DISTRIBUTIONS OF SAMPLING STATISTICS

6.1 INTRODUCTION

The science of statistics deals with drawing conclusions from observed data. For instance, a typical situation in a technological study arises when one is confronted with a large collection, or *population*, of items that have measurable values associated with them. By suitably *sampling* from this collection, and then analyzing the sampled items, one hopes to be able to draw some conclusions about the collection as a whole.

To use sample data to make inferences about an entire population, it is necessary to make some assumptions about the relationship between the two. One such assumption, which is often quite reasonable, is that there is an underlying (population) probability distribution such that the measurable values of the items in the population can be thought of as being independent random variables having this distribution. If the sample data are then chosen in a random fashion, then it is reasonable to suppose that they too are independent values from the distribution.

Definition

If $X_1, \ldots, X_n$ are independent random variables having a common distribution F, then we say that they constitute a *sample* (sometimes called a *random sample*) from the distribution F.

In most applications, the population distribution F will not be completely specified and one will attempt to use the data to make inferences about F. Sometimes it will be supposed that F is specified up to some unknown parameters (for instance, one might suppose that F was a normal distribution function having an unknown mean and variance, or that it is a Poisson distribution function whose mean is not given), and at other times it might be assumed that almost nothing is known about F (except maybe for assuming that it is a continuous, or a discrete, distribution). Problems in which the form of the underlying distribution is specified up to a set of unknown parameters are called *parametric* inference

problems, whereas those in which nothing is assumed about the form of F are called *nonparametric* inference problems.

EXAMPLE 6.1a Suppose that a new process has just been installed to produce computer chips, and suppose that the successive chips produced by this new process will have useful lifetimes that are independent with a common unknown distribution F. Physical reasons sometimes suggest the parametric form of the distribution F; for instance, it may lead us to believe that F is a normal distribution, or that F is an exponential distribution. In such cases, we are confronted with a parametrical statistical problem in which we would want to use the observed data to estimate the parameters of F. For instance, if F were assumed to be a normal distribution, then we would want to estimate its mean and variance; if F were assumed to be exponential, we would want to estimate its mean. In other situations, there might not be any physical justification for supposing that F has any particular form; in this case the problem of making inferences about F would constitute a nonparametric inference problem. ■

In this chapter, we will be concerned with the probability distributions of certain statistics that arise from a sample, where a *statistic* is a random variable whose value is determined by the sample data. Two important statistics that we will discuss are the sample mean and the sample variance. In Section 6.2, we consider the sample mean and derive its expectation and variance. We note that when the sample size is at least moderately large, the distribution of the sample mean is approximately normal. This follows from the central limit theorem, one of the most important theoretical results in probability, which is discussed in Section 6.3. In Section 6.4, we introduce the sample variance and determine its expected value. In Section 6.5, we suppose that the population distribution is normal and present the joint distribution of the sample mean and the sample variance. In Section 6.6, we suppose that we are sampling from a finite population of elements and explain what it means for the sample to be a "random sample." When the population size is large in relation to the sample size, we often treat it as if it were of infinite size; this is illustrated and its consequences are discussed.

6.2 THE SAMPLE MEAN

Consider a population of elements, each of which has a numerical value attached to it. For instance, the population might consist of the adults of a specified community and the value attached to each adult might be his or her annual income, or height, or age, and so on. We often suppose that the value associated with any member of the population can be regarded as being the value of a random variable having expectation μ and variance σ^2. The quantities μ and σ^2 are called the *population mean* and the *population variance*, respectively. Let $X_1, X_2, \ldots, X_n$ be a sample of values from this population. The sample mean is defined by

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

Since the value of the sample mean $\overline{X}$ is determined by the values of the random variables in the sample, it follows that $\overline{X}$ is also a random variable. Its expected value and variance are obtained as follows:

$$\begin{aligned} E[\overline{X}] &= E\left[\frac{X_1 + \cdots + X_n}{n}\right] \\ &= \frac{1}{n}(E[X_1] + \cdots + E[X_n]) \\ &= \mu \end{aligned}$$

and

$$\begin{aligned} \text{Var}(\overline{X}) &= \text{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) \\ &= \frac{1}{n^2}[\text{Var}(X_1) + \cdots + \text{Var}(X_n)] \quad \text{by independence} \\ &= \frac{n\sigma^2}{n^2} \\ &= \frac{\sigma^2}{n} \end{aligned}$$

where μ and σ^2 are the population mean and variance, respectively. Hence, the expected value of the sample mean is the population mean μ whereas its variance is $1/n$ times the population variance. As a result, we can conclude that $\overline{X}$ is also centered about the population mean μ, but its spread becomes more and more reduced as the sample size increases. Figure 6.1 plots the probability density function of the sample mean from a standard normal population for a variety of sample sizes.

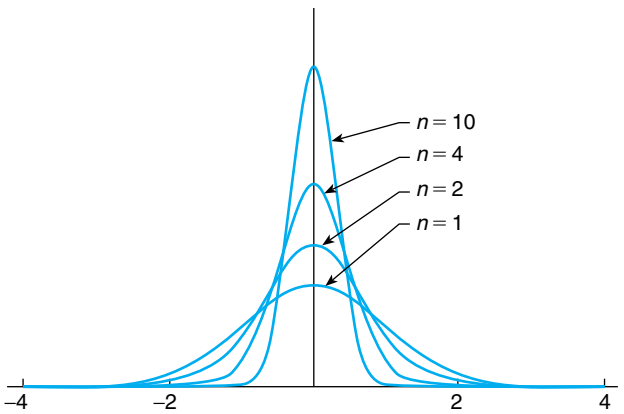


FIGURE 6.1 *Densities of sample means from a standard normal population.*

6.3 THE CENTRAL LIMIT THEOREM

In this section, we will consider one of the most remarkable results in probability — namely, the *central limit theorem*. Loosely speaking, this theorem asserts that the sum of a large number of independent random variables has a distribution that is approximately normal. Hence, it not only provides a simple method for computing approximate probabilities for sums of independent random variables, but it also helps explain the remarkable fact that the empirical frequencies of so many natural populations exhibit a bell-shaped (that is, a normal) curve.

In its simplest form, the central limit theorem is as follows:

Theorem 6.3.1 The Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then for n large, the distribution of

$$X_1 + \cdots + X_n$$

is approximately normal with mean $n\mu$ and variance $n\sigma^2$.

It follows from the central limit theorem that

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

is approximately a standard normal random variable; thus, for n large,

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < x\right\} \approx P\{Z < x\}$$

where Z is a standard normal random variable.

EXAMPLE 6.3a An insurance company has 25,000 automobile policy holders. If the yearly claim of a policy holder is a random variable with mean 320 and standard deviation 540, approximate the probability that the total yearly claim exceeds 8.3 million.

SOLUTION Let X denote the total yearly claim. Number the policy holders, and let X_i denote the yearly claim of policy holder i. With $n = 25{,}000$, we have from the central limit theorem that $X = \sum_{i=1}^{n} X_i$ will have approximately a normal distribution with mean $320 \times 25{,}000 = 8 \times 10^6$ and standard deviation $540\sqrt{25{,}000} = 8.5381 \times 10^4$. Therefore,

$$\begin{aligned} P\{X > 8.3 \times 10^6\} &= P\left\{\frac{X - 8 \times 10^6}{8.5381 \times 10^4} > \frac{8.3 \times 10^6 - 8 \times 10^6}{8.5381 \times 10^4}\right\} \\ &= P\left\{\frac{X - 8 \times 10^6}{8.5381 \times 10^4} > \frac{.3 \times 10^6}{8.5381 \times 10^4}\right\} \end{aligned}$$

$$\approx P\{Z > 3.51\} \qquad \text{where } Z \text{ is a standard normal}$$
$$\approx .00023$$

Thus, there are only 2.3 chances out of 10,000 that the total yearly claim will exceed 8.3 million. ■

EXAMPLE 6.3b Civil engineers believe that W, the amount of weight (in units of 1,000 pounds) that a certain span of a bridge can withstand without structural damage resulting, is normally distributed with mean 400 and standard deviation 40. Suppose that the weight (again, in units of 1,000 pounds) of a car is a random variable with mean 3 and standard deviation .3. How many cars would have to be on the bridge span for the probability of structural damage to exceed .1?

SOLUTION Let P_n denote the probability of structural damage when there are n cars on the bridge. That is,

$$\begin{aligned} P_n &= P\{X_1 + \cdots + X_n \geq W\} \\ &= P\{X_1 + \cdots + X_n - W \geq 0\} \end{aligned}$$

where X_i is the weight of the ith car, $i = 1, \ldots, n$. Now it follows from the central limit theorem that $\sum_{i=1}^{n} X_i$ is approximately normal with mean $3n$ and variance $.09n$. Hence, since W is independent of the X_i, $i = 1, \ldots, n$, and is also normal, it follows that $\sum_{i=1}^{n} X_i - W$ is approximately normal, with mean and variance given by

$$E\left[\sum_{1}^{n} X_i - W\right] = 3n - 400$$

$$\text{Var}\left(\sum_{1}^{n} X_i - W\right) = \text{Var}\left(\sum_{1}^{n} X_i\right) + \text{Var}(W) = .09n + 1{,}600$$

Thus,

$$\begin{aligned} P_n &= P\left\{\frac{X_1 + \cdots + X_n - W - (3n - 400)}{\sqrt{.09n + 1{,}600}} \geq \frac{-(3n - 400)}{\sqrt{.09n + 1{,}600}}\right\} \\ &\approx P\left\{Z \geq \frac{400 - 3n}{\sqrt{.09n + 1{,}600}}\right\} \end{aligned}$$

where Z is a standard normal random variable. Now $P\{Z \geq 1.28\} \approx .1$, and so if the number of cars n is such that

$$\frac{400 - 3n}{\sqrt{.09n + 1{,}600}} \leq 1.28$$

or

$$n \geq 117$$

then there is at least 1 chance in 10 that structural damage will occur. ■

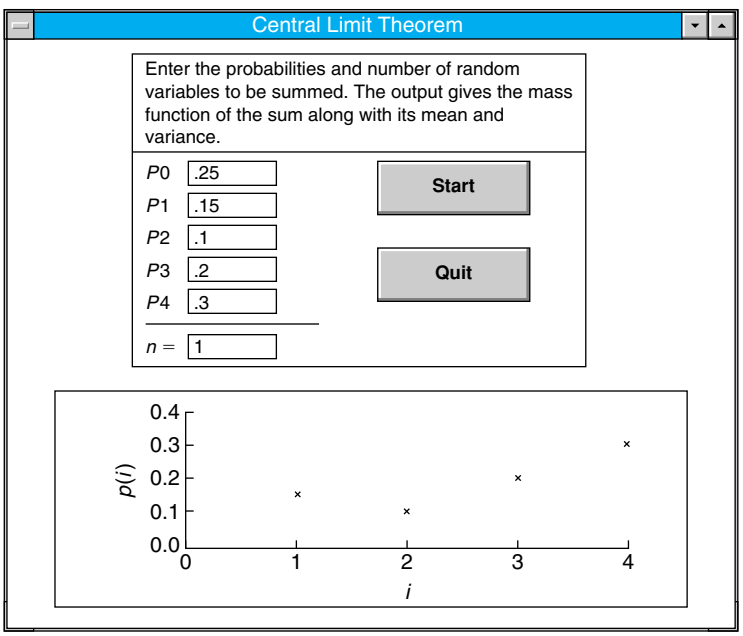
The central limit theorem is illustrated by Program 6.1 on the text disk. This program plots the probability mass function of the sum of n independent and identically distributed random variables that each take on one of the values 0, 1, 2, 3, 4. When using it, one enters the probabilities of these five values, and the desired value of n. Figures 6.2(a)–(f) give the resulting plot for a specified set of probabilities when $n = 1, 3, 5, 10, 25, 100$.

One of the most important applications of the central limit theorem is in regard to binomial random variables. Since such a random variable X having parameters (n, p) represents the number of successes in n independent trials when each trial is a success with probability p, we can express it as

$$X = X_1 + \cdots + X_n$$

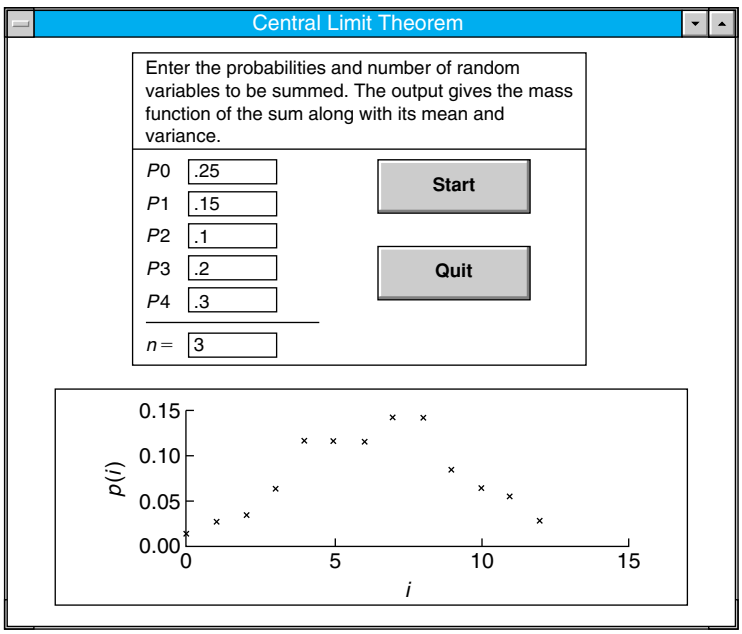
where

$$X_i = \begin{cases} 1 & \text{if the } i\text{th trial is a success} \\ 0 & \text{otherwise} \end{cases}$$

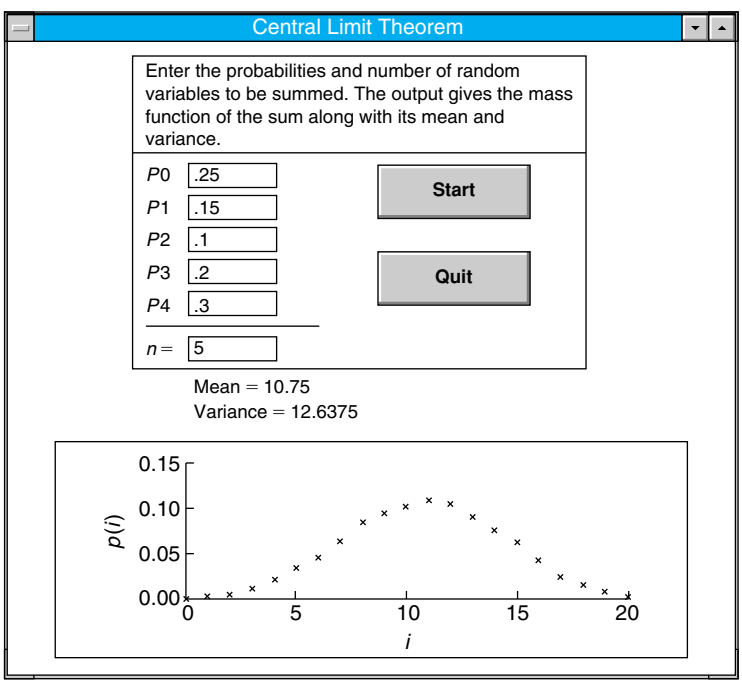


(a)

FIGURE 6.2 *(a) $n = 1$, (b) $n = 3$, (c) $n = 5$, (d) $n = 10$, (e) $n = 25$, (f) $n = 100$.*

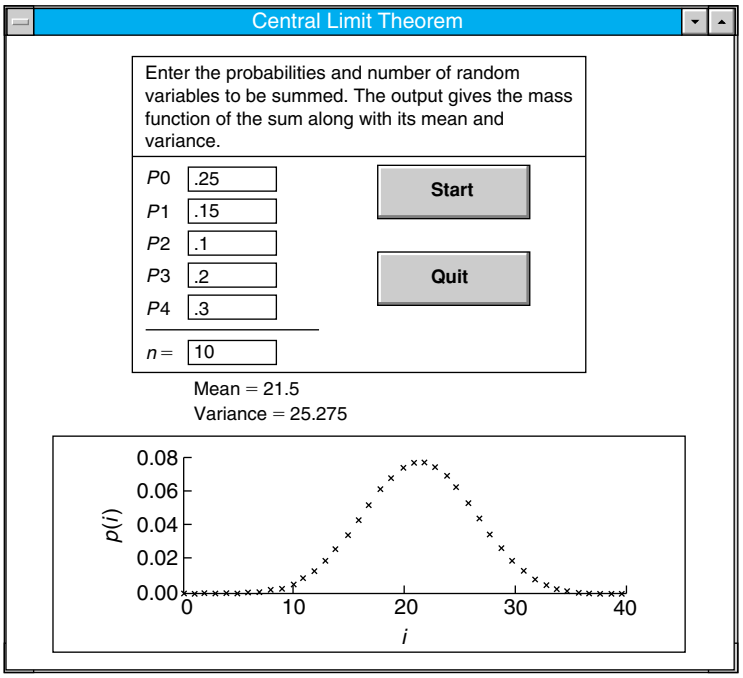


(b)

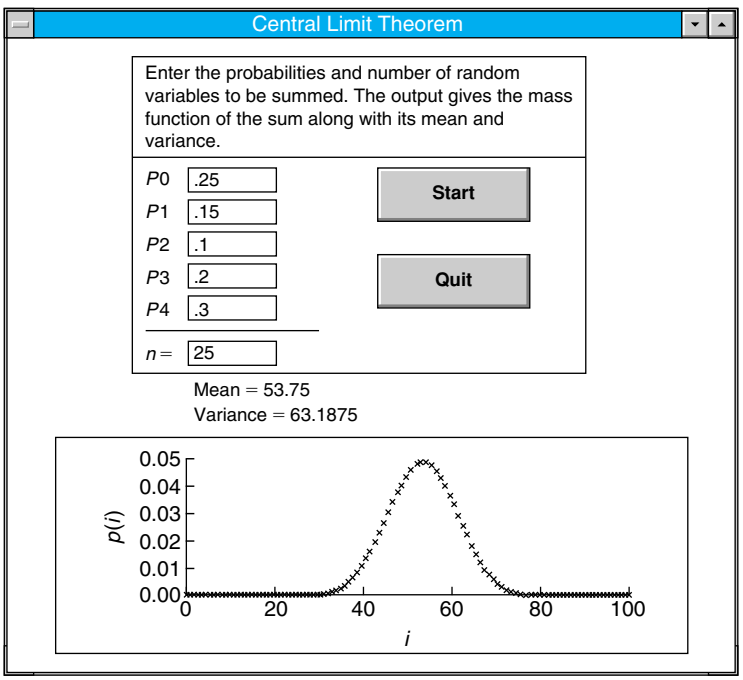


(c)

FIGURE 6.2 *(continued)*

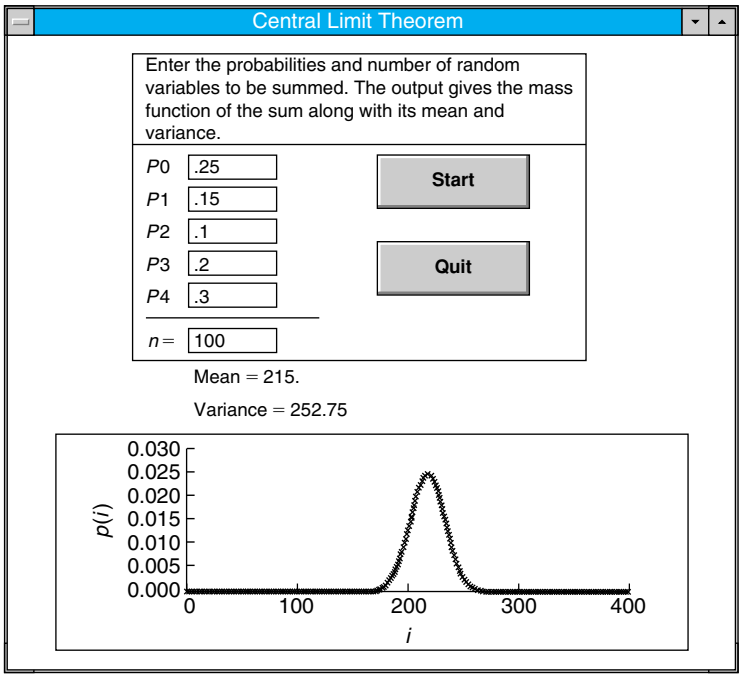


(d)



(e)

FIGURE 6.2 *(continued)*



(f)

FIGURE 6.2 *(continued)*

Because

$$E[X_i] = p, \qquad \text{Var}(X_i) = p(1-p)$$

it follows from the central limit theorem that for n large

$$\frac{X - np}{\sqrt{np(1-p)}}$$

will approximately be a standard normal random variable [see Figure 6.3, which graphically illustrates how the probability mass function of a binomial (n, p) random variable becomes more and more "normal" as n becomes larger and larger].

EXAMPLE 6.3c The ideal size of a first-year class at a particular college is 150 students. The college, knowing from past experience that, on the average, only 30 percent of those accepted for admission will actually attend, uses a policy of approving the applications of 450 students. Compute the probability that more than 150 first-year students attend this college.

SOLUTION Let X denote the number of students that attend; then assuming that each accepted applicant will independently attend, it follows that X is a binomial random

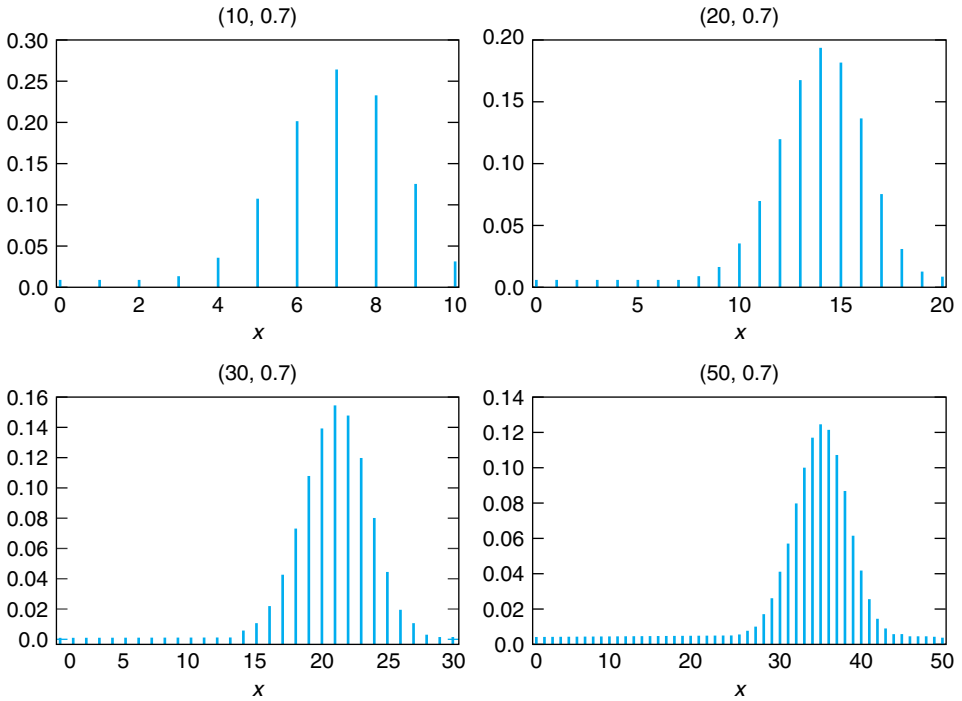


FIGURE 6.3 *Binomial probability mass functions converging to the normal density.*

variable with parameters $n = 450$ and $p = .3$. Since the binomial is a discrete and the normal a continuous distribution, it is best to compute $P\{X = i\}$ as $P\{i - .5 < X < i + .5\}$ when applying the normal approximation (this is called the continuity correction). This yields the approximation

$$P\{X > 150.5\} = P\left\{\frac{X - (450)(.3)}{\sqrt{450(.3)(.7)}} \geq \frac{150.5 - (450)(.3)}{\sqrt{450(.3)(.7)}}\right\}$$

$$\approx P\{Z > 1.59\} = .06$$

Hence, only 6 percent of the time do more than 150 of the first 450 accepted actually attend. ■

It should be noted that we now have two possible approximations to binomial probabilities: The Poisson approximation, which yields a good approximation when n is large and p small, and the normal approximation, which can be shown to be quite good when $np(1 - p)$ is large. [The normal approximation will, in general, be quite good for values of n satisfying $np(1 - p) \geq 10$.]

6.3.1 Approximate Distribution of the Sample Mean

Let $X_1, \ldots, X_n$ be a sample from a population having mean μ and variance σ^2. The central limit theorem can be used to approximate the distribution of the sample mean

$$\overline{X} = \sum_{i=1}^{n} X_i/n$$

Since a constant multiple of a normal random variable is also normal, it follows from the central limit theorem that $\overline{X}$ will be approximately normal when the sample size n is large. Since the sample mean has expected value μ and standard deviation $\sigma/\sqrt{n}$, it then follows that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

has approximately a standard normal distribution.

EXAMPLE 6.3d The weights of a population of workers have mean 167 and standard deviation 27.

(a) If a sample of 36 workers is chosen, approximate the probability that the sample mean of their weights lies between 163 and 170.

(b) Repeat part (a) when the sample is of size 144.

SOLUTION Let Z be a standard normal random variable.

(a) It follows from the central limit theorem that $\overline{X}$ is approximately normal with mean 167 and standard deviation $27/\sqrt{36} = 4.5$. Therefore, with Z being a standard normal random variable,

$$\begin{aligned} P\{163 < \overline{X} < 170\} &= P\left\{\frac{163 - 167}{4.5} < \frac{\overline{X} - 167}{4.5} < \frac{170 - 167}{4.5}\right\} \\ &\approx P\{-.8889 < Z < .8889\} \\ &= P\{Z < .8889\} - P\{Z < -.8889\} \\ &= 2P\{Z < .8889\} - 1 \\ &\approx .6259 \end{aligned}$$

(b) For a sample of size 144, the sample mean will be approximately normal with mean 167 and standard deviation $27/\sqrt{144} = 2.25$. Therefore,

$$\begin{aligned} P\{163 < \overline{X} < 170\} &= P\left\{\frac{163 - 167}{2.25} < \frac{\overline{X} - 167}{2.25} < \frac{170 - 167}{2.25}\right\} \\ &= P\left\{-1.7778 < \frac{\overline{X} - 167}{4.5} < 1.7778\right\} \\ &\approx 2P\{Z < 1.7778\} - 1 \\ &\approx .9246 \end{aligned}$$

Thus increasing the sample size from 36 to 144 increases the probability from .6259 to .9246. ■

EXAMPLE 6.3e An astronomer wants to measure the distance from her observatory to a distant star. However, due to atmospheric disturbances, any measurement will not yield the exact distance d. As a result, the astronomer has decided to make a series of measurements and then use their average value as an estimate of the actual distance. If the astronomer believes that the values of the successive measurements are independent random variables with a mean of d light years and a standard deviation of 2 light years, how many measurements need she make to be at least 95 percent certain that her estimate is accurate to within $\pm .5$ light years?

SOLUTION If the astronomer makes n measurements, then $\overline{X}$, the sample mean of these measurements, will be approximately a normal random variable with mean d and standard deviation $2/\sqrt{n}$. Thus, the probability that it will lie between $d \pm .5$ is obtained as follows:

$$\begin{aligned} P\{-.5 < \overline{X} - d < .5\} &= P\left\{\frac{-.5}{2/\sqrt{n}} < \frac{\overline{X} - d}{2/\sqrt{n}} < \frac{.5}{2/\sqrt{n}}\right\} \\ &\approx P\{-\sqrt{n}/4 < Z < \sqrt{n}/4\} \\ &= 2P\{Z < \sqrt{n}/4\} - 1 \end{aligned}$$

where Z is a standard normal random variable.

Thus, the astronomer should make n measurements, where n is such that

$$2P\{Z < \sqrt{n}/4\} - 1 \geq .95$$

or, equivalently,

$$P\{Z < \sqrt{n}/4\} \geq .975$$

Since $P\{Z < 1.96\} = .975$, it follows that n should be chosen so that

$$\sqrt{n}/4 \geq 1.96$$

That is, at least 62 observations are necessary. ■

6.3.2 How Large a Sample Is Needed?

The central limit theorem leaves open the question of how large the sample size n needs to be for the normal approximation to be valid, and indeed the answer depends on the population distribution of the sample data. For instance, if the underlying population distribution is normal, then the sample mean $\overline{X}$ will also be normal regardless of the sample size. A general rule of thumb is that one can be confident of the normal approximation whenever the sample size n is at least 30. That is, practically speaking, no matter how nonnormal the underlying population distribution is, the sample mean of a sample of size at least 30 will be approximately normal. In most cases, the normal approximation is

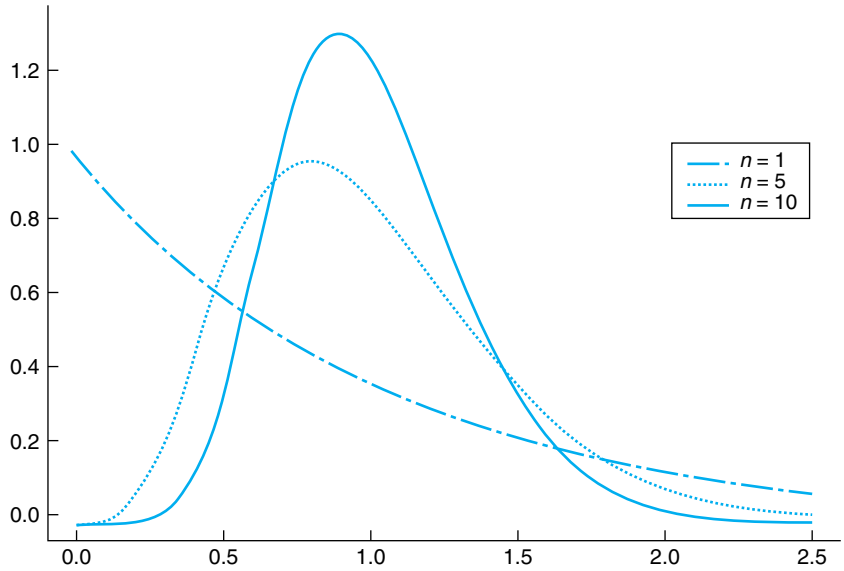


FIGURE 6.4 *Densities of the average of n exponential random variables having mean 1.*

valid for much smaller sample sizes. Indeed, a sample of size 5 will often suffice for the approximation to be valid. Figure 6.4 presents the distribution of the sample means from an exponential population distribution for samples of sizes $n = 1, 5, 10$.

6.4 THE SAMPLE VARIANCE

Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Let $\overline{X}$ be the sample mean, and recall the following definition from Section 2.3.2.

Definition

The statistic S^2, defined by

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

is called the *sample variance*. $S = \sqrt{S^2}$ is called the *sample standard deviation*.

To compute $E[S^2]$, we use an identity that was proven in Section 2.3.2: For any numbers $x_1, \ldots, x_n$

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$. It follows from this identity that

$$(n-1)S^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

Taking expectations of both sides of the preceding yields, upon using the fact that for any random variable W, $E[W^2] = \text{Var}(W) + (E[W])^2$,

$$\begin{aligned}(n-1)E[S^2] &= E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\overline{X}^2] \\ &= nE[X_1^2] - nE[\overline{X}^2] \\ &= n\text{Var}(X_1) + n(E[X_1])^2 - n\text{Var}(\overline{X}) - n(E[\overline{X}])^2 \\ &= n\sigma^2 + n\mu^2 - n(\sigma^2/n) - n\mu^2 \\ &= (n-1)\sigma^2\end{aligned}$$

or

$$E[S^2] = \sigma^2$$

That is, the expected value of the sample variance S^2 is equal to the population variance σ^2.

6.5 SAMPLING DISTRIBUTIONS FROM A NORMAL POPULATION

Let $X_1, X_2, \ldots, X_n$ be a sample from a normal population having mean μ and variance σ^2. That is, they are independent and $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$. Also let

$$\overline{X} = \sum_{i=1}^{n} X_i/n$$

and

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

denote the sample mean and sample variance, respectively. We would like to compute their distributions.

6.5.1 Distribution of the Sample Mean

Since the sum of independent normal random variables is normally distributed, it follows that $\overline{X}$ is normal with mean

$$E[\overline{X}] = \sum_{i=1}^{n} \frac{E[X_i]}{n} = \mu$$

and variance

$$\text{Var}(\overline{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) = \sigma^2/n$$

That is, $\overline{X}$, the average of the sample, is normal with a mean equal to the population mean but with a variance reduced by a factor of $1/n$. It follows from this that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

is a standard normal random variable.

6.5.2 Joint Distribution of $\overline{X}$ and S^2

In this section, we not only obtain the distribution of the sample variance S^2, but we also discover a fundamental fact about normal samples — namely, that $\overline{X}$ and S^2 are independent with $(n-1)S^2/\sigma^2$ having a chi-square distribution with $n-1$ degrees of freedom.

To start, for numbers $x_1, \ldots, x_n$, let $y_i = x_i - \mu, i = 1, \ldots, n$. Then as $\bar{y} = \bar{x} - \mu$, it follows from the identity

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

that

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \mu)^2 - n(\bar{x} - \mu)^2$$

Now, if $X_1, \ldots, X_n$ is a sample from a normal population having mean μ and variance σ^2, then we obtain from the preceding identity that

$$\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{\sigma^2} + \frac{n(\overline{X} - \mu)^2}{\sigma^2}$$

or, equivalently,

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma^2}+\left[\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}\right]^2 \tag{6.5.1}$$

Because $(X_i-\mu)/\sigma, i=1,\ldots,n$ are independent standard normals, it follows that the left side of Equation 6.5.1 is a chi-square random variable with n degrees of freedom. Also, as shown in Section 6.5.1, $\sqrt{n}(\overline{X}-\mu)/\sigma$ is a standard normal random variable and so its square is a chi-square random variable with 1 degree of freedom. Thus Equation 6.5.1 equates a chi-square random variable having n degrees of freedom to the sum of two random variables, one of which is chi-square with 1 degree of freedom. But it has been established that the sum of two independent chi-square random variables is also chi-square with a degree of freedom equal to the sum of the two degrees of freedom. Thus, it would seem that there is a reasonable possibility that the two terms on the right side of Equation 6.5.1 are independent, with $\sum_{i=1}^{n}(X_i-\overline{X})^2/\sigma^2$ having a chi-square distribution with $n-1$ degrees of freedom. Since this result can indeed be established, we have the following fundamental result.

Theorem 6.5.1

If $X_1,\ldots,X_n$ is a sample from a normal population having mean μ and variance σ^2, then $\overline{X}$ and S^2 are independent random variables, with $\overline{X}$ being normal with mean μ and variance σ^2/n and $(n-1)S^2/\sigma^2$ being chi-square with $n-1$ degrees of freedom.

Theorem 6.5.1 not only provides the distributions of $\overline{X}$ and S^2 for a normal population but also establishes the important fact that they are independent. In fact, it turns out that this independence of $\overline{X}$ and S^2 is a unique property of the normal distribution. Its importance will become evident in the following chapters.

EXAMPLE 6.5a The time it takes a central processing unit to process a certain type of job is normally distributed with mean 20 seconds and standard deviation 3 seconds. If a sample of 15 such jobs is observed, what is the probability that the sample variance will exceed 12?

SOLUTION Since the sample is of size $n=15$ and $\sigma^2=9$, write

$$\begin{aligned}P\{S^2>12\}&=P\left\{\frac{14S^2}{9}>\frac{14}{9}\cdot 12\right\}\\&=P\{\chi^2_{14}>18.67\}\\&=1-.8221 \qquad \text{from Program 5.8.1a}\\&=.1779 \quad \blacksquare\end{aligned}$$

The following corollary of Theorem 6.5.1 will be quite useful in the following chapters.

Corollary 6.5.2

Let $X_1, \ldots, X_n$ be a sample from a normal population with mean μ. If $\overline{X}$ denotes the sample mean and S the sample standard deviation, then

$$\sqrt{n}\frac{(\overline{X} - \mu)}{S} \sim t_{n-1}$$

That is, $\sqrt{n}(\overline{X} - \mu)/S$ has a t-distribution with $n - 1$ degrees of freedom.

Proof

Recall that a t-random variable with n degrees of freedom is defined as the distribution of

$$\frac{Z}{\sqrt{\chi_n^2/n}}$$

where Z is a standard normal random variable that is independent of χ_n^2, a chi-square random variable with n degrees of freedom. Because Theorem 6.5.1 gives that $\sqrt{n}(\overline{X}-\mu)/\sigma$ is a standard normal that is independent of $(n-1)S^2/\sigma^2$, which is chi-square with $n-1$ degrees of freedom, we can conclude that

$$\frac{\sqrt{n}(\overline{X} - \mu)/\sigma}{\sqrt{S^2/\sigma^2}} = \sqrt{n}\frac{(\overline{X} - \mu)}{S}$$

is a t-random variable with $n - 1$ degrees of freedom. ■

6.6 SAMPLING FROM A FINITE POPULATION

Consider a population of N elements, and suppose that p is the proportion of the population that has a certain characteristic of interest; that is, Np elements have this characteristic, and $N(1-p)$ do not. A sample of size n from this population is said to be a *random sample* if it is chosen in such a manner that each of the $\binom{N}{n}$ population subsets of size n is equally likely to be the sample. For instance, if the population consists of the three elements a, b, c, then a random sample of size 2 is one that is chosen so that each of the subsets $\{a, b\}$, $\{a, c\}$, and $\{b, c\}$ is equally likely to be the sample. A random subset can be chosen sequentially by letting its first element be equally likely to be any of the N elements of the population, then letting its second element be equally likely to be any of the remaining $N - 1$ elements of the population, and so on.

Suppose now that a random sample of size n has been chosen from a population of size N. For $i = 1, \ldots, n$, let

$$X_i = \begin{cases} 1 & \text{if the } i\text{th member of the sample has the characteristic} \\ 0 & \text{otherwise} \end{cases}$$

Consider now the sum of the X_i; that is, consider

$$X = X_1 + X_2 + \cdots + X_n$$

Because the term X_i contributes 1 to the sum if the ith member of the sample has the characteristic and 0 otherwise, it follows that X is equal to the number of members of the sample that possess the characteristic. In addition, the sample mean

$$\overline{X} = X/n = \sum_{i=1}^{n} X_i/n$$

is equal to the proportion of the members of the sample that possess the characteristic.

Let us now consider the probabilities associated with the statistics X and $\overline{X}$. To begin, note that since each of the N members of the population is equally likely to be the ith member of the sample, it follows that

$$P\{X_i = 1\} = \frac{Np}{N} = p$$

Also,

$$P\{X_i = 0\} = 1 - P\{X_i = 1\} = 1 - p$$

That is, each X_i is equal to either 1 or 0 with respective probabilities p and $1 - p$.

It should be noted that the random variables $X_1, X_2, \ldots, X_n$ are not independent. For instance, since the second selection is equally likely to be any of the N members of the population, of which Np have the characteristic, it follows that the probability that the second selection has the characteristic is $Np/N = p$. That is, without any knowledge of the outcome of the first selection,

$$P\{X_2 = 1\} = p$$

However, the conditional probability that $X_2 = 1$, given that the first selection has the characteristic, is

$$P\{X_2 = 1 | X_1 = 1\} = \frac{Np - 1}{N - 1}$$

which is seen by noting that if the first selection has the characteristic, then the second selection is equally likely to be any of the remaining $N-1$ elements, of which $Np-1$ have the characteristic. Similarly, the probability that the second selection has the characteristic given that the first one does not is

$$P\{X_2 = 1 | X_1 = 0\} = \frac{Np}{N - 1}$$

Thus, knowing whether or not the first element of the random sample has the characteristic changes the probability for the next element. However, when the population size N

is large in relation to the sample size n, this change will be very slight. For instance, if $N = 1{,}000$, $p = .4$, then

$$P\{X_2 = 1|X_1 = 1\} = \frac{399}{999} = .3994$$

which is very close to the unconditional probability that $X_2 = 1$; namely,

$$P\{X_2 = 1\} = .4$$

Similarly, the probability that the second element of the sample has the characteristic given that the first does not is

$$P\{X_2 = 1|X_1 = 0\} = \frac{400}{999} = .4004$$

which is again very close to .4.

Indeed, it can be shown that when the population size N is large with respect to the sample size n, then $X_1, X_2, \ldots, X_n$ are approximately independent. Now if we think of each X_i as representing the result of a trial that is a success if X_i equals 1 and a failure otherwise, it follows that $X = \sum_{i=1}^{n} X_i$ can be thought of as representing the total number of successes in n trials. Hence, if the X_i were independent, then X would be a binomial random variable with parameters n and p. In other words, when the population size N is large in relation to the sample size n, then the distribution of the number of members of the sample that possess the characteristic is approximately that of a binomial random variable with parameters n and p.

REMARK

Of course, X is a hypergeometric random variable (Section 5.4); and so the preceding shows that a hypergeometric can be approximated by a binomial random variable when the number chosen is small in relation to the total number of elements.

> For the remainder of this text, we will suppose that the underlying population is large in relation to the sample size and we will take the distribution of X to be binomial.

By using the formulas given in Section 5.1 for the mean and standard deviation of a binomial random variable, we see that

$$E[X] = np \quad \text{and} \quad SD(X) = \sqrt{np(1-p)}$$

Moreover $\overline{X} = X/n$, the proportion of the sample that has the characteristic, has mean and variance given by

$$E[\overline{X}] = E[X]/n = p$$

and

$$\text{Var}(\overline{X}) = \text{Var}(X)/n^2 = p(1-p)/n$$

EXAMPLE 6.6a Suppose that 45 percent of the population favors a certain candidate in an upcoming election. If a random sample of size 200 is chosen, find

(a) the expected value and standard deviation of the number of members of the sample that favor the candidate;

(b) the probability that more than half the members of the sample favor the candidate.

SOLUTION

(a) The expected value and standard deviation of the proportion that favor the candidate are

$$E[X] = 200(.45) = 90, \quad SD(X) = \sqrt{200(.45)(1-.45)} = 7.0356$$

(b) Since X is binomial with parameters 200 and .45, the text disk gives the solution

$$P\{X \geq 101\} = .0681$$

If this program were not available, then the normal approximation to the binomial (Section 6.3) could be used:

$$\begin{aligned} P\{X \geq 101\} &= P\{X \geq 100.5\} \quad \text{(the continuity correction)} \\ &= P\left\{\frac{X-90}{7.0356} \geq \frac{100.5-90}{7.0356}\right\} \\ &\approx P\{Z \geq 1.4924\} \\ &\approx .0678 \end{aligned}$$

The solution obtained by the normal approximation is correct to 3 decimal places. ■

Even when each element of the population has more than two possible values, it still remains true that if the population size is large in relation to the sample size, then the sample data can be regarded as being independent random variables from the population distribution.

EXAMPLE 6.6b According to the U.S. Department of Agriculture's *World Livestock Situation*, the country with the greatest per capita consumption of pork is Denmark. In 2013, the amount of pork consumed by a person residing in Denmark had a mean value of 147 pounds with a standard deviation of 62 pounds. If a random sample of 25 Danes is chosen, approximate the probability that the average amount of pork consumed by the members of this group in 2013 exceeded 150 pounds.

SOLUTION If we let X_i be the amount consumed by the ith member of the sample, $i = 1, \ldots, 25$, then the desired probability is

$$P\left\{\frac{X_1 + \cdots + X_{25}}{25} > 150\right\} = P\{\overline{X} > 150\}$$

where $\overline{X}$ is the sample mean of the 25 sample values. Since we can regard the X_i as being independent random variables with mean 147 and standard deviation 62, it follows from the central limit theorem that their sample mean will be approximately normal with mean 147 and standard deviation 62/5. Thus, with Z being a standard normal random variable, we have

$$\begin{aligned} P\{\overline{X} > 150\} &= P\left\{\frac{\overline{X} - 147}{12.4} > \frac{150 - 147}{12.4}\right\} \\ &\approx P\{Z > .242\} \\ &\approx .404 \quad \blacksquare \end{aligned}$$

Problems

1. Suppose that X_1, X_2, X_3 are independent with the common probability mass function

$$P\{X_i = 0\} = .2, \quad P\{X_i = 1\} = .3, \quad P\{X_i = 3\} = .5, \quad i = 1, 2, 3$$

   (a) Plot the probability mass function of $\overline{X}_2 = \dfrac{X_1 + X_2}{2}$.

   (b) Determine $E[\overline{X}_2]$ and $\text{Var}(\overline{X}_2)$.

   (c) Plot the probability mass function of $\overline{X}_3 = \dfrac{X_1 + X_2 + X_3}{3}$.

   (d) Determine $E[\overline{X}_3]$ and $\text{Var}(\overline{X}_3)$.

2. If 10 fair dice are rolled, approximate the probability that the sum of the values obtained (which ranges from 10 to 60) is between 30 and 40 inclusive.

3. Approximate the probability that the sum of 16 independent uniform (0, 1) random variables exceeds 10.

4. A roulette wheel has 38 slots, numbered 0, 00, and 1 through 36. If you bet 1 on a specified number, you either win 35 if the roulette ball lands on that number or lose 1 if it does not. If you continually make such bets, approximate the

probability that

(a) you are winning after 34 bets;
(b) you are winning after 1,000 bets;
(c) you are winning after 100,000 bets.

Assume that each roll of the roulette ball is equally likely to land on any of the 38 numbers.

5. A highway department has enough salt to handle a total of 80 inches of snowfall. Suppose the daily amount of snow has a mean of 1.5 inches and a standard deviation of .3 inches.

(a) Approximate the probability that the salt on hand will suffice for the next 50 days.
(b) What assumption did you make in solving part (a)?
(c) Do you think this assumption is justified? Explain briefly.

6. Fifty numbers are rounded off to the nearest integer and then summed. If the individual roundoff errors are uniformly distributed between $-.5$ and .5, what is the approximate probability that the resultant sum differs from the exact sum by more than 3?

7. A six-sided die, in which each side is equally likely to appear, is repeatedly rolled until the total of all rolls exceeds 400. Approximate the probability that this will require more than 140 rolls.

8. The amount of time that a certain type of battery functions is a random variable with mean 5 weeks and standard deviation 1.5 weeks. Upon failure, it is immediately replaced by a new battery. Approximate the probability that 13 or more batteries will be needed in a year.

9. The lifetime of a certain electrical part is a random variable with mean 100 hours and standard deviation 20 hours. If 16 such parts are tested, find the probability that the sample mean is

(a) less than 104;
(b) between 98 and 104 hours.

10. A tobacco company claims that the amount of nicotine in its cigarettes is a random variable with mean 2.2 mg and standard deviation .3 mg. However, the sample mean nicotine content of 100 randomly chosen cigarettes was 3.1 mg. What is the approximate probability that the sample mean would have been as high or higher than 3.1 if the company's claims were true?

11. The lifetime (in hours) of a type of electric bulb has expected value 500 and standard deviation 80. Approximate the probability that the sample mean of n such bulbs is greater than 525 when

(a) $n = 4$;
(b) $n = 16$;

(c) $n = 36$;
(d) $n = 64$.

12. An instructor knows from past experience that student exam scores have mean 77 and standard deviation 15. At present the instructor is teaching two separate classes — one of size 25 and the other of size 64.

(a) Approximate the probability that the average test score in the class of size 25 lies between 72 and 82.
(b) Repeat part (a) for a class of size 64.
(c) What is the approximate probability that the average test score in the class of size 25 is higher than that of the class of size 64?
(d) Suppose the average scores in the two classes are 76 and 83. Which class, the one of size 25 or the one of size 64, do you think was more likely to have averaged 83?

13. If X is binomial with parameters $n = 150$, $p = .6$, compute the exact value of $P\{X \leq 80\}$ and compare with its normal approximation both **(a)** making use of and **(b)** not making use of the continuity correction.

14. Each computer chip made in a certain plant will, independently, be defective with probability .25. If a sample of 1,000 chips is tested, what is the approximate probability that fewer than 200 chips will be defective?

15. A club basketball team will play a 60-game season. Thirty-two of these games are against class A teams and 28 are against class B teams. The outcomes of all the games are independent. The team will win each game against a class A opponent with probability .5, and it will win each game against a class B opponent with probability .7. Let X denote its total number of victories in the season.

(a) Is X a binomial random variable?
(b) Let X_A and X_B denote, respectively, the number of victories against class A and class B teams. What are the distributions of X_A and X_B?
(c) What is the relationship between X_A, X_B, and X?
(d) Approximate the probability that the team wins 40 or more games.

16. Argue, based on the central limit theorem, that a Poisson random variable having mean λ will approximately have a normal distribution with mean and variance both equal to λ when λ is large. If X is Poisson with mean 100, compute the exact probability that X is less than or equal to 116 and compare it with its normal approximation both when a continuity correction is utilized and when it is not. The convergence of the Poisson to the normal is indicated in Figure 6.5.

17. Use the text disk to compute $P\{X \leq 10\}$ when X is a binomial random variable with parameters $n = 100, p = .1$. Now compare this with its **(a)** Poisson and

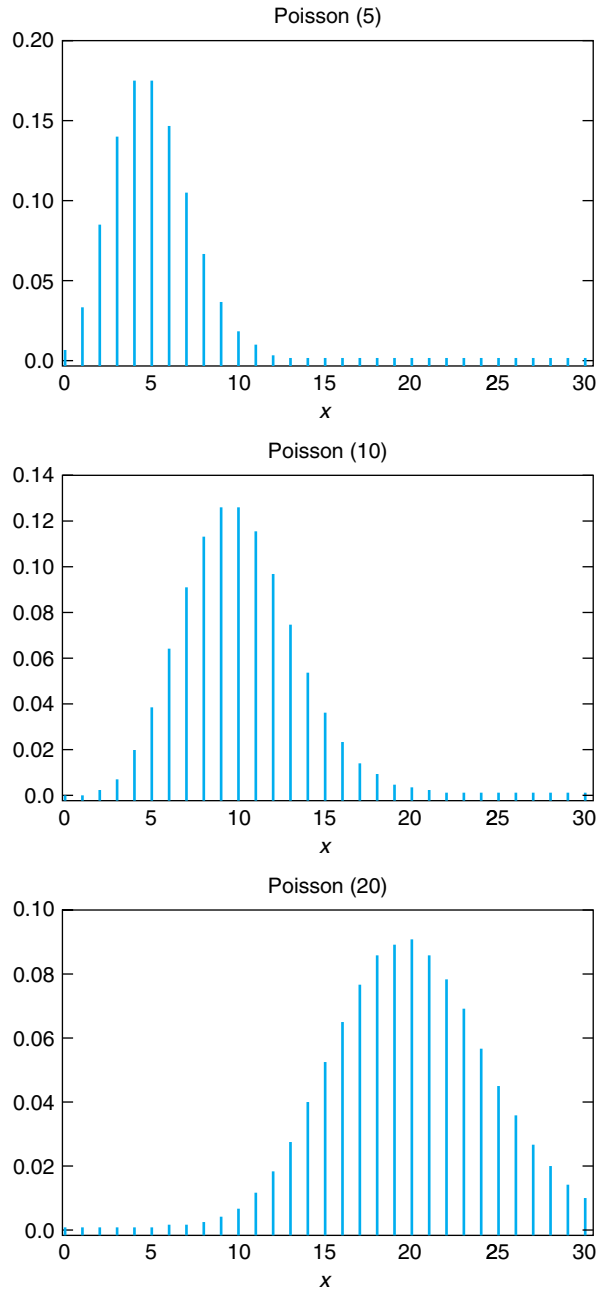


FIGURE 6.5 *Poisson probability mass functions.*

(b) normal approximation. In using the normal approximation, write the desired probability as $P\{X < 10.5\}$ so as to utilize the continuity correction.

18. The temperature at which a thermostat goes off is normally distributed with variance σ^2. If the thermostat is to be tested five times, find

(a) $P\{S^2/\sigma^2 \leq 1.8\}$
(b) $P\{.85 \leq S^2/\sigma^2 \leq 1.15\}$

where S^2 is the sample variance of the five data values.

19. In Problem 18, how large a sample would be necessary to ensure that the probability in part (a) is at least .95?

20. Consider two independent samples — the first of size 10 from a normal population having variance 4 and the second of size 5 from a normal population having variance 2. Compute the probability that the sample variance from the second sample exceeds the one from the first. (*Hint*: Relate it to the *F*-distribution.)

21. Twelve percent of the population is left-handed. Find the probability that there are between 10 and 14 left-handers in a random sample of 100 members of this population. That is, find $P\{10 \leq X \leq 14\}$, where X is the number of left-handers in the sample.

22. Fifty-two percent of the residents of a certain city are in favor of teaching evolution in high school. Find or approximate the probability that at least 50 percent of a random sample of size n is in favor of teaching evolution, when

(a) $n = 10$;
(b) $n = 100$;
(c) $n = 1{,}000$;
(d) $n = 10{,}000$.

23. The following table gives the percentages of individuals of a given city, categorized by gender, that follow certain negative health practices. Suppose a random sample of 300 men is chosen. Approximate the probability that

(a) at least 150 of them rarely eat breakfast;
(b) fewer than 100 of them smoke.

	Sleeps 6 Hours or Less per Night	Smoker	Rarely Eats Breakfast	Is 20 Percent or More Overweight
Men	22.7	28.4	45.4	29.6
Women	21.4	22.8	42.0	25.6

Source: U.S. National Center for Health Statistics, Health Promotion and Disease Prevention.

24. (Use the table from Problem 23.) Suppose a random sample of 300 women is chosen. Approximate the probability that

(a) at least 60 of them are overweight by 20 percent or more;
(b) fewer than 50 of them sleep 6 hours or less nightly.

25. (Use the table from Problem 23.) Suppose random samples of 300 women and of 300 men are chosen. Approximate the probability that more women than men rarely eat breakfast.

26. The following table uses data concerning the percentages of teenage male and female full-time workers whose annual salaries fall in different salary groupings. Suppose random samples of 1,000 men and 1,000 women were chosen. Use the table to approximate the probability that

(a) at least half of the women earned less than $20,000;
(b) more than half of the men earned $20,000 or more;
(c) more than half of the women and more than half of the men earned $20,000 or more;
(d) 250 or fewer of the women earned at least $25,000;
(e) at least 200 of the men earned $50,000 or more;
(f) more women than men earned between $20,000 and $24,999.

Earnings Range	Percentage of Women	Percentage of Men
$4,999 or less	2.8	1.8
$5,000 to $9,999	10.4	4.7
$10,000 to $19,999	41.0	23.1
$20,000 to $24,999	16.5	13.4
$25,000 to $49,999	26.3	42.1
$50,000 and over	3.0	14.9

Source: U.S. Department of Commerce, Bureau of the Census.

27. In 1995 the percentage of the labor force that belonged to a union was 14.9. If five workers had been randomly chosen in that year, what is the probability that none of them would have belonged to a union? Compare your answer to what it would be for the year 1945, when an all-time high of 35.5 percent of the labor force belonged to a union.

28. The sample mean and sample standard deviation of all San Francisco student scores on the most recent Scholastic Aptitude Test examination in mathematics were 517 and 120. Approximate the probability that a random sample of 144 students would have an average score exceeding

(a) 507;
(b) 517;
(c) 537;
(d) 550.

29. The average salary of newly graduated students with bachelor's degrees in chemical engineering is $53,600, with a standard deviation of $3,200. Approximate the probability that the average salary of a sample of 12 recently graduated chemical engineers exceeds $55,000.

30. A certain component is critical to the operation of an electrical system and must be replaced immediately upon failure. If the mean lifetime of this type of component is 100 hours and its standard deviation is 30 hours, how many of the components must be in stock so that the probability that the system is in continual operation for the next 2000 hours is at least .95?