4.3 Sample Proportions

Definition. A *population proportion* is the proportion of individuals in a population sharing a certain trait, denoted p. The *sample proportion* is the proportion of individuals in a sample sharing a certain trait, denoted $\hat{p}$.

The Sampling Distribution of $\hat{p}$

Note. How good is the statistic $\hat{p}$ as an estimate of the parameter p? To find out, we ask, "What would happen if we took many samples?" The sampling distribution of $\hat{p}$ answers this question. In the simulation examples in Section 4.1, we found:

- The sampling distribution of the sample proportion $\hat{p}$ has a shape that is close to normal.
- Its mean is close to the population proportion p.
- Its standard deviation gets smaller as the size of the sample gets larger.

Definition. Choose an SRS of size n from a large population with population proportion p having some characteristic of interest. Let p be the proportion of the sample having that characteristic. Then:

- The sampling distribution of $\hat{p}$ is *approximately normal* and is closer to a normal distribution when the sample size n is large.

- The *mean* of the sampling distribution is exactly p.
- The *standard deviation* of the sampling distribution is

$$\sqrt{\frac{p(1-p)}{n}}.$$

Note. As a rule of thumb, use the recipe for the standard deviation of $\hat{p}$ only when the population is at least 10 times as large as the sample.

Example 4.14. You ask an SRS of 1500 first-year college students whether they applied for admission to any other college. There are over 1.7 million first-year college students, so the rule of thumb is easily satisfied. In fact, 35% of all first-year students applied to colleges besides the one they are attending. What is the probability that your sample will give a result within 2 percentage points of this true value? We have an SRS of $n = 1500$ drawn from a population in which the proportion $p = .35$ applied to other colleges. The sample proportion $\hat{p}$ has mean 0.35 and standard deviation

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(.35)(.65)}{1500}} = .0123.$$

We want the probability that $\hat{p}$ falls between 0.33 and 0.37 (within 2 percentage points, or 0.02, of 0.35). This is a normal distribution calculation. Standardize $\hat{p}$ by subtracting its mean 0.35 and dividing by its standard deviation 0.123. That produces a new statistic that has the standard normal distribution. It is usual to call such a statistic Z:

$$Z = \frac{\hat{p} - .35}{.0123}.$$

Then draw a picture of the areas under the standard normal curve

(Figure 4.14 and TM-70), and use Table A (TM-139, TM-140) to find them. Here is the calculation.

$$\begin{aligned} P(.33 \le \hat{p} \le .37) &= P\left(\frac{.33-.35}{.0123} \le \frac{\hat{p}-.35}{.0123} \le \frac{.37-.35}{.0123}\right) \\ &= P(-1.63 \le Z \le 1.63) = .9484 - .0516 = .8968. \end{aligned}$$

We see that almost 90% of all samples will give a result within 2 percentage points of the truth about the population.

Using the Normal Approximation for $\hat{p}$

Note. As a second rule of thumb, we will use the normal approximation to the sampling distribution of $\hat{p}$ for values of n and p that satisfy $np \ge 10$ and $n(1-p) \ge 10$.

Example 4.15. One way of checking the effect of undercoverage, nonresponse, and other sources of error in a sample survey is to compare the sample with known facts about the population. About 11% of American adults are black. The proportion $\hat{p}$ of blacks in an SRS of 1500 adults should therefore be close to 11%. It is unlikely to be exactly 11% because of sampling variability. If a national sample contains only 9.2% blacks, should we suspect that the sampling procedure is somehow underrepresenting blacks? We will find the probability that a sample contains no more than 9.2% blacks when the population is 11% black. First, check our rule of thumb for using the normal approximation to the sampling distribution of $\hat{p}$: $np = (1500)(.11) = 165$ and $n(1-p) = (1500)(.89) = 1335$. Both are much larger than 10, so the

approximation will be quite accurate. The mean of $\hat{p}$ is $p = .11$. The standard deviation is

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(.11)(.89)}{1500}} = .00808.$$

Now do the normal probability calculation illustrated in Figure 4.15 (and TM-71):

$$P(\hat{p} \leq .092) = P\left(\frac{\hat{p} - .11}{.00808} \leq \frac{.092 - .11}{.00808}\right) = P(Z \leq -2.23) = .0129.$$

Only 1.29% of all samples would have so few blacks. Because it is unlikely that a sample would include so few blacks, we have good reason to suspect that the sampling procedure underrepresents blacks.

Sample Counts

Note. Sometimes we are interested in the *count* of special individuals in a sample rather than the proportion of such individuals. To deal with these problems, just restate them in term of proportions.