success probability of p.

(a) Find the exact probabilities of each of the following:

(i) $X \le 7$ when $n = 25, p = .4$

(ii) $11 \le X \le 16$ when $n = 20, p = .7$

(iii) $X \ge 9$ when $n = 16, p = .5$

(b) Use a normal approximation for each situation in part (a).

6.74 It is known from past experience that 7% of the tax bills are paid late. If 20,000 tax bills are sent out, find the probability that:

(a) Less than 1350 are paid late.

(b) 1480 or more are paid late.

6.75 A particular program, say, program A, previously drew 30% of the television audience. To determine if a recent rescheduling of the programs on a competing channel has adversely affected the audience of program A, a random sample of 400 viewers is to be asked whether or not they currently watch this program.

(a) If the percentage of viewers watching program A has not changed, what is the probability that fewer than 103 out of a sample of 400 will be found to watch the program?

(b) If the number of viewers of the program is actually found to be less than 103, will this strongly support the suspicion that the population percentage has dropped? Use your calculation from part (a).

6.76 The number of successes X has a binomial distribution. State whether or not the normal

approximation is appropriate in each of the following situations: (a) $n = 400, p = .23$ (b) $n = 20, p = .03$ (c) $n = 90, p = .98$.

6.77 Because 10% of the reservation holders are "no-shows," a U.S. airline sells 400 tickets for a flight that can accommodate 370 passengers.

(a) Find the probability that one or more reservation holders will not be accommodated on the flight.

(b) Find the probability of fewer than 350 passengers on the flight.

6.78 On a Saturday afternoon, 147 customers will be observed during check-out and the number paying by card, credit or debit, will be recorded. Records from the store suggest that 43% of customers pay by card. Approximate the probability that:

(a) More than 60 will pay by card.

(b) Between 60 and 70, inclusive, will pay by card.

6.79 In all of William Shakespeare's works, he used 884,647[4] different words. Of these, 14,376 appeared only once. In 1985 a 429-word poem was discovered that may have been written by Shakespeare. To keep the probability calculations simple, assume that the choices between a new word and one from the list of 884,647 are independent for each of the 429 words. Approximate the probability that a new word will not be on the list, by the relative frequency of words used once.

(a) Find the expected number of new words in the poem.

(b) Use the normal approximation to the binomial to determine the probability of finding 12 or more new words in the poem. Use the continuity correction.

(c) Use the normal approximation to the binomial to determine the probability of finding 2 or fewer new words in the poem. Use the continuity correction.

(d) Use the normal approximation to the binomial to determine the probability of finding more than 2 but less than 12 new words in the poem. On the basis of your answer, decide if 9 = actual number of new words not in the list is consistent with Shakespeare having written the poem or if it contradicts this claim.

[4]See R. Thisted and B. Efron. "Did Shakespeare write a newly-discovered poem?" *Biometrika*, **74,** (1987) pp. 445–455.

The Following Exercises Require a Computer

6.80 ***Normal-scores plot.*** Use a computer program to make a normal-scores plot for the volume of timber data in Table 4. Comment on the departure from normality displayed by the normal-scores plot.

We illustrate a normal-scores plot using MINITAB. With the data set in column 1, the MINITAB commands

Calc > Calculator.

Type *C2* in **Store.** Type *NSCOR*(*C1*) in **Expression.** Click **OK.**

Graph > Scatterplot. Select *Simple.* Type *C1* under **Y** variables and *C2* under **X** variables. Click **OK.**

will create a normal-scores plot for the observations in *C1*. (MINITAB uses a variant of the normal scores, m_i, that we defined)

*6.81 Use MINITAB or another package program to make a normal-scores plot of the malt extract data in Table D.8 of the Data Bank.

*6.82 Use MINITAB or another package program to make a normal-scores plot of the computer anxiety scores in Table D.4 of the Data Bank.

*6.83 ***Transformations and normal-scores plots.*** The MINITAB computer language makes it possible to easily transform data. With the data already set in column 1, the commands

Dialog box:

Calc > Calculator. Type C2 in **Store.** Type *LOGE*(*C1*) in **Expression.** Click **OK.**

Calc > Calculator. Type C3 in **Store.** Type *SQRT*(*C1*) in **Expression.** Click **OK.**

Calc > Calculator. Type C4 in **Store.** Type *SQRT*(*C2*) in **Expression.** Click **OK.**

will place the natural logarithm $\log_e x$ in *C2*, $\sqrt{x}$ in *C3*, and $x^{1/4}$ in *C4*. Normal-scores plots can then be constructed as in Exercise 6.80.

Refer to the lightning data in Exercise 2.121 of Chapter 2. Make a normal-scores plot of the

(a) original observations.

(b) fourth root of the original observations.

(c) Comment on the quality of the approximation by a normal distribution in both cases.