5.1 Toward Statistical Inference

PARAMETERS AND STATISTICS

A parameter is a number that describes the **population**. A parameter is a fixed number, but in practice, we do not know its value.

A statistic is a number that describes a **sample**. The value of a statistic is known when we have taken a sample, but it can change from sample to sample. We often use a statistic to estimate an unknown parameter.

Understanding the college student market. Since 1987, *Student Monitor* has published an annual market research study that provides clients with information about the college student market. The firm uses a random sample of 1200 students located throughout the United States.[1] One phase of the research focuses on computing and technology. The firm reports that undergraduates spend an average of 19.0 hours per week on the Internet and that 88% own a cell phone.

sample proportion

The sample mean $\bar{x}$ = 19.0 hours is a *statistic*. The corresponding *parameter* is the average (call it μ) of all undergraduates enrolled in four-year colleges and universities. Similarly, the **proportion of the sample** who own a cell phone

$$\hat{p} = \frac{1056}{1200} = 0.88 = 88\%$$

population proportion

is a *statistic*. The corresponding *parameter* is the **proportion** (call it p) of all undergraduates at four-year colleges and universities who own a cell phone. We don't know the values of the parameters μ and p, so we use the statistics $\bar{x}$ and $\hat{P}$, respectively, to estimate them.

- **Shape:** The histograms look Normal. Figure 5.3 is a Normal quantile plot of the values of $\hat{P}$ for our samples of size 100. It confirms that the distribution in Figure 5.1 is close to Normal. The 1000 values for samples of size 1200 in Figure 5.2 are even closer to Normal. The Normal curves drawn through the histograms describe the overall shape quite well.
- **Center:** In both cases, the values of the sample proportion $\hat{P}$ vary from sample to sample, but the values are centered at 0.9. Recall that $p = 0.9$ is the true population parameter. Some samples have a $\hat{P}$ less than 0.9 and some greater, but there is no tendency to be always low or always high. That is, $\hat{P}$ has no *bias* as an estimator of p. This is true for both large and small samples. (Want the details? The mean of the 1000 values of $\hat{P}$ is 0.8985 for samples of size 100 and 0.8994 for samples of size 1200. The median value of is exactly 0.9 for samples of both sizes.)
- **Spread:** The values of $\hat{P}$ from samples of size 1200 are much less spread out than the values from samples of size 100. In fact, the standard deviations are 0.0304 for Figure 5.1 and 0.0083 for Figure 5.2.

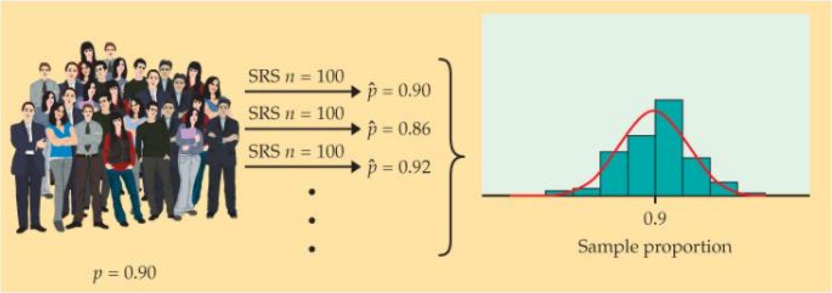


FIGURE 5.1 The results of many SRSs have a regular pattern, Example 5.3. Here we draw 1000 SRSs of size 100 from the same population. The population parameter is $p = 0.9$. The histogram shows the distribution of 1000 sample proportions.

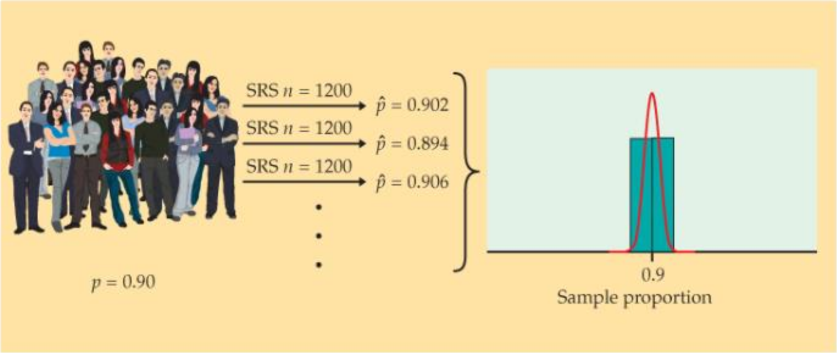


FIGURE 5.2 The distribution of the sample proportion for 1000 SRSs of size 1200 drawn from the same population as in Figure 5.1. The two histograms have the same scale. The statistic from the larger sample is less variable.

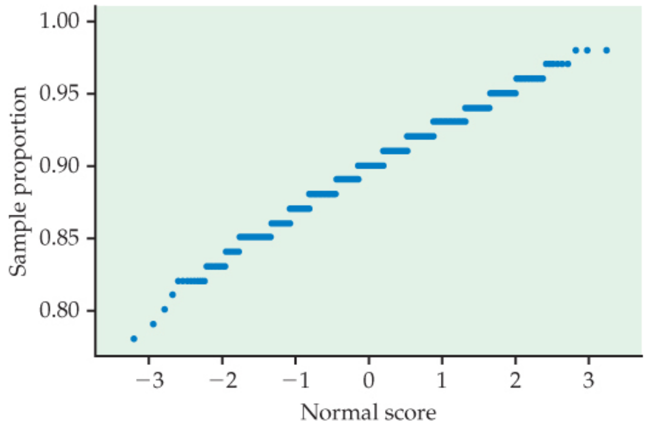


FIGURE 5.3 Normal quantile plot of the sample proportions in Figure 5.1. The distribution is close to Normal except for some clustering due to the fact that the sample proportions from a sample of size 100 can take only values that are a multiple of 0.01.

BIAS AND VARIABILITY

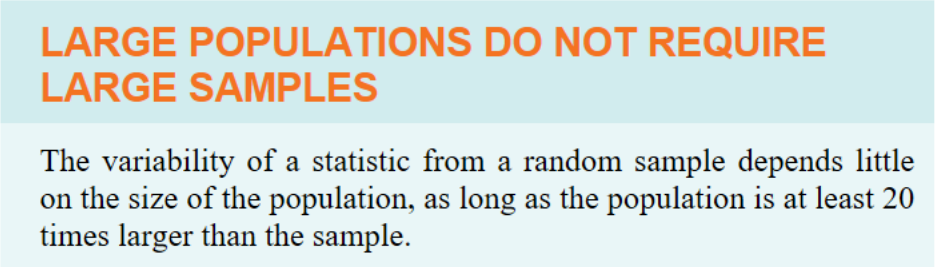
Bias concerns the center of the sampling distribution. A statistic used to estimate a parameter is an **unbiased estimator** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined by the sampling design and the sample size *n*. Statistics from larger probability samples have smaller spreads.

The **margin of error** is a numerical measure of the spread of a sampling distribution. It can be used to set bounds on the size of the likely error in using the statistic as an estimator of a population parameter.

To reduce bias, use random sampling. When we start with a list of the entire population, simple random sampling produces unbiased estimates—the values of a statistic computed from an **SRS** neither consistently overestimate nor consistently underestimate the value of the population parameter.

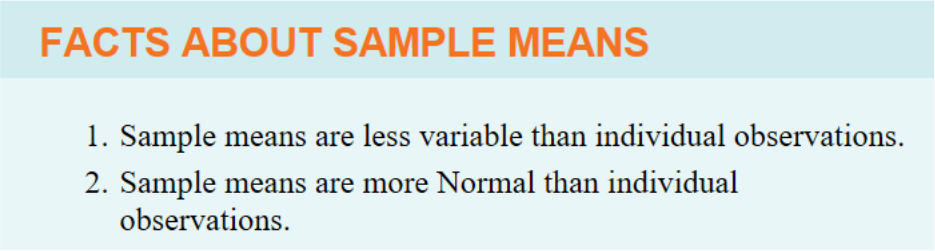
To reduce the variability of a statistic from an **SRS**, use a larger sample. You can make the variability as small as you want by taking a large enough sample.



LARGE POPULATIONS DO NOT REQUIRE LARGE SAMPLES

The variability of a statistic from a random sample depends little on the size of the population, as long as the population is at least 20 times larger than the sample.

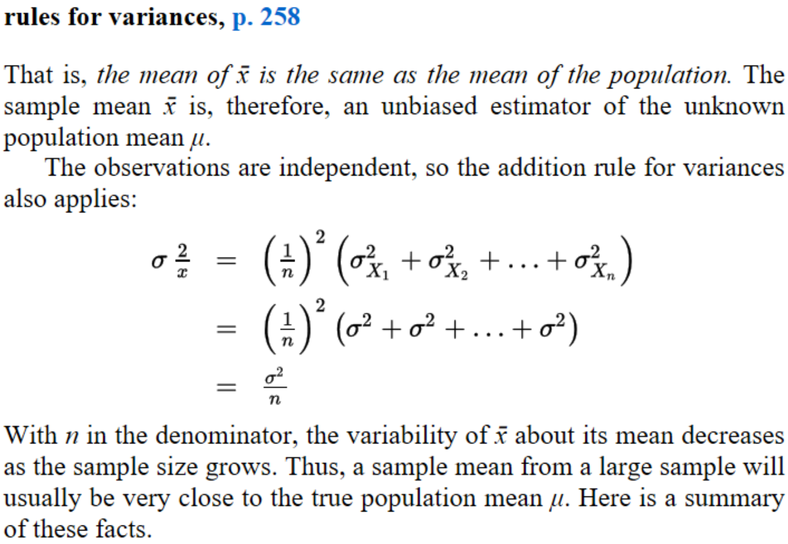
5.2 The Sampling Distribution of a Sample Mean



FACTS ABOUT SAMPLE MEANS

1. Sample means are less variable than individual observations.
2. Sample means are more Normal than individual observations.

If the population has mean μ, then μ is the mean of the distribution of each observation X_i. To get the mean of $\bar{x}$, we use the rules for means of random variables.



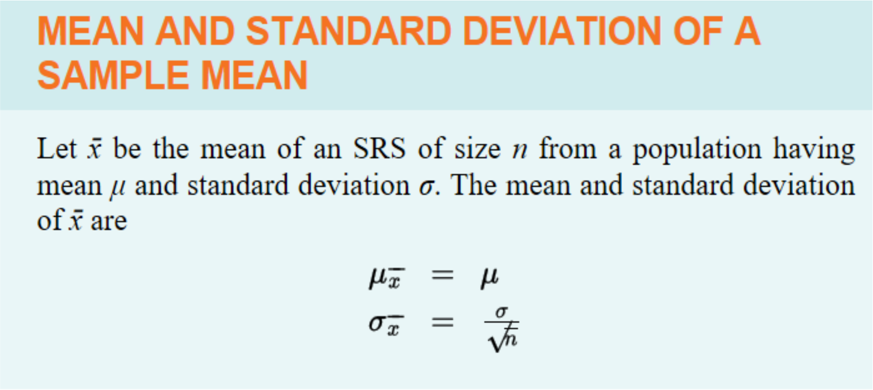
rules for variances, p. 258

That is, *the mean of $\bar{x}$ is the same as the mean of the population.* The sample mean $\bar{x}$ is, therefore, an unbiased estimator of the unknown population mean μ.

The observations are independent, so the addition rule for variances also applies:

$$\begin{aligned}\sigma^2_{\bar{x}} &= \left(\frac{1}{n}\right)^2 \left(\sigma^2_{X_1} + \sigma^2_{X_2} + \ldots + \sigma^2_{X_n}\right) \\ &= \left(\frac{1}{n}\right)^2 \left(\sigma^2 + \sigma^2 + \ldots + \sigma^2\right) \\ &= \frac{\sigma^2}{n}\end{aligned}$$

With n in the denominator, the variability of $\bar{x}$ about its mean decreases as the sample size grows. Thus, a sample mean from a large sample will usually be very close to the true population mean μ. Here is a summary of these facts.



MEAN AND STANDARD DEVIATION OF A SAMPLE MEAN

Let $\bar{x}$ be the mean of an SRS of size n from a population having mean μ and standard deviation σ. The mean and standard deviation of $\bar{x}$ are

$$\begin{aligned}\mu_{\bar{x}} &= \mu \\ \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}}\end{aligned}$$

EXAMPLE 5.6

Standard deviations for sample means of visit lengths. The standard deviation of the population of visit lengths in Figure 5.6 (a) is $\sigma = 41.84$ minutes. The length of a single visit will often be far from the population mean. If we choose an SRS of 15 visits, the standard deviation of their mean length is

$$\sigma_{\bar{x}} = \frac{41.84}{\sqrt{15}} = 10.80\text{minutes}$$

Averaging over more visits reduces the variability and makes it more likely that $\bar{x}$ is close to μ. Our sample size of 60 visits is 4 times 15, so the standard deviation will be half as large:

$$\sigma_{\bar{x}} = \frac{41.84}{\sqrt{60}} = 5.40\text{minutes}$$

SAMPLING DISTRIBUTION OF A SAMPLE MEAN

If a population has the $N(\mu, \sigma)$ distribution, then the sample mean $\bar{x}$ of n independent observations has the $N(\mu, \sigma/\sqrt{n})$ distribution.

CENTRAL LIMIT THEOREM

Draw an SRS of size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

EXAMPLE 5.8

How can we reduce the standard deviation? In the setting of Example 5.7, if we want to reduce the standard deviation of $\bar{x}$ by a factor of 2, we must take a sample four times as large, $n = 4 \times 60$, or 240. Then

$$\sigma_{\bar{x}} = \frac{41.84}{\sqrt{240}} = 2.70\text{minutes}$$

For samples of size 240, about 95% of the sample means will be within twice 2.70, or 5.40 minutes, of the population mean μ.

EXAMPLE 5.11

Time between snaps. Snapchat has more than 100 million daily users sending well over 400 million snaps a day.[6] Suppose that the time X between snaps received is governed by the exponential distribution with mean $\mu = 15$ minutes and standard deviation $\sigma = 15$ minutes. You record the next 50 times between snaps. What is the probability that their average exceeds 13 minutes?

The central limit theorem says that the sample mean time $\overline{x}$ (in minutes) between snaps has approximately the Normal distribution with mean equal to the population mean $\mu = 15$ minutes and standard deviation

$$\frac{\sigma}{\sqrt{50}} = \frac{15}{\sqrt{50}} = 2.12\text{minutes}$$

The sampling distribution of $\overline{x}$ is, therefore, approximately $N(15,2.12)$. Figure 5.10 shows this Normal curve (solid) and also the actual density curve of $\overline{x}$ (dashed).

The probability we want is $P(\overline{x} > 13.0)$. This is the area to the right of 13 under the solid Normal curve in Figure 5.10. A Normal distribution calculation gives

$$\begin{aligned} P(\overline{x} > 13.0) &= P\left(\frac{\overline{x}-15}{2.12} > \frac{13.0-15}{2.12}\right) \\ &= P(Z > -0.94) = 0.8264 \end{aligned}$$

Getting to and from campus. You live off campus and take the shuttle, provided by your apartment complex, to and from campus. Your time on the shuttle in minutes varies from day to day. The time going to campus X has the $N(20,4)$ distribution, and the time returning from campus Y varies according to the $N(18, 8)$ distribution. If they vary independently, what is the probability that you will be on the shuttle for less time going to campus?

The difference in times $X - Y$ is Normally distributed, with mean and variance

$$\mu_{X-Y} = \mu_x - \mu_Y = 20 - 18 = 2$$

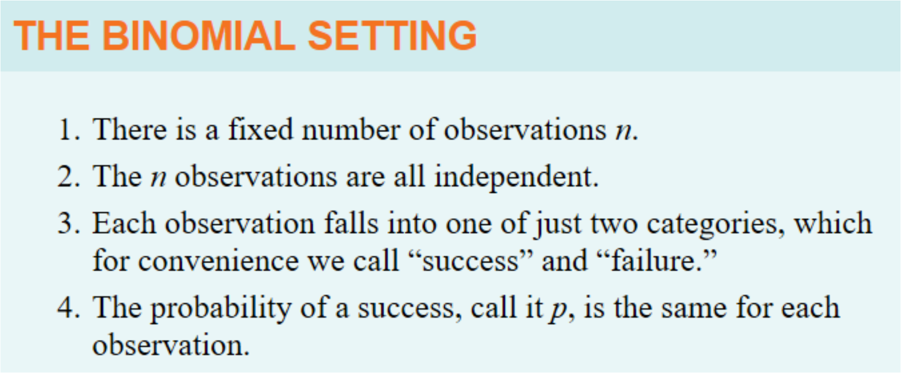
$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y = 4^2 + 8^2 = 80$$

Because $\sqrt{80} = 8.94$, $X - \mathrm{Y}$ has the $N(2, 8.94)$ distribution. Figure 5.12 illustrates the probability computation:

$$\begin{aligned} P(X < Y) &= P(X - Y < 0) \\ &= P\left(\frac{(X-Y)-2}{8.94} < \frac{0-2}{8.94}\right) \\ &= P(Z < -0.22) = 0.4129 \end{aligned}$$

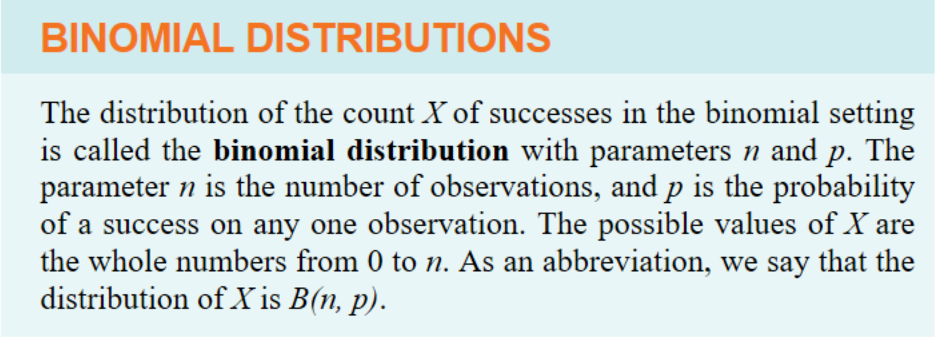
Although, on average, it takes longer to go to campus than return, the trip to campus will take less time on roughly two of every five days.

5.3 Sampling Distributions for Counts and Proportions



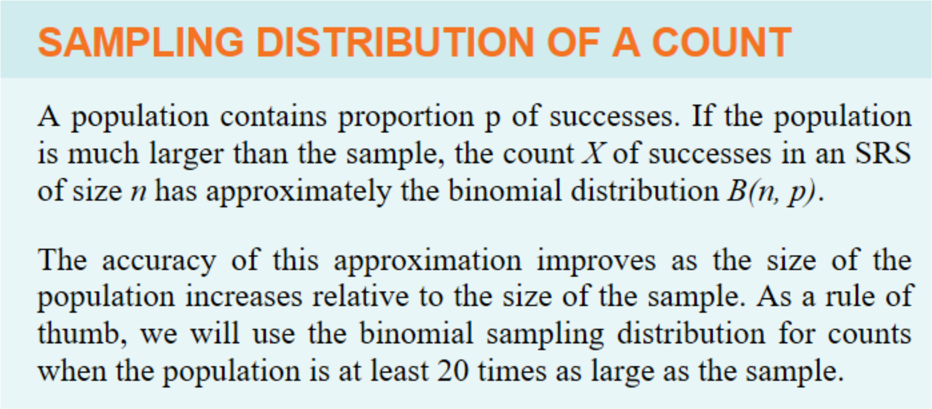
THE BINOMIAL SETTING

1. There is a fixed number of observations n.
2. The n observations are all independent.
3. Each observation falls into one of just two categories, which for convenience we call “success” and “failure.”
4. The probability of a success, call it p, is the same for each observation.



BINOMIAL DISTRIBUTIONS

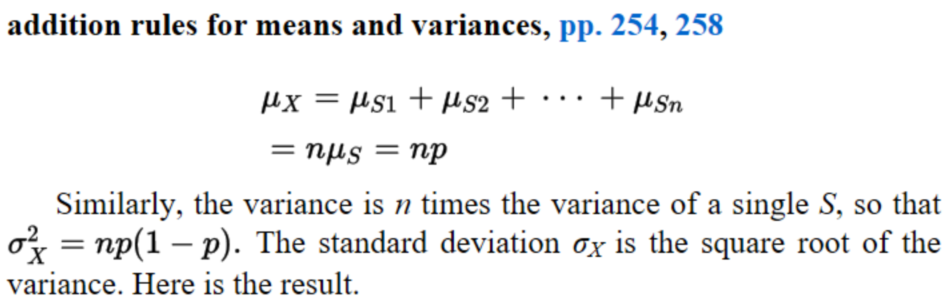
The distribution of the count X of successes in the binomial setting is called the **binomial distribution** with parameters n and p. The parameter n is the number of observations, and p is the probability of a success on any one observation. The possible values of X are the whole numbers from 0 to n. As an abbreviation, we say that the distribution of X is $B(n, p)$.



SAMPLING DISTRIBUTION OF A COUNT

A population contains proportion p of successes. If the population is much larger than the sample, the count X of successes in an SRS of size n has approximately the binomial distribution $B(n, p)$.

The accuracy of this approximation improves as the size of the population increases relative to the size of the sample. As a rule of thumb, we will use the binomial sampling distribution for counts when the population is at least 20 times as large as the sample.



addition rules for means and variances, pp. 254, 258

$$\begin{aligned}\mu_X &= \mu_{S1} + \mu_{S2} + \cdots + \mu_{Sn} \\ &= n\mu_S = np\end{aligned}$$

Similarly, the variance is n times the variance of a single S, so that $\sigma_X^2 = np(1-p)$. The standard deviation σ_X is the square root of the variance. Here is the result.

BINOMIAL MEAN AND STANDARD DEVIATION

If a count *X* has the binomial distribution *B(n, p)*, then

$$\mu_X = np$$
$$\sigma_X = \sqrt{np(1-p)}$$

EXAMPLE 5.23

The Helsinki Heart Study. The Helsinki Heart Study asked whether the anticholesterol drug gemfibrozil reduces heart attacks. In planning such an experiment, the researchers must be confident that the sample sizes are large enough to enable them to observe enough heart attacks. The Helsinki study planned to give gemfibrozil to about 2000 men aged 40 to 55 and a placebo to another 2000. The probability of a heart attack during the five-year period of the study for men this age is about 0.04. What are the mean and standard deviation of the number of heart attacks that will be observed in one group if the treatment does not change this probability?

There are 2000 independent observations, each having probability *p* = 0.04 of a heart attack. The count *X* of heart attacks has the *B*(2000, 0.04) distribution, so that

$$\mu_X = np = (2000)\ (0.04)\ = 80$$
$$\sigma_X = \sqrt{np(1-p)} = \sqrt{(2000)\ (0.04)\ (0.96)} = 8.76$$

The expected number of heart attacks is large enough to permit conclusions about the effectiveness of the drug. In fact, there were 84 heart attacks among the 2035 men actually assigned to the placebo, quite close to the mean. The gemfibrozil group of 2046 men suffered only 56 heart attacks. This is evidence that the drug reduces the chance of a heart attack. In a later chapter, we will learn how to determine if this is strong enough evidence to conclude the drug is effective.

Sample proportions

What proportion of a company's sales records have an incorrect sales tax classification? What percent of adults favor stronger laws restricting firearms? In statistical sampling, we often want to estimate the proportion p of "successes" in a population. Our estimator is the sample proportion of successes:

LOOK BACK

population proportion, p. 283

$$\begin{aligned} \hat{p} &= \frac{\text{count of successes in sample}}{\text{size of sample}} \\ &= \frac{X}{n} \end{aligned}$$

MEAN AND STANDARD DEVIATION OF A SAMPLE PROPORTION

Let $\hat{p}$ be the sample proportion of successes in an SRS of size n drawn from a large population having population proportion p of successes. The mean and standard deviation of $\hat{p}$ are

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

The formula for $\sigma_{\hat{p}}$ is exactly correct in the binomial setting. It is approximately correct for an SRS from a large population. We will use it when the population is at least 20 times as large as the sample

NORMAL APPROXIMATION FOR COUNTS AND PROPORTIONS

Draw an SRS of size n from a large population having population proportion p of successes. Let X be the count of successes in the sample and $\hat{p} = X/n$ be the sample proportion of successes. When n is large, the sampling distributions of these statistics are approximately Normal:

$$X\text{isapproximately}N\left(np, \sqrt{np\,(1-p)}\right)$$

$$\hat{p}\text{isapproximately}N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

As a rule of thumb, we will use this approximation for values of n and p that satisfy $np \geq 10$ and $n(1\text{-}p) \geq 10$.

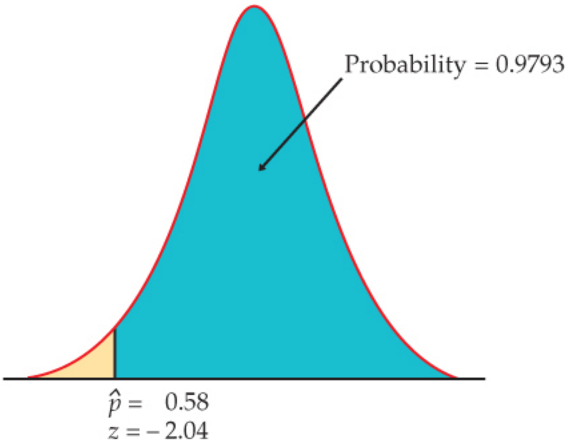
Compare the Normal approximation with the exact calculation. Let's compare the Normal approximation for the calculation of Example 5.24 with the exact calculation from software. We want to calculate $P(\hat{p} \geq 0.58)$ when the sample size is $n = 2500$ and the population proportion is $p = 0.6$. Example 5.25 shows that

$$\mu_{\hat{p}} = p = 0.6$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = 0.0098$$

Act as if $\hat{p}$ were Normal with mean 0.6 and standard deviation 0.0098. The approximate probability, as illustrated in Figure 5.18, is

$$\begin{aligned} P\left(\hat{p} \geq 0.58\right) &= P\left(\frac{\hat{p}-0.6}{0.0098} \geq \frac{0.58-0.6}{0.0098}\right) \\ &\doteq P\left(Z \geq -2.04\right) = 0.9793 \end{aligned}$$



EXAMPLE 5.27

Using the Normal approximation. The audit described in Example 5.19 examined an SRS of 150 sales records for compliance with sales tax laws. In fact, 8% of all the company's sales records have an incorrect sales tax classification. The count X of bad records in the sample has approximately the $B(150, 0.08)$ distribution.

According to the Normal approximation to the binomial distributions, the count X is approximately Normal with mean and standard deviation

$$\mu_X = np = (150)\,(0.08) = 12$$
$$\sigma_X = \sqrt{np\,(1-p)} = \sqrt{(150)\,(0.08)\,(0.92)} = 3.3226$$

The Normal approximation for the probability of no more than 10 misclassified records is the area to the left of $X = 10$ under the Normal curve. Using Table A,

$$\begin{aligned} P\,(X \le 10) &= P\left(\frac{X-12}{3.3226} \le \frac{10-12}{3.3226}\right) \\ &\doteq P\,(Z \le -0.60) = 0.2743 \end{aligned}$$

BINOMIAL COEFFICIENT

The number of ways of arranging k successes among n observations is given by the **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

for $k = 0, 1, 2, \ldots, n$.

factorial

The formula for binomial coefficients uses the **factorial** notation. The factorial $n!$ for any positive whole number n is

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

Also, $0! = 1$. Notice that the larger of the two factorials in the denominator of a binomial coefficient will cancel much of the $n!$ in the numerator. For example, the binomial coefficient we need for Example 5.28 is

$$\begin{aligned} \binom{5}{2} &= \frac{5!}{2!3!} \\ &= \frac{(5)(4)(3)(2)(1)}{(2)(1)\times(3)(2)(1)} \\ &= \frac{(5)(4)}{(2)(1)} = \frac{20}{2} = 10 \end{aligned}$$

This agrees with our previous calculation.

The notation $\binom{n}{k}$ *is not related to the fraction* $\frac{n}{k}$. A helpful way to remember its meaning is to read it as "binomial coefficient n choose k." Binomial coefficients have many uses in mathematics, but we are interested in them only as an aid to finding binomial probabilities. The binomial coefficient $\binom{n}{k}$ counts the number of ways in which k successes can be distributed among n observations. The binomial probability $P(X = k)$ is this count multiplied by the probability of any specific arrangement of the k successes. Here is the formula we seek.

BINOMIAL PROBABILITY

If X has the binomial distribution $B(n, p)$ with n observations and probability p of success on each observation, the possible values of X are 0, 1, 2, . . . , n. If k is any one of these values, the **binomial probability** is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

EXAMPLE 5.29

Using the binomial probability formula. The number X of misclassified sales records in the auditor's sample in Example 5.21 (page 316) has the $B(15,0.08)$ distribution. The probability of finding no more than one misclassified record is

$$\begin{aligned} P(X \leq 1) &= P(X=0) + P(X=1) \\ &= \binom{15}{0}(0.08)^0(0.92)^{15} + \binom{15}{1}(0.08)^1(0.92)^{14} \\ &= \frac{15!}{0!15!}\ (1)\ (0.2863) + \frac{15!}{1!14!}\ (0.08)\ (0.3112) \\ &= (1)\ (1)\ (0.2863) + \ (15)\ (0.08)\ (0.3112) \\ &= 0.2863 + 0.3734 = 0.6597 \end{aligned}$$

The calculation used the facts that 0! 5 1 and that $a^0 = 1$ for any number $a \neq 0$. The result agrees with that obtained from Table C in Example 5.21.

- A count X of successes has a **Poisson distribution** in the **Poisson setting**: the number of successes that occur in two nonoverlapping units of measure are independent; the probability that a success will occur in a unit of measure is the same for all units of equal size and is proportional to the size of the unit; the probability that more than one event occurs in a unit of measure is negligible for very small-sized units. In other words, the events occur one at a time.
- If X has the Poisson distribution with mean μ, then the standard deviation of X is $\sqrt{\mu}$, and the possible values of X are the whole numbers 0, 1, 2, 3, and so on.
- The **Poisson probability** that X takes any of these values is

  $$P\left(X = k\right) = \frac{e^{-\mu}\mu^k}{k!} \quad k = 0, 1, 2, 3, \ldots$$

  Sums of independent Poisson random variables also have the Poisson distribution. For example, in a Poisson model with mean μ per unit of measure, the count of successes in a units is a Poisson random variable with mean $a\mu$.