5.2 Sampling Distributions for Counts and Proportions

When you complete this section, you will be able to

- **Determine when the count X can be modeled using the binomial distribution.**
- **Determine when the sampling distribution of X can be modeled using the binomial distribution.**
- **Calculate the mean and standard deviation of X when it has the $B(n, p)$ distribution.**
- **Explain the differences in the sampling distributions of a count X and the associated sample proportion $\hat{p} = X/n$.**
- **Determine when one can utilize the Normal approximation to describe the sampling distribution of the count or the sampling distribution of the sample proportion.**
- **Use the Normal approximation for counts and proportions to perform probability calculations about the statistics.**

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.

Binomial mean and standard deviation

If a count X has the $B(n, p)$ distribution, what are the mean μ_X and the standard deviation σ_X? We can guess the mean. If we expect 46% of the students to have fallen asleep in class due to poor sleep, the mean number in 12 students should be 46% of 12, or 5.5. That's μ_X when X has the $B(12, 0.46)$ distribution.

Intuition suggests more generally that the mean of the $B(n, p)$ distribution should be np. Can we show that this is correct and also obtain a short formula for the standard deviation? Because binomial distributions are discrete probability distributions, we could find the mean and variance by using the definitions in Section 4.4. Here is an easier way.

LOOK BACK
means and variances of random variables, p. 263

A binomial random variable X is the count of successes in n independent observations that each have the same probability p of success. Let the random variable S_i indicate whether the ith observation is a success or failure by taking the values $S_i = 1$ if a success occurs and $S_i = 0$ if the outcome is a failure. The S_i are independent because the observations are, and each S_i has the same simple distribution:

Outcome	1	0
Probability	p	$1 - p$

LOOK BACK
mean and variance of a discrete random variable, p. 279

From the definition of the mean of a discrete random variable, we know that the mean of each S_i is

$$\mu_S = (1)(p) + (0)(1 - p) = p$$

Similarly, the definition of the variance shows that $\sigma_S^2 = p(1 - p)$. Because each S_i is 1 for a success and 0 for a failure, to find the total number of successes X we add the S_i's:

$$X = S_1 + S_2 + \cdots + S_n$$

Apply the addition rules for means and variances to this sum. To find the mean of X we add the means of the S_i's:

$$\begin{aligned}\mu_X &= \mu_{S_1} + \mu_{S_2} + \cdots + \mu_{S_n} \\ &= n\mu_S = np\end{aligned}$$

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

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

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:

proportion

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

Be sure to distinguish between the proportion $\hat{p}$ and the count X. The count takes whole-number values between 0 and n, but a proportion is always a number between 0 and 1. In the binomial setting, the count X has a binomial distribution. The proportion $\hat{p}$ does *not* have a binomial distribution. We can, however, do probability calculations about $\hat{p}$ by restating them in terms of the count X and using binomial methods.

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.

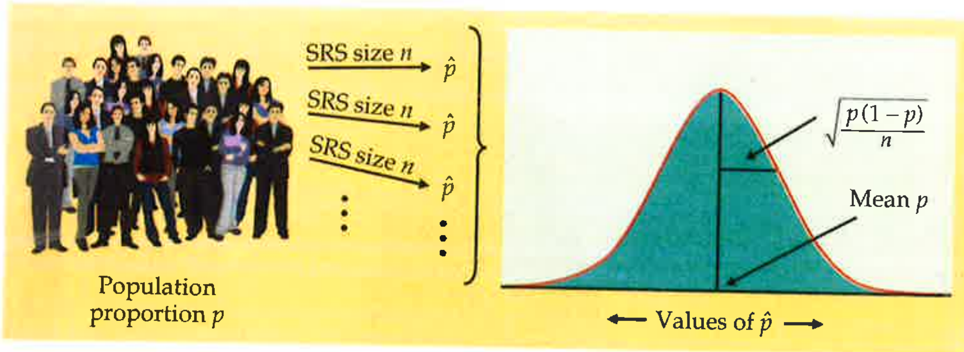
NORMAL APPROXIMATION FOR COUNTS AND PROPORTIONS

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:

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

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

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



The accuracy of the Normal approximations improves as the sample size n increases. They are most accurate for any fixed n when p is close to 1/2, and least accurate when p is near 0 or 1. You can compare binomial distributions with their Normal approximations by using the *Normal Approximation to Binomial* applet. This applet allows you to change n or p while watching the effect on the binomial probability histogram and the Normal curve that approximates it.

Figure 5.12 summarizes the distribution of a sample proportion in a form that emphasizes the big idea of a sampling distribution. Just as with Figure 5.6, the general framework for constructing a sampling distribution is shown on the left.

- **Take many random samples of size n from a population that contains proportion p of successes.**
- **Find the sample proportion $\hat{p}$ for each sample.**
- **Collect all the $\hat{p}$'s and display their distribution.**

The sampling distribution of $\hat{p}$ is shown on the right. Keep this figure in mind as you move toward statistical inference.

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

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

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