Chapter

5 Sampling Distributions

Objectives

In this chapter we will develop the idea of a sampling distribution, which is central to classical statistical inference. In particular, we will

- describe sampling distributions.
- show how the sample size is related to the accuracy of the sample mean.
- explore the Central Limit Theorem.
- demonstrate how the normal distribution can be used to approximate the binomial distribution.

5.1 Basic Ideas

An important goal of data analysis is to distinguish between features of the data that reflect real biological facts and features that may reflect only chance effects. As explained in Sections 1.3 and 2.8, the random sampling model provides a framework for making this distinction. The underlying reality is visualized as a population, the data are viewed as a random sample from the population, and chance effects are regarded as sampling error—that is, discrepancy between the sample and the population.

In this chapter we develop the theoretical background that will enable us to place specific limits on the degree of sampling error to be expected in a study. (Although in Chapter 1 we distinguished between an experimental study and an observational study, for the present discussion we will call any scientific investigation a *study*.) As in earlier chapters, we continue to confine the discussion to the simple context of a study with only one group (one sample).

Sampling Variability

The variability among random samples from the same population is called **sampling variability**. A probability distribution that characterizes some aspect of sampling variability is termed a **sampling distribution**. Usually a random sample will resemble the population from which it came. Of course, we have to expect a certain amount of discrepancy between the sample and the population. A sampling distribution tells us how close the resemblance between the sample and the population is likely to be.

In this chapter we will discuss several aspects of sampling variability and study an important sampling distribution. From this point forward, we will assume that the sample size is a negligibly small fraction of the population size. This assumption simplifies the theory because it guarantees that the process of drawing the sample does not change the population composition in any appreciable way.

The Meta-Study

According to the random sampling model, we regard the data in a study as a random sample from a population. Generally we obtain only a single random sample, which comes from a very large population. However, to visualize sampling variability we must broaden our frame of reference to include not merely one sample, but all the possible samples that might be drawn from the population. This wider frame of reference we will call the **meta-study**. A meta-study consists of indefinitely many repetitions, or replications, of the same study.* Thus, if the study consists of drawing a random sample of size n from some population, the corresponding meta-study involves drawing *repeated* random samples of size n from the same population. The process of repeated drawing is carried on indefinitely, with the members of each sample being replaced before the next sample is drawn. The study and the meta-study are schematically represented in Figure 5.1.1.

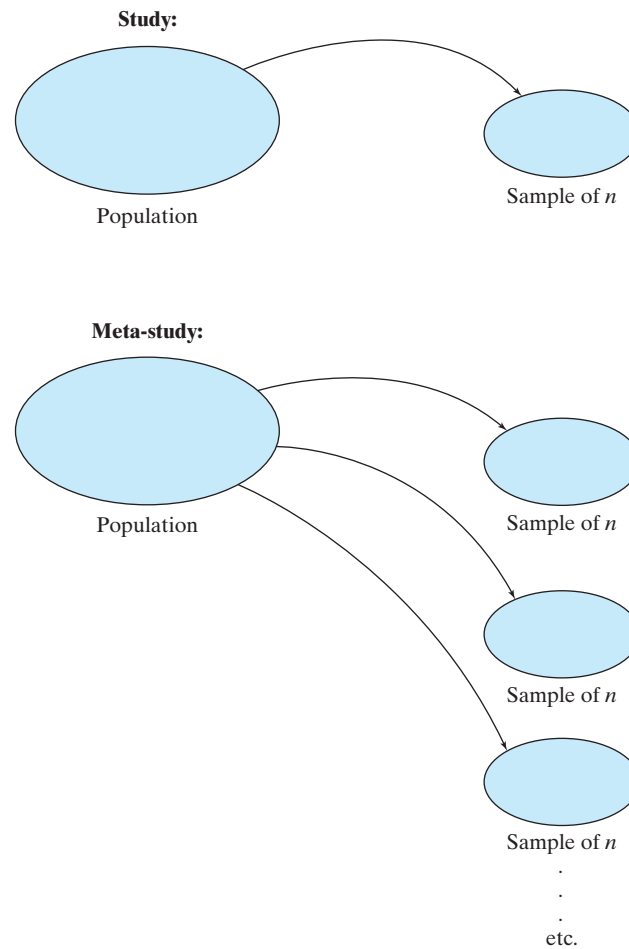


Figure 5.1.1 Schematic representation of study and meta-study

*The term *meta-study* is not a standard term. It is unrelated to the term *meta-analysis*, which denotes a particular type of statistical analysis.

The following two examples illustrate the notion of a meta-study.

Example 5.1.1 Rat Blood Pressure A study consists of measuring the change in blood pressure in each of $n = 10$ rats after administering a certain drug. The corresponding meta-study would consist of repeatedly choosing groups of $n = 10$ rats from the same population and making blood pressure measurements under the same conditions. ■

Example 5.1.2 Bacterial Growth A study consists of observing bacterial growth in $n = 5$ petri dishes that have been treated identically. The corresponding meta-study would consist of repeatedly preparing groups of five petri dishes and observing them in the same way. ■

Note that a meta-study is a theoretical construct rather than an operation that is actually performed by a researcher.

The meta-study concept provides a link between sampling variability and probability. Recall from Chapter 3 that the probability of an event can be interpreted as the long-run relative frequency of occurrence of the event. Choosing a random sample is a chance operation; the meta-study consists of many repetitions of this chance operation, and so *probabilities concerning a random sample can be interpreted as relative frequencies in a meta-study.* Thus, the meta-study is a device for explicitly visualizing a sampling distribution: The sampling distribution describes the variability, for a chosen statistic, among the many random samples in a meta-study.

We consider a small (and artificial) example to illustrate the idea of a sampling distribution.

Example 5.1.3 Knee Replacement Consider a population of women age 65 to 75 who are experiencing pain in their knees and are candidates for knee replacement surgery. A woman might have replacement surgery done on one knee at a cost of $35,000, both knees at a cost of $60,000 (a "double replacement," which is less expensive than two single replacements), or neither knee. Consider the perspective of an insurance company regarding a sample of $n = 3$ women it insures: What is the total cost for treating these three? The smallest the total could be is zero—if all three women skip surgery—while the largest possible cost would be $180,000—if all three women have double replacements. To keep things relatively simple, suppose that one-fourth of women age 65 to 75 elect a double knee replacement, one-half elect a single knee replacement, and one-fourth choose not to have surgery.

The complete list of possible samples is given in Table 5.1.1, along with the sample total (in thousands of dollars) in each case and the probability of each case arising. For example, the probability that all three women skip surgery ("None, None, None") is $(1/4) \times (1/4) \times (1/4) = 1/64$ while the probability that the first two women skip surgery and the third has a single knee operation ("None, None, Single") is $(1/4) \times (1/4) \times (2/4) = 2/64$. There are 10 possible values for the sample total: 0, 35, 60, 70, 95, 105, 120, 130, 155, and 180. The first and third columns of Table 5.1.2 give the sampling distribution of the sample total by combining the samples that yield the same total and summing their probabilities. For example, there are three ways for the total to be 70, each of which has probability 4/64; these sum to 12/64.

Table 5.1.1 Total knee replacement costs for all possible samples of size $n = 3$

Sample	Costs (in units of $1,000)	Sample total	Probability
None, None, None	0,0,0	0	1/64
None, None, Single	0,0,35	35	2/64
None, None, Double	0,0,60	60	1/64
None, Single, None	0,35,0	35	2/64
None, Single, Single	0,35,35	70	4/64
None, Single, Double	0,35,60	95	2/64
None, Double, None	0,60,0	60	1/64
None, Double, Single	0,60,35	95	2/64
None, Double, Double	0,60,60	120	1/64
Single, None, None	35,0,0	35	2/64
Single, None, Single	35,0,35	70	4/64
Single, None, Double	35,0,60	95	2/64
Single, Single, None	35,35,0	70	4/64
Single, Single, Single	35,35,35	105	8/64
Single, Single, Double	35,35,60	130	4/64
Single, Double, None	35,60,0	95	2/64
Single, Double, Single	35,60,35	130	4/64
Single, Double, Double	35,60,60	155	2/64
Double, None, None	60,0,0	60	1/64
Double, None, Single	60,0,35	95	2/64
Double, None, Double	60,0,60	120	1/64
Double, Single, None	60,35,0	95	2/64
Double, Single, Single	60,35,35	130	4/64
Double, Single, Double	60,35,60	155	2/64
Double, Double, None	60,60,0	120	1/64
Double, Double, Single	60,60,35	155	2/64
Double, Double, Double	60,60,60	180	1/64

The second column of Table 5.1.2 shows the sample mean (rounded to one decimal place) so that the last two columns of the table give the sampling distribution of the sample mean. These two distributions, shown graphically in Figure 5.1.2, are scaled versions of each other. An insurance company might speak in terms of total cost, but this is equivalent to looking at average cost. ■

Relationship to Statistical Inference

Knowing a sampling distribution allows one to make probability statements about possible samples. For example, for the setting in Example 5.1.3 the insurance company might ask, What is the probability that the total knee replacement costs for a sample of three women will be less than $110,000? We can answer this question by

Table 5.1.2 Sampling distribution of total surgery costs for samples of size $n = 3$

Sample total	Sample mean	Probability
0	0.0	1/64
35	11.7	6/64
60	20.0	3/64
70	23.3	12/64
95	31.7	12/64
105	35.0	8/64
120	40.0	3/64
130	43.3	12/64
155	51.7	6/64
180	60.0	1/64

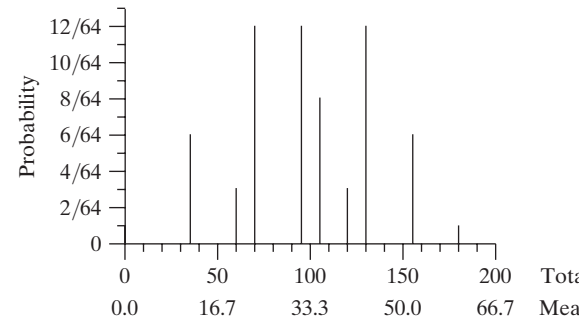


Figure 5.1.2 Graph of the sampling distribution of total surgery costs for samples of size $n = 3$

adding the probabilities of the first six outcomes listed in Table 5.1.2; the sum is 42/64. We will expand upon this idea as we formally develop ideas of statistical inference.

Exercises 5.1.1–5.1.4

5.1.1 Consider taking a random sample of size 3 from the knee replacement population of Example 5.1.3. What is the probability that the total cost for those in the sample will be greater than \$125,000?

5.1.2 Consider taking a random sample of size 3 from the knee replacement population of Example 5.1.3. What is the probability that the total cost for those in the sample will be between \$80,000 and \$125,000?

5.1.3 Consider taking a random sample of size 3 from the knee replacement population of Example 5.1.3. What is the probability that the mean cost for those in the sample will be between \$40,000 and \$100,000?

5.1.4 Consider a hypothetical population of dogs in which there are four possible weights, all of which are equally likely: 42, 48, 52, or 58 pounds. If a sample of size $n = 2$ is drawn from this population, what is the sampling distribution of the total weight of the two dogs selected? That is, what are the possible values for the total and what are the probabilities associated with each of those values?

5.2 The Sample Mean

For a quantitative variable, the sample and the population can be described in various ways—by the mean, the median, the standard deviation, and so on. The natures (e.g., shape, center, spread) of the sampling distributions for these descriptive measures are not all the same. In this section we will focus primarily on the sampling distribution of the sample mean.

The Sampling Distribution of $\overline{Y}$

The sample mean $\bar{y}$ can be used, not only as a description of the data in the sample, but also as an estimate of the population mean μ. It is natural to ask, "How close to μ is $\bar{y}$?" We cannot answer this question for the mean $\bar{y}$ of a particular

sample, but we can answer it if we think in terms of the random sampling model and regard the sample mean as a random variable $\overline{Y}$. The question then becomes: "How close to μ is $\overline{Y}$ *likely* to be?" and the answer is provided by the **sampling distribution of** $\overline{Y}$—that is, the probability distribution that describes sampling variability in $\overline{Y}$.

To visualize the sampling distribution of $\overline{Y}$, imagine the meta-study as follows: Random samples of size n are repeatedly drawn from a fixed population with mean μ and standard deviation σ; each sample has its own mean $\bar{y}$. The variation of the $\bar{y}$'s among the samples is specified by the sampling distribution of $\overline{Y}$. This relationship is indicated schematically in Figure 5.2.1.

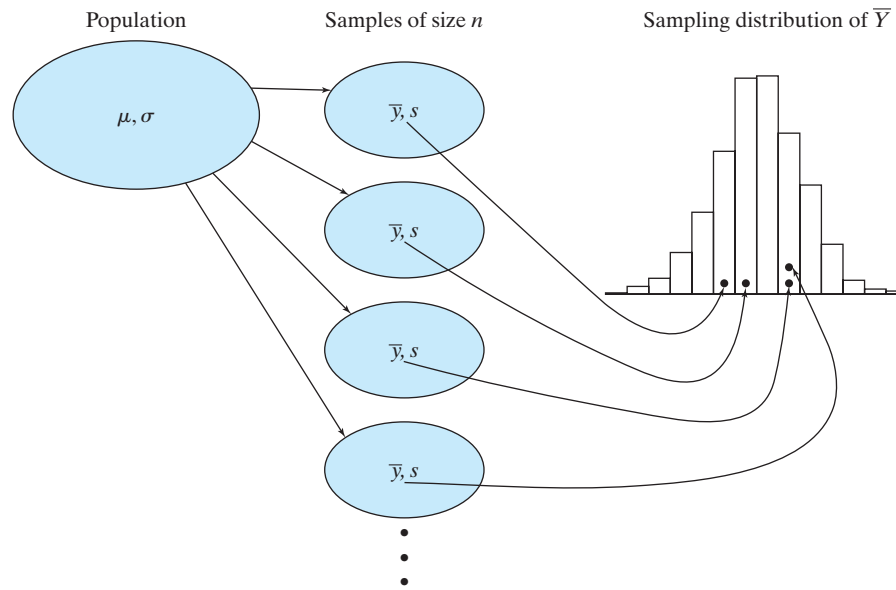


Figure 5.2.1 Schematic representation of the sampling distribution of $\overline{Y}$

When we think of $\overline{Y}$ as a random variable, we need to be aware of two basic facts. The first of these is intuitive: On average, the sample mean equals the population mean. That is, the average of the sampling distribution of $\overline{Y}$ is μ. The second fact is not obvious: The standard deviation of $\overline{Y}$ is equal to the standard deviation of Y divided by the square root of the sample size. That is, the standard deviation of $\overline{Y}$ is $\sigma/\sqrt{n}$.

Example 5.2.1

Serum Cholesterol The serum cholesterol levels of 12- to 14-year-olds follow a normal distribution with mean $\mu = 162$ mg/dl and standard deviation $\sigma = 28$ mg/dl.[1] If we take a random sample, then we expect the sample mean to be near 162, with the means of some samples being larger than 162 and the means of some samples being smaller than 162. As the preceding formula indicates, the amount of variability in the sample mean depends on the variability of cholesterol levels of the population, σ. If the population is very homogeneous (everyone has nearly the same cholesterol value so that σ is small), then samples and hence sample means would all be very similar and thus exhibit low variability. If the population is very heterogenous (σ is large), then samples (and hence sample mean values) would vary more. While researchers have little control over the value of σ, we can control the sample size, n, and n affects the amount of variability in the sample mean. If we take a sample of

size $n = 9$, then the standard deviation of the sample mean is $\frac{28}{\sqrt{9}} = \frac{28}{3} = 9.3$. This means, loosely speaking, that the sample mean, $\overline{Y}$, will vary from one to sample to the next by about 9.3 mg/dl.* If we took larger random samples of size $n = 25$, then the standard deviation of the sample mean would be smaller: $\frac{28}{\sqrt{25}} = \frac{28}{5} = 5.6$, which means that $\overline{Y}$ would vary from one sample to the next by about 5.6. As the sample size goes up, the variability in the sample mean $\overline{Y}$ goes down. ■

We now state as a theorem the basic facts about the sampling distribution of $\overline{Y}$. The theorem can be proved using the methods of mathematical statistics; we will state it without proof. The theorem describes the sampling distribution of $\overline{Y}$ in terms of its mean (denoted by $\mu_{\overline{Y}}$), its standard deviation (denoted by $\sigma_{\overline{Y}}$), and its shape.**

Theorem 5.2.1: The Sampling Distribution of $\overline{Y}$

1. Mean The mean of the sampling distribution of $\overline{Y}$ is equal to the population mean. In symbols,

$$\mu_{\overline{Y}} = \mu$$

2. Standard deviation The standard deviation of the sampling distribution of $\overline{Y}$ is equal to the population standard deviation divided by the square root of the sample size. In symbols,

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$

3. Shape
   (a) If the population distribution of Y is normal, then the sampling distribution of $\overline{Y}$ is normal, regardless of the sample size n.
   (b) *Central Limit Theorem* If n is large, then the sampling distribution of $\overline{Y}$ is approximately normal, even if the population distribution of Y is not normal.

Parts 1 and 2 of Theorem 5.2.1 specify the relationship between the mean and standard deviation of the population being sampled, and the mean and standard deviation of the sampling distribution of $\overline{Y}$. Part 3(a) of the theorem states that, if the observed variable Y follows a normal distribution in the population being sampled, then the sampling distribution of $\overline{Y}$ is also a normal distribution. These relationships are indicated in Figure 5.2.2.

---

*Strictly speaking, the standard deviation measures deviation from the mean, not the difference between consecutive observations.

We are assuming here that the population is infinitely large or, equivalently, that we are sampling with replacement, so that we never exhaust the population. If we sample without replacement from a finite population then an adjustment is needed to get the right value for $\sigma_{\overline{Y}}$. Here $\sigma_{\overline{Y}}$ is given by $\frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$. The term $\sqrt{\frac{N-n}{N-1}}$ is called the **finite population correction factor. Note that if the sample size n is 10% of the population size N, then the correction factor is $\sqrt{\frac{0.9N}{N-1}} \approx 0.95$, so the adjustment is small. Thus, if n is small, in comparison to N, then the finite population correction factor is close to 1 and can be ignored.

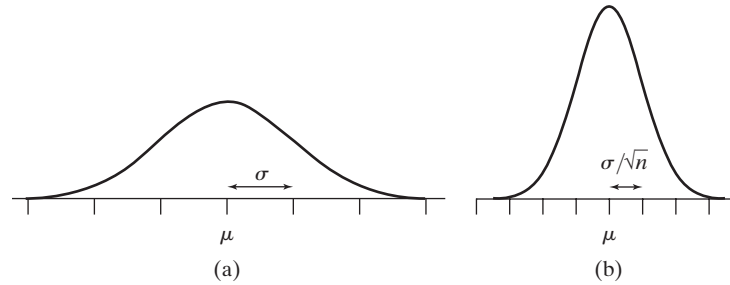


Figure 5.2.2 (a) The population distribution of a normally distributed variable Y; (b) the sampling distribution of $\overline{Y}$ in samples from the population of part (a)

The following example illustrates the meaning of parts 1, 2, and 3(a) of Theorem 5.2.1.

Example 5.2.2

Weights of Seeds A large population of seeds of the princess bean *Phaseotus vulgaris* is to be sampled. The weights of the seeds in the population follow a normal distribution with mean $\mu = 500$ mg and standard deviation $\sigma = 120$ mg.[2] Suppose now that a random sample of four seeds is to be weighed, and let $\overline{Y}$ represent the mean weight of the four seeds. Then, according to Theorem 5.2.1, the sampling distribution of $\overline{Y}$ will be a normal distribution with mean and standard deviation as follows:

$$\mu_{\overline{Y}} = \mu = 500 \text{ mg}$$

and

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{120}{\sqrt{4}} = 60 \text{ mg}$$

Thus, on average the sample mean will equal 500 mg, but the variability from one sample of size 4 to the next sample of size 4 is such that about two-thirds of the time $\overline{Y}$ will be within 60 mg of 500 mg, that is, between $500 - 60 = 440$ mg and $500 + 60 = 560$ mg. Likewise, allowing for 2 standard deviations, we expect that $\overline{Y}$ will be within 120 mg of 500 mg or between $500 - 120 = 380$ mg and $500 + 120 = 620$ mg about 95% of the time. The sampling distribution of $\overline{Y}$ is shown in Figure 5.2.3; the ticks are 1 standard deviation apart. ■

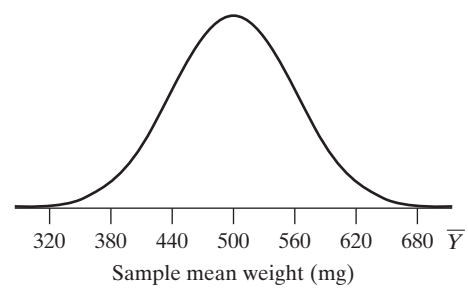


Figure 5.2.3 Sampling distribution of $\overline{Y}$ for Example 5.2.2

The sampling distribution of $\overline{Y}$ expresses the relative likelihood of the various possible values of $\overline{Y}$. For example, suppose we want to know the probability that the mean weight of the four seeds will be greater than 550 mg. This probability is shown as the shaded area in Figure 5.2.4. Notice that the value of $\bar{y} = 550$ must be converted to the Z scale using the standard deviation $\sigma_{\overline{Y}} = 60$, not $\sigma = 120$.

$$z = \frac{\bar{y} - \mu_{\overline{Y}}}{\sigma_{\overline{Y}}} = \frac{550 - 500}{60} = 0.83$$

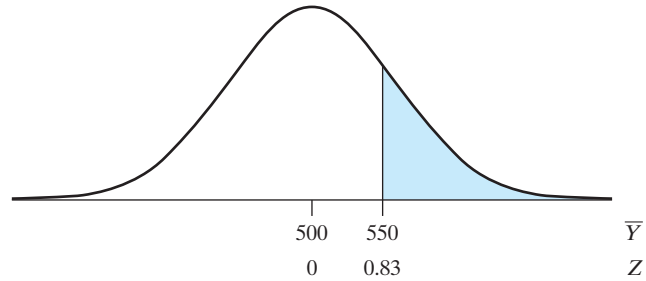


Figure 5.2.4 Calculation of $\Pr\{\overline{Y} > 550\}$ for Example 5.2.2

From Table 3, $z = 0.83$ corresponds to an area of 0.7967. Thus,

$$\Pr\{\overline{Y} > 550\} = \Pr\{Z > 0.83\} = 1 - 0.7967$$
$$= 0.2033 \approx 0.20$$

This probability can be interpreted in terms of a meta-study as follows: If we were to choose many random samples of four seeds each from the population, then about 20% of the samples would have a mean weight exceeding 550 mg.

Part 3(b) of Theorem 5.2.1 is known as the **Central Limit Theorem**. The Central Limit Theorem states that, *no matter what distribution Y may have in the population*,* if the sample size is large enough, then the sampling distribution of $\overline{Y}$ will be approximately a normal distribution.

The Central Limit Theorem is of fundamental importance because it can be applied when (as often happens in practice) the form of the population distribution is not known. It is because of the Central Limit Theorem (and other similar theorems) that the normal distribution plays such a central role in statistics.

It is natural to ask how "large" a sample size is required by the Central Limit Theorem: How large must n be in order that the sampling distribution of $\overline{Y}$ be well approximated by a normal curve? The answer is that the required n depends on the shape of the population distribution. If the shape is normal, any n will do. If the shape is moderately nonnormal, a moderate n is adequate. If the shape is highly nonnormal, then a rather large n will be required. (Some specific examples of this phenomenon are given in the optional Section 5.3.)

Remark We stated in Section 5.1 that the theory of this chapter is valid if the sample size is small compared to the population size. But the Central Limit Theorem is a statement about large samples. This may seem like a contradiction: How can a large sample be a small sample? In practice, there is no contradiction. In a typical biological application, the population size might be 10^6; a sample of size $n = 100$ would be a small fraction of the population but would nevertheless be large enough for the Central Limit Theorem to be applicable (in most situations).

Dependence on Sample Size

Consider the possibility of choosing random samples of various sizes from the same population. The sampling distribution of $\overline{Y}$ will depend on the sample size n in two ways. First, its standard deviation is

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$

*Technically, the Central Limit Theorem requires that the distribution of Y have a standard deviation. In practice this condition is always met.

and this is inversely proportional to $\sqrt{n}$. Second, if the population distribution is not normal, then the *shape* of the sampling distribution of $\overline{Y}$ depends on n, being more nearly normal for larger n. However, if the population distribution is normal, then the sampling distribution of $\overline{Y}$ is always normal, and only the standard deviation depends on n.

The more important of the two effects of sample size is the first: Larger n gives a smaller value of $\sigma_{\overline{Y}}$ and consequently a smaller expected sampling error if $\bar{y}$ is used as an estimate of μ. The following example illustrates this effect for sampling from a normal population.

Example 5.2.3

Weights of Seeds Figure 5.2.5 shows the sampling distribution of $\overline{Y}$ for samples of various sizes from the princess bean population of Example 5.2.2. Notice that for larger n the sampling distribution is more concentrated around the population mean $\mu = 500$ mg. As a consequence, the probability that $\overline{Y}$ is close to it is larger for larger n. For instance, consider the probability that $\overline{Y}$ is within ± 50 mg of μ, that is, $\Pr\{450 \leq \overline{Y} \leq 550\}$. Table 5.2.1 shows how this probability depends on n. ■

Table 5.2.1

n	$\Pr\{450 \leq \overline{Y} \leq 550\}$
4	0.59
9	0.79
16	0.91
64	0.999

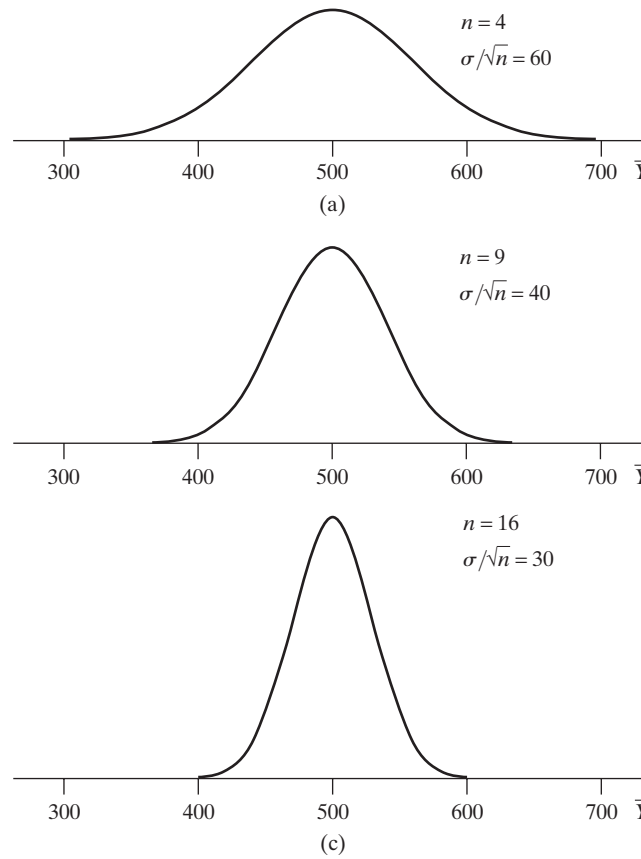


Figure 5.2.5 Sampling distribution of $\overline{Y}$ for various sample sizes n

Example 5.2.3 illustrates how the closeness of $\overline{Y}$ to μ depends on sample size. The mean of a larger sample is not *necessarily* closer to it than the mean of a smaller sample, but it has a *greater probability* of being close. It is in this sense that a larger sample provides more information about the population mean than a smaller sample.

Populations, Samples, and Sampling Distributions

In thinking about Theorem 5.2.1, it is important to distinguish clearly among three different distributions related to a quantitative variable Y: (1) the distribution of Y in the population; (2) the distribution of Y in a sample of data, and (3) the sampling distribution of $\overline{Y}$. The means and standard deviations of these distributions are summarized in Table 5.2.2.

Table 5.2.2

Distribution	Mean	Standard deviation
Y in population	μ	σ
Y in sample	$\bar{y}$	s
$\overline{Y}$ (in meta-study)	$\mu_{\overline{Y}} = \mu$	$\sigma_{\overline{Y}} = \dfrac{\sigma}{\sqrt{n}}$

The following example illustrates the distinction among the three distributions.

Example 5.2.4

Weights of Seeds For the princess bean population of Example 5.2.2, the population mean and standard deviation are $\mu = 500$ mg and $\sigma = 120$ mg; the population distribution of Y = weight is represented in Figure 5.2.6(a). Suppose we weigh a random sample of $n = 25$ seeds from the population and obtain the data in Table 5.2.3.

For the data in Table 5.2.3, the sample mean is $\bar{y} = 526.1$ mg and the sample standard deviation is $s = 113.7$ mg. Figure 5.2.6(b) shows a histogram of the data; this histogram represents the distribution of Y in the sample. The sampling distribution of $\overline{Y}$ is a theoretical distribution which relates, not to the particular sample shown in the histogram, but rather to the meta-study of repeated samples of size $n = 25$. The mean and standard deviation of the sampling distribution are

$$\mu_{\overline{Y}} = 500 \text{ mg and } \sigma_{\overline{Y}} = 120/\sqrt{25} = 24 \text{ mg}$$

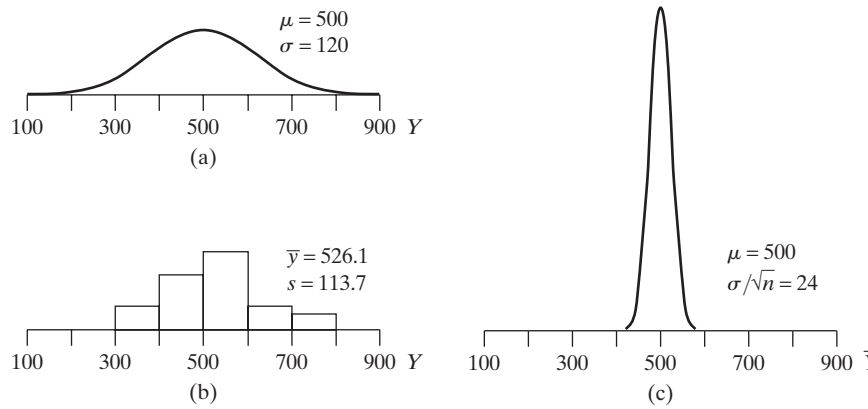


Figure 5.2.6 Three distributions related to Y = seed weight of princess beans: (a) population distribution of Y; (b) distribution of 25 observations of Y; (c) sampling distribution of $\overline{Y}$ for $n = 25$

Table 5.2.3 Weights of twenty-five princess bean seeds

Weight (mg)						
343	755	431	480	516	469	694
659	441	562	597	502	612	549
348	469	545	728	416	536	581
433	583	570	334			

The sampling distribution is represented in Figure 5.2.6(c). Notice that the distributions in Figures 5.2.6(a) and (b) are more or less similar; in fact, the distribution in (b) is an estimate (based on the data in Table 5.2.3) of the distribution in (a). By contrast, the distribution in (c) is much narrower, because it represents a distribution of *means* rather than of individual observations. ■

Other Aspects of Sampling Variability

The preceding discussion has focused on sampling variability in the sample mean, $\overline{Y}$. Two other important aspects of sampling variability are (1) sampling variability in the sample standard deviation, s and (2) sampling variability in the *shape* of the sample, as represented by the sample histogram. Rather than discuss these aspects formally, we illustrate them with the following example.

Example 5.2.5

Weights of Seeds In Figure 5.2.6(b) we displayed a random sample of 25 observations from the princess bean population of Example 5.2.2; now we display in Figure 5.2.7 eight additional random samples from the same population. (All nine samples were actually simulated using a computer.) Notice that, even though the samples were drawn from a normal population [pictured in Figure 5.2.6(a)], there is very substantial variation in the forms of the histograms. Notice also that there is considerable variation in the sample standard deviations. Of course, if the sample size were larger (say, $n = 100$ rather than $n = 25$), there would be less sampling variation; the histograms would tend to resemble a normal curve more closely, and the standard deviations would tend to be closer to the population value ($\sigma = 120$). ■

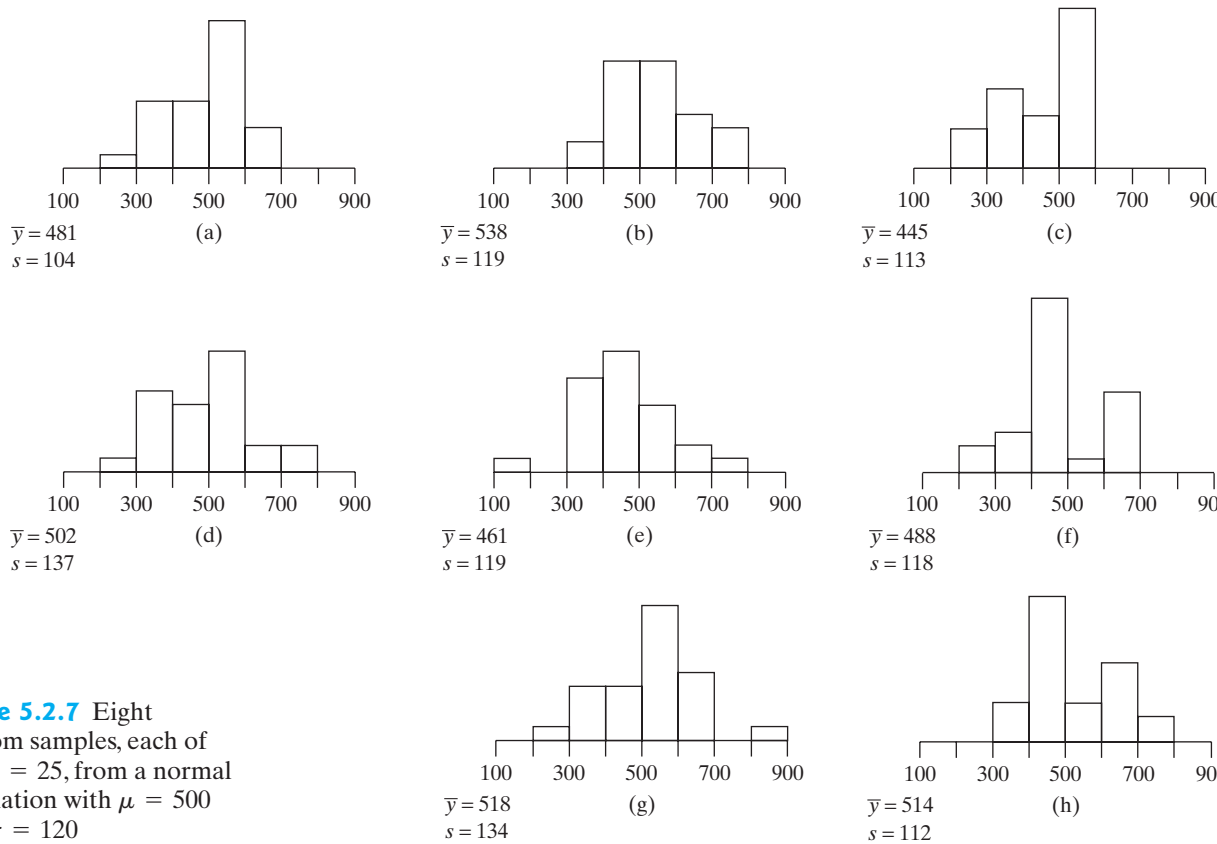


Figure 5.2.7 Eight random samples, each of size $n = 25$, from a normal population with $\mu = 500$ and $\sigma = 120$

Exercises 5.2.1–5.2.19

5.2.1 (Sampling exercise) Refer to Exercise 1.3.5. The collection of 100 ellipses shown there can be thought of as representing a natural population of the organism *C. ellipticus*. Use your judgment to choose a sample of 5 ellipses that you think should be reasonably representative of the population. (In order to best simulate the analogous judgment in a real-life setting, you should make your choice intuitively, without any detailed preliminary study of the population.) With a metric ruler, measure the length of each ellipse in your sample. Measure only the body, excluding any tail bristles; measurements to the nearest millimeter will be adequate. Compute the mean and standard deviation of the five lengths. To facilitate the pooling of results from the entire class, express the mean and standard deviation in millimeters, keeping two decimal places.

5.2.2 (Sampling exercise) Proceed as in Exercise 5.2.1, but use random sampling rather than "judgment" sampling. To do this, choose 10 random digits (from Table 1 or your calculator). Let the first 2 digits be the number of the first ellipse that goes into your sample, and so on. The 10 random digits will give you a random sample of five ellipses.

5.2.3 (Sampling exercise) Proceed as in Exercise 5.2.2, but choose a random sample of 20 ellipses.

5.2.4 Refer to Exercise 5.2.2. The following scheme is proposed for choosing a sample of 5 ellipses from the population of 100 ellipses. (i) Choose a point at random in the ellipse "habitat" (that is, the figure); this could be done crudely by dropping a pencil point on the page, or much better by overlaying the page with graph paper and using random digits. (ii) If the chosen point is inside an ellipse, include that ellipse in the sample, otherwise start again at step (i). (iii) Continue until 5 ellipses have been selected. Explain why this scheme is not equivalent to random sampling. In what direction is the scheme biased—that is, would it tend to produce a $\bar{y}$ that is too large, or a $\bar{y}$ that is too small?

5.2.5 The serum cholesterol levels of a population of 12- to 14-year-olds follow a normal distribution with mean 162 mg/dl and standard deviation 28 mg/dl (as in Example 4.1.1).

(a) What percentage of the 12- to 14-year-olds have serum cholesterol values between 152 and 172 mg/dl?

(b) Suppose we were to choose at random from the population a large number of groups of nine 12- to 14-year-olds each. In what percentage of the groups would the group mean cholesterol value be between 152 and 172 mg/dl?

(c) If $\overline{Y}$ represents the mean cholesterol value of a random sample of nine 12- to 14-year-olds from the population, what is $\Pr\{152 \leq \overline{Y} \leq 172\}$?

5.2.6 An important indicator of lung function is forced expiratory volume (FEV), which is the volume of air that a person can expire in one second. Dr. Hernandez plans to measure FEV in a random sample of n young women from a certain population, and to use the sample mean $\bar{y}$ as an estimate of the population mean. Let E be the event that Hernandez's sample mean will be within ± 100 ml of the population mean. Assume that the population distribution is normal with mean 3,000 ml and standard deviation 400 ml.[3] Find $\Pr\{E\}$ if

(a) $n = 15$

(b) $n = 60$

(c) How does $\Pr\{E\}$ depend on the sample size? That is, as n increases, does $\Pr\{E\}$ increase, decrease, or stay the same?

5.2.7 Refer to Exercise 5.2.6. Assume that the population distribution of FEV is normal with standard deviation 400 ml.

(a) Find $\Pr\{E\}$ if $n = 15$ and the population mean is 2,800 ml.

(b) Find $\Pr\{E\}$ if $n = 15$ and the population mean is 2,600 ml.

(c) How does $\Pr\{E\}$ depend on the population mean?

5.2.8 The heights of a certain population of corn plants follow a normal distribution with mean 145 cm and standard deviation 22 cm (as in Exercise 4.S.4).

(a) What percentage of the plants are between 135 and 155 cm tall?

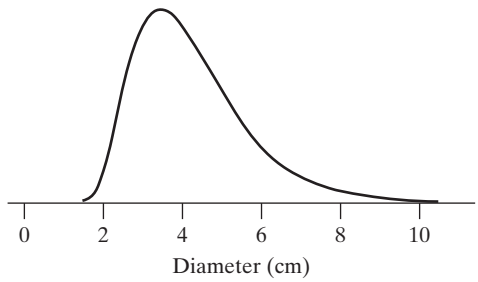
(b) Suppose we were to choose at random from the population a large number of samples of 16 plants each. In what percentage of the samples would the sample mean height be between 135 and 155 cm?

(c) If $\overline{Y}$ represents the mean height of a random sample of 16 plants from the population, what is $\Pr\{135 \leq \overline{Y} \leq 155\}$?

(d) If $\overline{Y}$ represents the mean height of a random sample of 36 plants from the population, what is $\Pr\{135 \leq \overline{Y} \leq 155\}$?

5.2.9 The basal diameter of a sea anemone is an indicator of its age. The density curve shown here represents the distribution of diameters in a certain large population of anemones; the population mean diameter is 4.2 cm, and the standard deviation is 1.4 cm.[4] Let $\overline{Y}$ represent the

mean diameter of 25 anemones randomly chosen from the population.



(a) Find the approximate value of $\Pr\{4 \le \overline{Y} \le 5\}$.

(b) Why is your answer to part (a) approximately correct even though the population distribution of diameters is clearly not normal? Would the same approach be equally valid for a sample of size 2 rather than 25? Why or why not?

5.2.10 In a certain population of fish, the lengths of the individual fish follow approximately a normal distribution with mean 54.0 mm and standard deviation 4.5 mm. We saw in Example 4.3.1 that in this situation 65.68% of the fish are between 51 and 60 mm long. Suppose a random sample of four fish is chosen from the population. Find the probability that

(a) all four fish are between 51 and 60 mm long.

(b) the mean length of the four fish is between 51 and 60 mm.

5.2.11 In Exercise 5.2.10, the answer to part (b) was larger than the answer to part (a). Argue that this must necessarily be true, no matter what the population mean and standard deviation might be. [*Hint:* Can it happen that the event in part (a) occurs but the event in part (b) does not?]

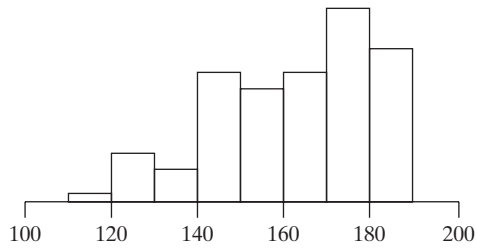
5.2.12 Professor Smith conducted a class exercise in which students ran a computer program to generate random samples from a population that had a mean of 50 and a standard deviation of 9 mm. Each of Smith's students took a random sample of size n and calculated the sample mean. Smith found that about 68% of the students had sample means between 48.5 and 51.5 mm. What was n? (Assume that n is large enough that the Central Limit Theorem is applicable.)

5.2.13 A certain assay for serum alanine aminotransferase (ALT) is rather imprecise. The results of repeated assays of a single specimen follow a normal distribution with mean equal to the ALT concentration for that specimen and standard deviation equal to 4 U/l (as in Exercise 4.S.15). Suppose a hospital lab measures many specimens every day, and specimens with reported ALT values of 40 or more are flagged as "unusually high." If a patient's true ALT concentration is 35 U/l, find the probability that his specimen will be flagged as "unusually high"

(a) if the reported value is the result of a single assay.

(b) if the reported value is the mean of three independent assays of the same specimen.

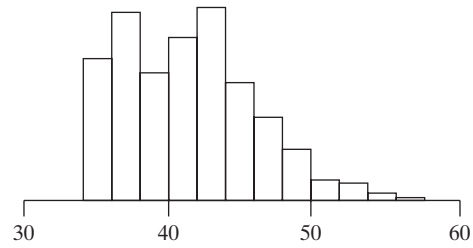
5.2.14 The mean of the distribution shown in the following histogram is 162 and the standard deviation is 18. Consider taking random samples of size $n = 9$ from this distribution and calculating the sample mean, $\bar{y}$, for each sample.



(a) What is the mean of the sampling distribution of $\overline{Y}$?

(b) What is the standard deviation of the sampling distribution of $\overline{Y}$?

5.2.15 The mean of the distribution shown in the following histogram is 41.5 and the standard deviation is 4.7. Consider taking random samples of size $n = 4$ from this distribution and calculating the sample mean, $\bar{y}$, for each sample.



(a) What is the mean of the sampling distribution of $\overline{Y}$?

(b) What is the standard deviation of the sampling distribution of $\overline{Y}$?

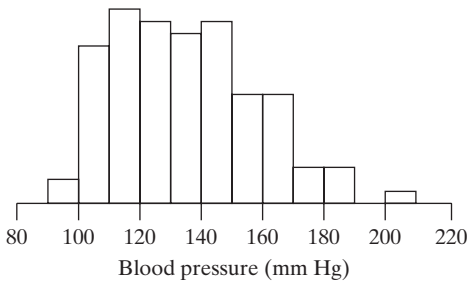
5.2.16 Refer to the histogram in Exercise 5.2.15. Suppose that 100 random samples are taken from this population and the sample mean is calculated for each sample. If we were to make a histogram of the distribution of the sample means from 100 samples, what kind of shape would we expect the histogram to have

(a) if $n = 2$ for each random sample?

(b) if $n = 25$ for each random sample?

5.2.17 Refer to the histogram in Exercise 5.2.15. Suppose that 100 random samples are taken from this population and the sample mean is calculated for each sample. If we were to make a histogram of the distribution of the sample means from 100 samples, what kind of shape would we expect the histogram to have if $n = 1$ for each random sample? That is, what does the sampling distribution of the mean look like when the sample size is $n = 1$?

5.2.18 A medical researcher measured systolic blood pressure in 100 middle-aged men.[5] The results are displayed in the accompanying histogram; note that the distribution is rather skewed. According to the Central Limit Theorem, would we expect the distribution of blood pressure readings to be less skewed (and more bell shaped) if it were based on $n = 400$ rather than $n = 100$ men? Explain.



5.2.19 The partial pressure of oxygen, PaO_2, is a measure of the amount of oxygen in the blood. Assume that the distribution of PaO_2 levels among newborns has an average of 38 (mm Hg) and a standard deviation of 9.[6] If we take a sample of size $n = 25$,

(a) what is the probability that the sample average will be greater than 36?

(b) what is the probability that the sample average will be greater than 41?

5.3 Illustration of the Central Limit Theorem (Optional)

The importance of the normal distribution in statistics is due largely to the Central Limit Theorem and related theorems. In this section we take a closer look at the Central Limit Theorem. According to the Central Limit Theorem, the sampling distribution of $\overline{Y}$ is approximately normal if n is large. If we consider larger and larger samples from a fixed nonnormal population, then the sampling distribution of $\overline{Y}$ will be more nearly normal for larger n. The following examples show the Central Limit Theorem at work for two nonnormal distributions: a moderately skewed distribution (Example 5.3.1) and a highly skewed distribution (Example 5.3.2).

Example 5.3.1

Eye Facets The number of facets in the eye of the fruitfly *Drosophila melanogaster* is of interest in genetic studies. The distribution of this variable in a certain *Drosophila* population can be approximated by the density function shown in Figure 5.3.1. The distribution is moderately skewed; the population mean and standard deviation are $\mu = 64$ and $\sigma = 22$.[7]

Figure 5.3.2 shows the sampling distribution of $\overline{Y}$ for samples of various sizes from the eye-facet population. In order to clearly show the shape of these distributions, we have plotted them to different scales; the horizontal scale is stretched more for larger n. Notice that the distributions are somewhat skewed to the right, but the skewness is diminished for larger n; for $n = 32$ the distribution looks very nearly normal. ■

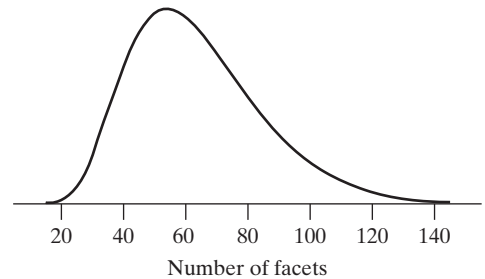


Figure 5.3.1 Distribution of eye-facet number in a *Drosophila* population

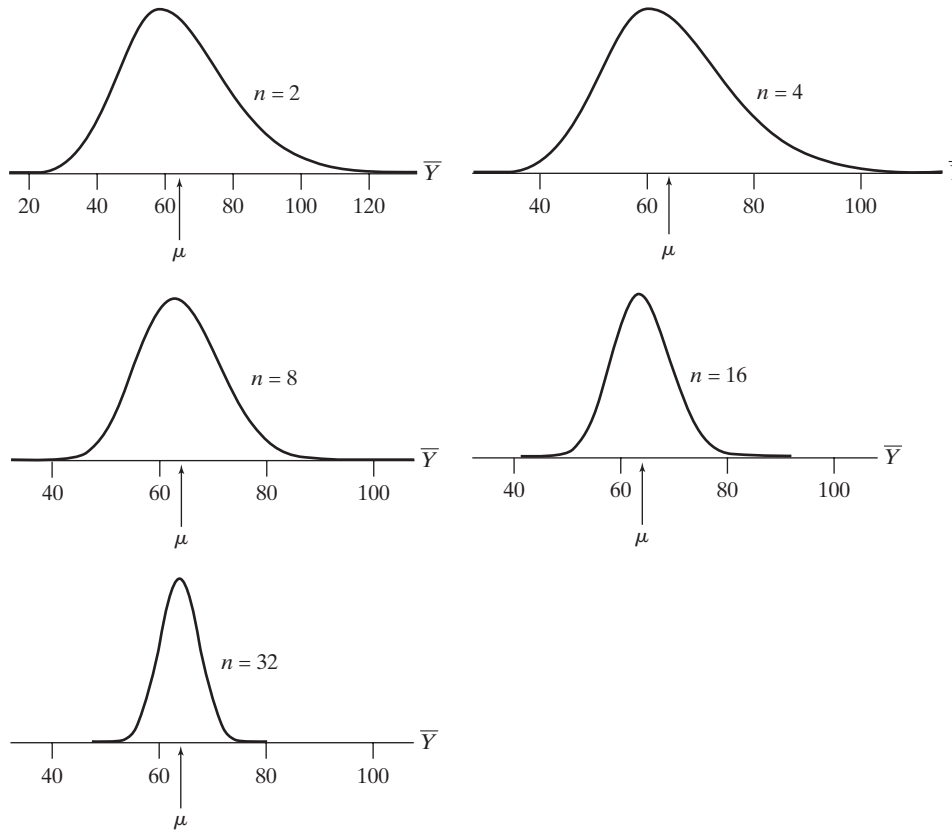


Figure 5.3.2 Sampling distributions of $\overline{Y}$ for samples from the *Drosophila* eye-facet population

Example 5.3.2

Reaction Time A psychologist measured the time required for a person to reach up from a fixed position and operate a pushbutton with his or her forefinger. The distribution of time scores (in milliseconds) for a single person is represented by the density shown in Figure 5.3.3. About 10% of the time, the subject fumbled, or missed the button on the first thrust; the resulting delayed times appear as the second peak of the distribution.[8] The first peak is centered at 115 ms and the second at 450 ms; because of the two peaks, the overall distribution is violently skewed. The population mean and standard deviation are $\mu = 148$ ms and $\sigma = 105$ ms, respectively.

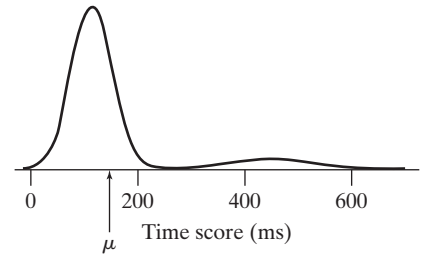


Figure 5.3.3 Distribution of time scores in a button-pushing task

Figure 5.3.4 shows the sampling distribution of $\overline{Y}$ for samples of various sizes from the time-score distribution. To show the shape clearly, the Y scale has been stretched more for larger n. Notice that for small n the distribution has several modes. As n increases, these modes are reduced to bumps and finally disappear, and the distribution becomes increasingly symmetric. ■

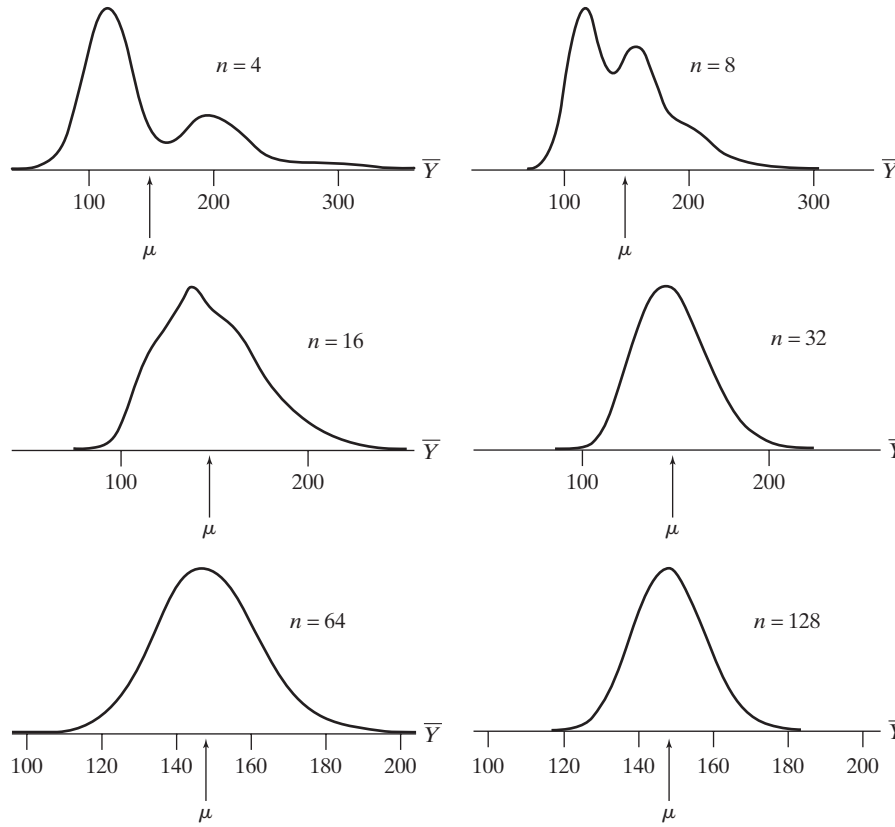


Figure 5.3.4 Sampling distributions of $\overline{Y}$ for samples from the time-score population

Examples 5.3.1 and 5.3.2 illustrate the fact, mentioned in Section 5.2, that the meaning of the requirement "n is large" in the Central Limit Theorem depends on the shape of the population distribution. Approximate normality of the sampling distribution of $\overline{Y}$ will be achieved for a moderate n if the population distribution is only moderately nonnormal (as in Example 5.3.1), while a highly nonnormal population (as in Example 5.3.2) will require a larger n.

Note, however, that Example 5.3.2 indicates the remarkable strength of the Central Limit Theorem. The skewness of the time-score distribution is so extreme that one might be reluctant to consider the mean as a summary measure. Even in this "worst case," you can see the effect of the Central Limit Theorem in the relative smoothness and symmetry of the sampling distribution for $n = 64$.

The Central Limit Theorem may seem rather like magic. To demystify it somewhat, we look at the time-score sampling distributions in more detail in the following example.

Example 5.3.3

Reaction Time Consider the sampling distributions of $\overline{Y}$ displayed in Figure 5.3.4. Consider first the distribution for $n = 4$, which is the distribution of the mean of four button-pressing times. The high peak at the left of the distribution represents cases in which the subject did not fumble any of the 4 thrusts, so that all four times were about 115 ms; such an outcome would occur about 66% of the time [from the binomial distribution, because $(0.9)^4 = 0.66$]. The next lower peak represents cases in which 3 thrusts took about 115 ms each, while one was fumbled and took about 450 ms. (Notice that the average of three 115's and one 450 is about 200, which is the center of the second peak.) Similarly, the third peak (which is barely visible)

represents cases in which the subject fumbled 2 of the 4 thrusts. The peaks representing 3 and 4 fumbles are too low to be visible in the plot.

Now consider the plot for $n = 8$. The first peak represents 8 good thrusts (no fumbles), the second represents 7 good thrusts and 1 fumble, the third represents 6 good thrusts and 2 fumbles, and so on. The fourth and later peaks are blended together. For $n = 16$ it is more likely to see 15 good thrusts and 1 fumble than 16 good thrusts (as you can verify from the binomial distribution) and thus there is a bump, corresponding to 16 good thrusts, below the overall peak, which corresponds to 15 good thrusts; the bump to the right of the peak corresponds to 14 good thrusts and 2 fumbles. For $n = 32$, the most likely outcome is 3 fumbles and 29 good thrusts; this outcome gives a mean time of about

$$\frac{(3)(450) + (29)(115)}{32} \approx 146 \text{ ms}$$

which is the location of the central peak. For similar reasons, the distribution for larger n is centered at about 148 ms, which is the population mean. ■

Exercises 5.3.1–5.3.3

5.3.1 Refer to Example 5.3.3. In the sampling distribution of $\overline{Y}$ for $n = 4$ (Figure 5.3.4), approximately what is the area under

(a) the first peak?

(b) the second peak?
(*Hint:* Use the binomial distribution.)

5.3.2 Refer to Example 5.3.3. Consider the sampling distribution of $\overline{Y}$ for $n = 2$ (which is not shown in Figure 5.3.4).

(a) Make a rough sketch of the sampling distribution. How many peaks does it have? Show the location (on the Y-axis) of each peak.

(b) Find the approximate area under each peak. (*Hint:* Use the binomial distribution.)

5.3.3 Refer to Example 5.3.3. Consider the sampling distribution of $\overline{Y}$ for $n = 1$ (which is not shown in Figure 5.3.4). Make a rough sketch of the sampling distribution. How many peaks does it have? Show the location (on the Y-axis) of each peak.

5.4 The Normal Approximation to the Binomial Distribution (Optional)

The Central Limit Theorem tells us that the sampling distribution of a mean becomes bell shaped as the sample size increases. Suppose we have a large dichotomous population for which we label the two types of outcomes as "1" (for "success") and "0" (for "failure"). If we take a sample and calculate the average number of 1's, then this sample average is just the sample proportion of 1's—commonly labeled as $\hat{P}$—and is governed by the Central Limit Theorem. This means that if the sample size n is large, then the distribution of $\hat{P}$ will be approximately normal.

Note that if we know the number of 1's (i.e., the number of successes in n trials), then we know the proportion of 1's and vice versa. Thus, the normal approximation to the binomial distribution can be expressed in two equivalent ways: in terms of the number of successes, Y, or in terms of the proportion of successes, $\hat{P}$. We state both forms in the following theorem. In this theorem, n represents the sample size (or, more generally, the number of independent trials) and p represents the population proportion (or, more generally, the probability of success in each independent trial).

Theorem 5.4.1: Normal Approximation to Binomial Distribution

(a) If n is large, then the binomial distribution of the number of successes, Y, can be approximated by a normal distribution with

$$\text{Mean} = np$$

and

$$\text{Standard deviation} = \sqrt{np(1-p)}$$

(b) If n is large, then the sampling distribution of $\hat{P}$ can be approximated by a normal distribution with

$$\text{Mean} = p$$

and

$$\text{Standard deviation} = \sqrt{\frac{p(1-p)}{n}}$$

Remarks

1. Appendix 5.1 provides more detailed explanation of the relationship between the normal approximation to the binomial and the Central Limit Theorem.
2. As shown in Appendix 3.2, for a population of 0's and 1's, where the proportion of 1's is given by p, the standard deviation is $\sigma = \sqrt{p(1-p)}$. Theorem 5.2.1 stated that the standard deviation of a mean is given by $\frac{\sigma}{\sqrt{n}}$. We think of $\hat{P}$ in part (b) of Theorem 5.2.1 as a special kind of sample average, for the setting in which all of the data are 0's and 1's. Thus, Theorem 5.2.1 tells us that the standard deviation of $\hat{P}$ should be $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$, or $\sqrt{\frac{p(1-p)}{n}}$, which agrees with the result stated in Theorem 5.4.1(b).

The following example illustrates the use of Theorem 5.4.1.

Example 5.4.1 Normal Approximation to Binomial We consider a binomial distribution with $n = 50$ and $p = 0.3$. Figure 5.4.1(a) shows this binomial distribution, using spikes to represent probabilities; superimposed is a normal curve with

$$\text{Mean} = np = (50)(0.3) = 15$$

and

$$\text{SD} = \sqrt{np(1-p)} = \sqrt{(50)(0.3)(0.7)} = 3.24$$

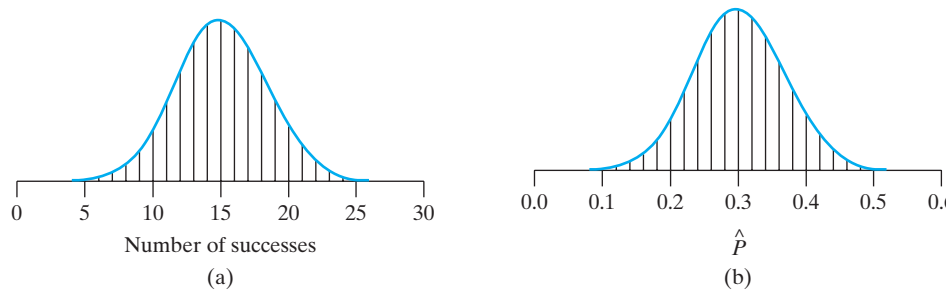


Figure 5.4.1 The normal approximation (blue curve) to the binomial distribution (black spikes) with $n = 50$ and $p = 0.3$

Note that the curve fits the distribution fairly well. Figure 5.4.1(b) shows the sampling distribution of $\hat{P}$ for $n = 50$ and $p = 0.3$; superimposed is a normal curve with

$$\text{Mean} = p = 0.3$$

and

$$\text{SD} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.3)(0.7)}{50}} = 0.0648$$

Note that Figure 5.4.1(b) is just a relabeled version of Figure 5.4.1(a).

To illustrate the use of the normal approximation, let us find the probability that 50 independent trials result in at least 18 successes. We could use the binomial formula to find the probability of exactly 18 successes in 50 trials and add this to the probability of exactly 19 successes, exactly 20 successes, and so on:

$$\begin{aligned}\Pr\{\text{at least 18 successes}\} &= {}_{50}C_{18}(0.3)^{18}(1-0.3)^{50-18} \\ &\quad + {}_{50}C_{19}(0.3)^{19}(1-0.3)^{50-19} + \ldots \\ &= 0.0772 + 0.0558 + \ldots = 0.2178\end{aligned}$$

This probability can be visualized as the area above and to the right of the "18" in Figure 5.4.2. The normal approximation to the probability is the corresponding area under the normal curve, which is shaded in Figure 5.4.2. The z value that corresponds to 18 is

$$z = \frac{18 - 15}{3.2404} = 0.93$$

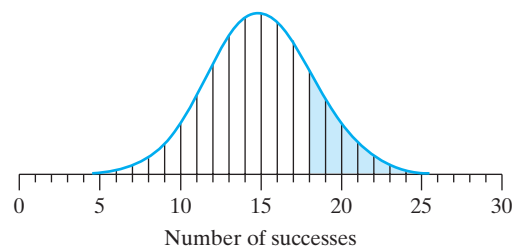


Figure 5.4.2 Normal approximation to the probability of at least 18 successes

From Table 3, we find that the area is $1 - 0.8238 = 0.1762$, which is reasonably close to the exact value of 0.2178. This approximation can be improved by accounting for the fact that the binomial distribution is discrete and the normal distribution is continuous as we shall see below. ■

The Continuity Correction

As we have seen in Chapter 4, because the normal distribution is continuous, probabilities are computed areas under the normal curve, rather than being the height of the normal curve at any particular value. Because of this, to compute $\Pr\{Y = 18\}$, the probability of 18 successes, we think of "18" as covering the space from 17.5 to 18.5 and thus we consider the area under the normal curve between 17.5 and 18.5; this is illustrated in Figure 5.4.3. Likewise, to get a more accurate approximation in Example 5.4.1, we can use 17.5 in place of 18 when finding the z value. Each of these is an example of a continuity correction.

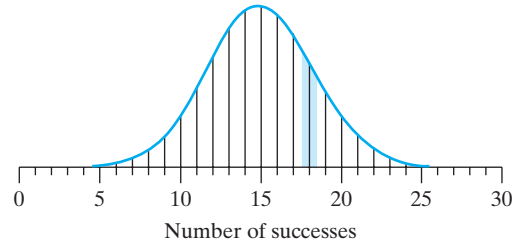


Figure 5.4.3 Normal approximation to the probability of exactly 18 successes

Example 5.4.2

Applying continuity correction within the normal approximation, the probability of at least 18 successes in 50 trials, when $p = 0.3$, is approximated by finding

$$z = \frac{17.5 - 15}{3.2404} = 0.77$$

From Table 3, we find that the area above 0.77 is $1 - 0.7794 = 0.2206$, which agrees quite well with the exact value of 0.2178. This area is displayed in Figure 5.4.4. ■

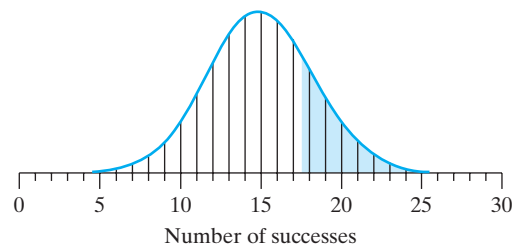


Figure 5.4.4 Improved normal approximation to the probability of at least 18 successes

Example 5.4.3

To illustrate part (b) of Theorem 5.4.1, we again assume that $n = 50$ and $p = 0.3$. Consider finding the probability that at least 40% of the 50 trials in a binomial experiment with $p = 0.3$ result in successes. That is, we wish to find $\Pr\{\hat{P} \geq 0.40\}$. The normal approximation to this probability is the shaded area in Figure 5.4.5. Using continuity correction, the boundary of the area is $\hat{p} = 19.5/50 = 0.39$, which corresponds on the Z scale to

$$z = \frac{0.39 - 0.30}{0.0648} = 1.39$$

The resulting approximation (from Table 3) is then

$$\Pr\{\hat{P} \geq 0.40\} \approx 1 - 0.9177 = 0.0823$$

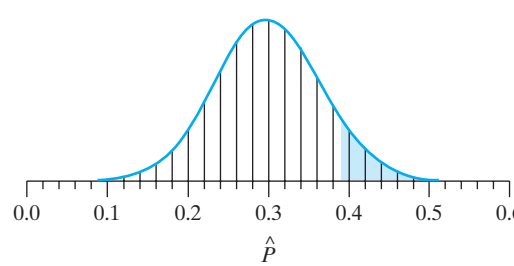


Figure 5.4.5 Normal approximation to $\Pr\{\hat{P} \geq 0.40\}$

which agrees very well with the exact value of 0.0848 (found by using the binomial formula). ■

Remark Any problem involving the normal approximation to the binomial can be solved in two ways: in terms of Y, using part (a) of Theorem 5.4.1, or in terms of $\hat{P}$, using part (b) of the theorem. Although it is natural to state questions in terms of proportions (e.g., "What is $\Pr\{\hat{P} > 0.70\}$?"), it is often easier to solve problems in terms of the binomial count Y (e.g., "What is $\Pr\{Y > 35\}$?"), particularly when using continuity correction. The following example illustrates the approach of converting a question about a sample proportion into a question about the number of successes for a binomial random variable.

Example 5.4.4

Consider a binomial distribution with $n = 50$ and $p = 0.3$. The sample proportion of successes, out of the 50 trials, is $\hat{P}$. Figure 5.4.1(b) shows the sampling distribution of $\hat{P}$ with a normal curve superimposed.

Suppose we wish to find the probability that $0.24 \le \hat{P} \le 0.36$. Since $\hat{P} = Y/50$, this is the probability that $0.24 \le Y/50 \le 0.36$, which is the same as the probability that $12 \le Y \le 18$. That is, $\Pr\{0.24 \le \hat{P} \le 0.36\} = \Pr\{12 \le Y \le 18\}$.

We know that Y has a binomial distribution with mean $= np = (50)(0.3) = 15$ and SD $= \sqrt{np(1-p)} = \sqrt{(50)(0.3)(0.7)} = 3.24$. Using continuity correction, we would find the Z scale values of

$$z = \frac{11.5 - 15}{3.24} = -1.08$$

and

$$z = \frac{18.5 - 15}{3.24} = 1.08$$

Then, using Table 3, we have $\Pr\{0.24 \le \hat{P} \le 0.36\} = \Pr\{12 \le Y \le 18\} \approx 0.8599 - 0.1401 = 0.7198$. ■

How Large Must n Be?

Theorem 5.4.1 states that the binomial distribution can be approximated by a normal distribution if n is "large." It is helpful to know how large n must be in order for the approximation to be adequate. The required n depends on the value of p. If $p = 0.5$, then the binomial distribution is symmetric and the normal approximation is quite good even for n as small as 10. However, if $p = 0.1$, the binomial distribution for $n = 10$ is quite skewed, and is poorly fitted by a normal curve; for larger n the skewness is diminished and the normal approximation is better. A simple rule of thumb is the following:

> The normal approximation to the binomial distribution is fairly good if both np and $n(1-p)$ are at least equal to 5.

For example, if $n = 50$ and $p = 0.3$, as in Example 5.4.4, then $np = 15$ and $n(1-p) = 35$; since $15 \ge 5$ and $35 \ge 5$, the rule of thumb indicates that the normal approximation is fairly good.

Exercises 5.4.1–5.4.13

5.4.1 A fair coin is to be tossed 20 times. Find the probability that 10 of the tosses will fall heads and 10 will fall tails,

(a) using the binomial distribution formula.

(b) using the normal approximation with the continuity correction.

5.4.2 In the United States, 44% of the population has type O blood. Suppose a random sample of 12 persons is taken. Find the probability that 6 of the persons will have type O blood (and 6 will not)

(a) using the binomial distribution formula.

(b) using the normal approximation.

5.4.3 Refer to Exercise 5.4.2. Find the probability that at most 6 of the persons will have type O blood by using the normal approximation

(a) without the continuity correction.

(b) with the continuity correction.

5.4.4 An epidemiologist is planning a study on the prevalence of oral contraceptive use in a certain population.[9] She plans to choose a random sample of n women and to use the sample proportion of oral contraceptive users ($\hat{P}$) as an estimate of the population proportion (p). Suppose that in fact $p = 0.12$. Use the normal approximation (with the continuity correction) to determine the probability that $\hat{P}$ will be within ± 0.03 of p if

(a) $n = 100$.

(b) $n = 200$.

[*Hint:* If you find using part (b) of Theorem 5.4.1 to be difficult here, try using part (a) of the theorem instead.]

5.4.5 In a study of how people make probability judgments, college students (with no background in probability or statistics) were asked the following question.[10] A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

For a period of one year, each hospital recorded the days on which at least 60% of the babies born were boys. Which hospital do you think recorded more such days?

- The larger hospital
- The smaller hospital
- About the same (i.e., within 5% of each other)

(a) Imagine that you are a participant in the study. Which answer would you choose, based on intuition alone?

(b) Determine the correct answer by using the normal approximation (without the continuity correction) to calculate the appropriate probabilities.

5.4.6 Consider random sampling from a dichotomous population with $p = 0.3$, and let E be the event that $\hat{P}$ is within ± 0.05 of p. Use the normal approximation (without the continuity correction) to calculate $\Pr\{E\}$ for a sample of size $n = 400$.

5.4.7 Refer to Exercise 5.4.6. Calculate $\Pr\{E\}$ for $n = 40$ (rather than 400) without the continuity correction.

5.4.8 Refer to Exercise 5.4.6. Calculate $\Pr\{E\}$ for $n = 40$ (rather than 400) with the continuity correction.

5.4.9 A certain cross between sweet-pea plants will produce progeny that are either purple flowered or white flowered;[11] the probability of a purple-flowered plant is $p = \frac{9}{16}$. Suppose n progeny are to be examined, and let $\hat{P}$ be the sample proportion of purple-flowered plants. It might happen, by chance, that $\hat{P}$ would be closer to $\frac{1}{2}$ than to $\frac{9}{16}$. Find the probability that this misleading event would occur if

(a) $n = 1$.

(b) $n = 64$.

(c) $n = 320$.

(Use the normal approximation without the continuity correction.)

5.4.10 Cytomegalovirus (CMV) is a (generally benign) virus that infects one-half of young adults.[12] If a random sample of 10 young adults is taken, find the probability that between 30% and 40% (inclusive) of those sampled will have CMV,

(a) using the binomial distribution formula.

(b) using the normal approximation with the continuity correction.

5.4.11 In a certain population of mussels (*Mytilus edulis*), 80% of the individuals are infected with an intestinal parasite.[13] A marine biologist plans to examine 100 randomly chosen mussels from the population. Find the probability that 85% or more of the sampled mussels will be infected, using the normal approximation without the continuity correction.

5.4.12 Refer to Exercise 5.4.11. Find the probability that 85% or more of the sampled mussels will be infected, using the normal approximation with the continuity correction.

5.4.13 Refer to Exercise 5.4.11. Suppose that the biologist takes a random sample of size 50. Find the probability that fewer than 35 of the sampled mussels will be infected, using the normal approximation

(a) without the continuity correction.

(b) with the continuity correction.

5.5 Perspective

In this chapter we have presented the concept of a sampling distribution and have focused on the sampling distribution of $\overline{Y}$. Of course, there are many other important sampling distributions, such as the sampling distribution of the sample standard deviation and the sampling distribution of the sample median.

Let us take another look at the random sampling model in the light of Chapter 5. As we have seen, a *random* sample is not necessarily a *representative* sample.* But using sampling distributions, one can specify the degree of representativeness to be expected in a random sample. For instance, it is intuitively plausible that a larger sample is likely to be more representative than a smaller sample from the same population. In Sections 5.1 and 5.2 we saw how a sampling distribution can make this vague intuition precise by specifying the probability that a specified degree of representativeness will be achieved by a random sample. Thus, sampling distributions provide what has been called "certainty about uncertainty."[14]

In Chapter 6 we will see for the first time how the theory of sampling distributions can be put to practical use in the analysis of data. We will find that, although the calculations of Chapter 5 seem to require the knowledge of unknowable quantities (such as μ and σ), when analyzing data one can nevertheless estimate the probable magnitude of sampling error using only information contained in the sample itself.

In addition to their application to data analysis, sampling distributions provide a basis for comparing the relative merits of different methods of analysis. For example, consider sampling from a normal population with mean μ. Of course, the sample mean $\overline{Y}$ is an estimator of μ. But since a normal distribution is symmetric, it is also the population median, so the sample *median* is also an estimator of μ. How, then, can we decide which estimator is better? This question can be answered in terms of sampling distributions, as follows: Statisticians have determined that, if the population is normal, the sample median is inferior to the sample mean in the sense that its sampling distribution, while centered at μ, has a standard deviation larger than $\frac{\sigma}{\sqrt{n}}$.

Consequently, the sample median is less efficient (as an estimator of μ) than the sample mean; for a given sample size n, the sample median provides less information about μ than does the sample mean. (If the population is not normal, however, the sample median can be much more efficient than the mean.)

*It is true, however, that sometimes the investigator can force the sample to be representative with respect to some variable (not the one under study) whose population distribution is known; for example, a stratified random sample as discussed in Section 1.3. The methods of analysis given in this book, however, are only appropriate for *simple* random samples and cannot be applied without suitable modification.

Supplementary Exercises 5.S.1–5.S.12

(*Note:* Exercises preceded by an asterisk refer to optional sections.)

5.S.1 In an agricultural experiment, a large field of wheat was divided into many plots (each plot being 7×100 ft) and the yield of grain was measured for each plot. These plot yields followed approximately a normal distribution with mean 88 lb and standard deviation 7 lb (as in Exercise 4.3.5). Let $\overline{Y}$ represent the mean yield of five plots chosen at random from the field. Find $\Pr\{\overline{Y} > 90\}$.

5.S.2 Consider taking a random sample of size 14 from the population of students at a certain college and measuring the diastolic blood pressure each of the 14 students. In the context of this setting, explain what is meant by the sampling distribution of the sample mean.

5.S.3 Refer to the setting of Exercise 5.S.2. Suppose that the population mean is 70 mmHg and the population standard deviation is 10 mmHg. If the sample size is 14, what is the standard deviation of the sampling distribution of the sample mean?

5.S.4 The heights of men in a certain population follow a normal distribution with mean 69.7 inches and standard deviation 2.8 inches.[15]

(a) If a man is chosen at random from the population, find the probability that he will be more than 72 inches tall.

(b) If two men are chosen at random from the population, find the probability that (i) both of them will be more than 72 inches tall; (ii) their mean height will be more than 72 inches.

5.S.5 Suppose a botanist grows many individually potted eggplants, all treated identically and arranged in groups of four pots on the greenhouse bench. After 30 days of growth, she measures the total leaf area Y of each plant. Assume that the population distribution of Y is approximately normal with mean = 800 cm^2 and SD = 90 cm^2.[16]

(a) What percentage of the plants in the population will have leaf area between 750 cm^2 and 850 cm^2?

(b) Suppose each group of four plants can be regarded as a random sample from the population. What percentage of the groups will have a group mean leaf area between 750 cm^2 and 850 cm^2?

5.S.6 Refer to Exercise 5.S.5. In a real greenhouse, what factors might tend to invalidate the assumption that each group of plants can be regarded as a random sample from the same population?

***5.S.7** Consider taking a random sample of size 25 from a population in which 42% of the people have type A blood. What is the probability that the sample proportion with type A blood will be greater than 0.44? Use the normal approximation to the binomial with continuity correction.

5.S.8 The activity of a certain enzyme is measured by counting emissions from a radioactively labeled molecule. For a given tissue specimen, the counts in consecutive 10-second time periods may be regarded (approximately) as repeated independent observations from a normal distribution (as in Exercise 4.S.1). Suppose the mean 10-second count for a certain tissue specimen is 1,200 and the standard deviation is 35. For that specimen, let Y represent a 10-second count and let $\overline{Y}$ represent the mean of six 10-second counts. Both Y and $\overline{Y}$ are unbiased—they each have an average of 1,200—but that doesn't imply that they are equally good. Find $\Pr\{1{,}175 \le Y \le 1{,}225\}$ and $\Pr\{1{,}175 \le \overline{Y} \le 1{,}225\}$, and compare the two. Does the comparison indicate that counting for one minute and dividing by 6 would tend to give a more precise result than merely counting for a single 10-second time period? How?

5.S.9 In a certain lab population of mice, the weights at 20 days of age follow approximately a normal distribution with mean weight = 8.3 gm and standard deviation = 1.7 gm.[17] Suppose many litters of 10 mice each are to be weighed. If each litter can be regarded as a random sample from the population, what percentage of the litters will have a total weight of 90 gm or more? (*Hint:* How is the total weight of a litter related to the mean weight of its members?)

5.S.10 Refer to Exercise 5.S.9. In reality, what factors would tend to invalidate the assumption that each litter can be regarded as a random sample from the same population?

5.S.11 Consider taking a random sample of size 25 from a population of plants, measuring the weight of each plant, and adding the weights to get a sample total. In the context of this setting, explain what is meant by the sampling distribution of the sample total.

5.S.12 The skull breadths of a certain population of rodents follow a normal distribution with a standard deviation of 10 mm. Let $\overline{Y}$ be the mean skull breadth of a random sample of 64 individuals from this population, and let μ be the population mean skull breadth.

(a) Suppose μ = 50 mm. Find $\Pr\{\overline{Y}$ is within ±2 mm of $\mu\}$.

(b) Suppose μ = 100 mm. Find $\Pr\{\overline{Y}$ is within ±2 mm of $\mu\}$.

(c) Suppose μ is unknown. Can you find $\Pr\{\overline{Y}$ is within ±2 mm of $\mu\}$? If so, do it. If not, explain why not.