8.1 Inference for a Single Proportion

In statistical terms, we are concerned with inference about the probability p of a success in the binomial setting. The sample proportion of successes $\hat{p} = X/n$ estimates the unknown population proportion p. If the population is much larger than the sample (at least 20 times as large), the count X has approximately the binomial distribution $B(n, p)$.[1]

Robotics and jobs. A Pew survey asked a panel of experts whether or not they thought that networked, automated, artificial intelligence (AI), and robotic devices will have displaced more jobs than they have created (net jobs) by 2025.[2]

The sample size is the number of experts who responded to the Pew survey question, $n = 1896$. The report on the survey tells us that 48% of the respondents said they "believe net jobs will decrease by 2025 due to networked, automated, artificial intelligence (AI), and robotic devices." Thus, the sample proportion is $\hat{p} = 0.48$. We can calculate the count X from the information given; it is the sample size times the proportion responding Yes, $X = n\hat{p} = 1896\,(0.48) = 910$.

Large-sample confidence interval for a single proportion

The unknown population proportion p is estimated by the sample proportion $\hat{p} = X/n$. If the sample size n is sufficiently large, the sampling distribution of $\hat{p}$ is approximately Normal, with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{p\,(1-p)\,/n}$. This means that approximately 95% of the time $\hat{p}$ will be within $2\sqrt{p\,(1-p)\,/n}$ of the unknown population proportion p.

LARGE-SAMPLE CONFIDENCE INTERVAL FOR A POPULATION PROPORTION

Choose an SRS of size n from a large population with an unknown proportion p of successes. The **sample proportion** is

$$\hat{p} = \frac{X}{n}$$

where X is the number of successes. The **standard error of $\hat{\mathbf{p}}$** is

$$\mathrm{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

and the **margin of error** for confidence level C is

$$m = z^{*}\mathrm{SE}_{\hat{p}}$$

where the critical value z^* is the value for the standard Normal density curve with area C between $-z^*$ and z^*.

An **approximate level C confidence interval** for p is

$$\hat{p} \pm m$$

Use this interval for 90% ($z^* = 1.645$), 95% ($z^* = 1.96$), or 99% ($z^* = 2.576$) confidence when the number of successes and the number of failures are both at least 10.

Inference for robotics and jobs. The sample survey in Example 8.1 found that 910 of a sample of 1896 experts reported that they think net jobs will decrease by 2025 because of robots and related technology developments. Thus, the sample size is $n = 1896$ and the count is $X = 910$. The sample proportion is

$$\hat{p} = \frac{X}{n} = \frac{910}{1896} = 0.47996$$

The standard error is

$$\text{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.47996(1-0.47996)}{1896}} = 0.011474$$

The z critical value for 95% confidence is $z^* = 1.96$, so the margin of error is

$$m = 1.96\text{SE}_{\hat{p}} = (1.96)\ (0.011474) = 0.022489$$

The confidence interval is

$$\hat{p} \pm m = 0.480 \pm 0.022$$

We are 95% confident that between 45.8% and 50.2% of CEOs would report that they think net jobs will decrease by 2025 because of robots and related technology developments.

Afin d'améliorer l'approximation dans les cas où le nombre de succès et d'échecs est trop faible (alors que n est d'au moins 10), il est possible de faire comme s'il y avait 4 observations de plus, soit 2 succès et 2 échecs de plus, et ainsi d'utiliser $\tilde{p} = (X + 2)/(n + 4)$ pour estimer la proportion de succès. Pour calculer l'intervalle de confiance, on utilise $SE_{\tilde{p}} = \sqrt{(\tilde{p}(1 - \tilde{p}))/(n + 4)}$, c'est-à-dire qu'on fait ce qu'on faisait avant mais on utilise $X + 2$ succès parmi $n + 4$ essais.

L'intervalle de confiance de niveau *approximatif* C pour la vraie proportion p est

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}},$$

où z^* est le point critique supérieur de niveau $(1 - C)/2$ d'une normale standard.

A nutrition researcher planning some bone health experiments would like to include some equol producers and some nonproducers among her subjects. A preliminary sample of 12 female subjects were measured, and four were found to be equol producers. We would like to estimate the proportion of equol producers in the population from which this researcher will draw her subjects.

The plus four estimate of the proportion of equol producers is

$$\tilde{p} = \frac{4+2}{12+4} = \frac{6}{16} = 0.375$$

For a 95% confidence interval, we use Table D to find $z^* = 1.96$. We first compute the standard error

$$\begin{aligned} \text{SE}_{\tilde{p}} &= \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} \\ &= \sqrt{\frac{(0.375)(1-0.375)}{16}} \\ &= 0.12103 \end{aligned}$$

and then the margin of error

$$\begin{aligned} m &= z^* \text{SE}_{\tilde{p}} \\ &= (1.96)(0.12103) \\ &= 0.237 \end{aligned}$$

So the confidence interval is

$$\begin{aligned} \tilde{p} \pm m &= 0.375 \pm 0.237 \\ &= (0.138, 0.612) \end{aligned}$$

We estimate with 95% confidence that between 14% and 61% of women from this population are equol producers. Note that the interval is very wide because the sample size is very small. Compare this result with the large-sample confidence interval.

z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

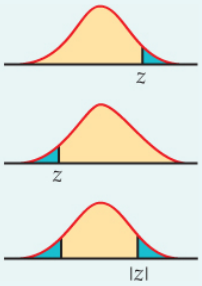
LARGE-SAMPLE SIGNIFICANCE TEST FOR A POPULATION PROPORTION

Draw an SRS of size n from a large population with an unknown proportion p of successes. To test the hypothesis H_0: $p = p_0$, compute the ***z* statistic**

$$z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

In terms of a standard Normal random variable Z, the approximate P-value for a test of H_0 against

H_a: $p > p_0$	is	$P(Z \geq z)$
H_a: $p < p_0$	is	$P(Z \leq z)$
H_a: $p \neq p_0$	is	$2P(Z \geq \lvert z \rvert)$



We recommend the large-sample z significance test as long as the expected number of successes, np_0, and the expected number of failures, $n(1 - p_0)$, are both at least 10.

Comparing two sunblock lotions. Your company produces a sunblock lotion designed to protect the skin from both UVA and UVB exposure to the sun. You hire a company to compare your product with the product sold by your major competitor. The testing company exposes skin on the backs of a sample of 20 people to UVA and UVB rays and measures the protection provided by each product. For 13 of the subjects, your product provided better protection, while for the other 7 subjects, your competitor's product provided better protection. Do you have evidence to support a commercial claiming that your product provides superior UVA and UVB protection? For the data we have $n = 20$ subjects and $X = 13$ successes. The parameter p is the proportion of people who would receive superior UVA and UVB protection from your product. To answer the claim question, we test

$$H_0: p = 0.5$$

$$H_a: p \neq 0.5$$

The expected numbers of successes (your product provides better protection) and failures (your competitor's product provides better protection) are $20 \times 0.5 = 10$ and $20 \times 0.5 = 10$. Both are at least 10, so we can use the z test. The sample proportion is

$$\hat{p} = \frac{X}{n} = \frac{13}{20} = 0.65$$

The test statistic is

$$z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.65 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{20}}} = 1.34$$

From Table A, we find $P(Z < 1.34) = 0.9099$, so the probability in the upper tail is $1 - 0.9099 = 0.0901$. The P-value is the area in both tails, $P = 2 \times 0.0901 = 0.1802$.

We conclude that the sunblock testing data are compatible with the hypothesis of no difference between your product and your competitor's product ($\hat{p} = 0.65$, $z = 1.34$, $P = 0.18$). The data do not support your proposed advertising claim.

P-VALUE

The probability, assuming H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the ***P*-value** of the test. The smaller the P-value, the stronger the evidence against H_0 provided by the data.

Recall that the margin of error for the large-sample confidence interval for a population proportion is

$$m = z^* \mathrm{SE}_{\hat{p}} = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The level C confidence interval for a proportion p will have a margin of error approximately equal to a specified value m when the sample size satisfies

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1 - p^*)$$

Here, z^* is the critical value for confidence level C, and p^* is a guessed value for the proportion of successes in the future sample.

The margin of error will be less than or equal to m if p^* is chosen to be 0.5. Substituting $p^* = 0.5$ into the formula above gives

$$n = \frac{1}{4}\left(\frac{z^*}{m}\right)^2$$

Planning a survey of students. A large university is interested in assessing student satisfaction with the overall campus environment. The plan is to distribute a questionnaire to an SRS of students, but before proceeding, the university wants to determine how many students to sample. The questionnaire asks about a student's degree of satisfaction with various student services, each measured on a five-point scale. The university is interested in the proportion p of students who are satisfied (that is, who choose either "satisfied" or "very satisfied," the two highest levels on the five-point scale).

The university wants to estimate p with 95% confidence and a margin of error less than or equal to 3%, or 0.03. For planning purposes, it is willing to use $p^* = 0.5$. To find the sample size required,

$$n = \frac{1}{4}\left(\frac{z^*}{m}\right)^2 = \frac{1}{4}\left(\frac{1.96}{0.03}\right)^2 = 1067.1$$

Round up to get $n = 1068$. (Always round up. Rounding down would give a margin of error slightly greater than 0.03.)

Similarly, for a 2.5% margin of error, we have (after rounding up)

$$n = \frac{1}{4}\left(\frac{1.96}{0.025}\right)^2 = 1537$$

and for a 2% margin of error,

$$n = \frac{1}{4}\left(\frac{1.96}{0.02}\right)^2 = 2401$$

Margins of error. In the Rec Sports survey, the margin of error of a 95% confidence interval for any value of $\hat{p}$ and $n = 200$ is

$$\begin{aligned} m &= z^* \mathrm{SE}_{\hat{p}} \\ &= 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{200}} \\ &= 0.139\sqrt{\hat{p}\,(1-\hat{p})} \end{aligned}$$

The results for various values of $\hat{p}$ are

$\hat{p}$	m
0.05	0.030
0.10	0.042
0.20	0.056
0.30	0.064
0.40	0.068
0.50	0.070
0.60	0.068
0.70	0.064
0.80	0.056
0.90	0.042
0.95	0.030

Rec Sports judged these margins of error to be acceptable, and it contacted 225 students, hoping to achieve a sample size of 200 for its survey.

Choosing a sample size for a significance test

power, p. 391

In Chapter 6, we also introduced the idea of power for a significance test. These ideas apply to the significance test for a proportion that we studied in this section. There are some more complicated details, but the basic ideas are the same. Fortunately, software can take care of the details, and we can concentrate on the input and output.

To find the required sample size, we need to specify

- The value of p_0 in the null hypothesis H_0: $p = p_0$.
- The alternative hypothesis, two-sided (H_a: $p \neq p_0$), one-sided (H_a: $p > p_0$ or H_a: $p < p_0$).

- A value of p for the alternative hypothesis.
- The Type I error (α, the probability of rejecting the null hypothesis when it is true); usually we choose 5% ($\alpha = 0.05$) for the Type I error.
- Power (probability of rejecting the null hypothesis when it is false); usually we choose 80% (0.80) for power.

SECTION 8.1 SUMMARY

- Inference about a population proportion p from an SRS of size n is based on the **sample proportion** $\hat{p} = X/n$. When n is large, $\hat{p}$ has approximately the Normal distribution with mean p and standard deviation $\sqrt{p(1-p)/n}$.
- For large samples, the **margin of error for confidence level *C*** is

$$m = z^* \mathrm{SE}_{\hat{p}}$$

where the critical value z^* is the value for the standard Normal density curve with area C between $-z^*$ and z^*, and the **standard error of $\hat{p}$** is

$$\mathrm{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- The **level *C* large-sample confidence interval** is

$$\hat{p} \pm m$$

We recommend using this interval for 90%, 95%, and 99% confidence whenever the number of successes and the number of failures are both at least 10. When sample sizes are smaller, alternative procedures such as the **plus four estimate of the population proportion** are recommended.

- The **sample size** required to obtain a confidence interval of approximate margin of error m for a proportion is found from

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*)$$

where p^* is a guessed value for the proportion and z^* is the standard Normal critical value for the desired level of confidence. To ensure that the margin of error of the interval is less than or equal to m no matter what $\hat{p}$ may be, use

$$n = \frac{1}{4}\left(\frac{z^*}{m}\right)^2$$

- Tests of H_0: $p = p_0$ are based on the ***z* statistic**

$$z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

with P-values calculated from the $N(0, 1)$ distribution. Use this procedure when the expected number of successes, np_0, and the expected number of failures, $n(1 - p_0)$, are both greater than 10.

- Software can be used to determine the sample sizes for significance tests.

8.2 Comparing Two Proportions

We call the two groups being compared Population 1 and Population 2 and the two population proportions of "successes" p_1 and p_2. The data consist of two independent SRSs, of size n_1 from Population 1 and size n_2 from Population 2. The proportion of successes in each sample estimates the corresponding population proportion. Here is the notation we will use in this section:

Population	Population proportion	Sample size	Count of successes	Sample proportion
1	p_1	n_1	X_1	$\hat{p}_1 = X_1/n_1$
2	p_2	n_2	X_2	$\hat{p}_2 = X_2/n_2$

To compare the two populations, we use the difference between the two sample proportions:

$$D = \hat{p}_1 - \hat{p}_2$$

When both sample sizes are sufficiently large, the sampling distribution of the difference D is approximately Normal.

Inference procedures for comparing proportions are z procedures based on the Normal approximation and on standardizing the difference D. The first step is to obtain the mean and standard deviation of D. By the addition rule for means, the mean of D is the difference of the means:

$$\mu_D = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p_1 - p_2$$

addition rule for means, p. 254
addition rule for variances, p. 258

That is, the difference $D = \hat{p}_1 - \hat{p}_2$ between the sample proportions is an unbiased estimator of the population difference $p_1 - p_2$. Similarly, the addition rule for variances tells us that the variance of D is the *sum* of the variances:

$$\begin{aligned}\sigma_D^2 &= \sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_2}^2 \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\end{aligned}$$

Therefore, when n_1 and n_2 are large, D is approximately Normal with mean $\mu_D = p_1 - p_2$ and standard deviation

$$\sigma_D = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

LARGE-SAMPLE CONFIDENCE INTERVAL FOR COMPARING TWO PROPORTIONS

Choose an SRS of size n_1 from a large population having proportion p_1 of successes and an independent SRS of size n_2 from another population having proportion p_2 of successes. The estimate of the difference in the population proportions is

$$D = \hat{p}_1 - \hat{p}_2$$

The **standard error of *D*** is

$$\text{SE}_D = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

and the **margin of error** for confidence level C is

$$m = z^* \text{SE}_D$$

where the critical value z^* is the value for the standard Normal density curve with area C between $-z^*$ and z^*. An **approximate level *C* confidence interval** for $p_1 - p_2$ is

$$D \pm m$$

Use this method for 90%, 95%, or 99% confidence when the number of successes and the number of failures in each sample are both 10 or more.

Who uses Instagram? A recent study compared the proportions of young women and men who use Instagram.[15] A total of 1069 young women and men were surveyed. These are the cases for the study. The response variable is User with values Yes and No. The explanatory variable is Sex with values "Men" and "Women." Here are the data:

Sex	n	X	$\hat{p} = X/n$
Women	537	328	0.6108
Men	532	234	0.4398
Total	1069	562	0.5257

In this table, the $\hat{p}$ column gives the sample proportions of women and men who use Instagram. The proportion for the total sample is given in the last entry in this column.

Let's find a 95% confidence interval for the difference between the proportions of women and of men who use Instagram. We first find the difference in the proportions:

$$\begin{aligned} D &= \hat{p}_1 - \hat{p}_2 \\ &= 0.6108 - 0.4398 \\ &= 0.1710 \end{aligned}$$

Then we calculate the standard error of D:

$$\begin{aligned} \text{SE}_D &= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= \sqrt{\frac{(0.6108)(1-0.6108)}{537} + \frac{(0.4398)(1-0.4398)}{532}} \\ &= 0.0301 \end{aligned}$$

For 95% confidence, we have $z^* = 1.96$, so the margin of error is

$$\begin{aligned} m &= z^* \text{SE}_D \\ &= (1.96)(0.0301) \\ &= 0.0590 \end{aligned}$$

The 95% confidence interval is

$$\begin{aligned} D \pm m &= 0.1710 \pm 0.0590 \\ &= (0.112, 0.230) \end{aligned}$$

With 95% confidence, we can say that the difference in the proportions is between 0.112 and 0.230. Alternatively, we can report that the difference between the percent of women who are Instagram users and the percent of men who are Instagram users is 17.1%, with a 95% margin of error of 5.9%.

In this example, men and women were not sampled separately. The sample sizes are, in fact, random and reflect the gender distributions of the subjects who responded to the survey. Two-sample significance tests and confidence intervals are still approximately correct in this situation.

In the preceding example, we chose women to be the first population. Had we chosen men to be the first population, the estimate of the difference would be negative (−0.1710). Because it is easier to discuss positive numbers, we generally choose the first population to be the one with the higher proportion.

EXEMPLE:

A pilot study included 12 girls and 12 boys from a population that will be used for a large experiment. Four of the boys and three of the girls had Tanner scores of 4 or 5, a high level of sexual maturity. Let's find a 95% confidence interval for the difference between the proportions of boys and girls who have high (4 or 5) Tanner scores in this population. The numbers of successes and failures in both groups are not all at least 10, so the large-sample approach is not recommended. On the other hand, the sample sizes are both at least 5, so the plus four method is appropriate.

The plus four estimate of the population proportion for boys is

$$\tilde{p}_1 = \frac{X_1+1}{n_1+2} = \frac{4+1}{12+2} = 0.3571$$

For girls, the estimate is

$$\tilde{p}_2 = \frac{X_2+1}{n_2+2} = \frac{3+1}{12+2} = 0.2857$$

Therefore, the estimate of the difference is

$$\tilde{D} = \tilde{p}_1 - \tilde{p}_2 = 0.3571 - 0.2857 = 0.071$$

The standard error of $\tilde{D}$ is

$$\begin{aligned} \mathrm{SE}_{\tilde{D}} &= \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}} \\ &= \sqrt{\frac{(0.3571)(1-0.3571)}{12+2} + \frac{(0.2857)(1-0.2857)}{12+2}} \\ &= 0.1760 \end{aligned}$$

For 95% confidence, z^* = 1.96 and the margin of error is

$$m = z^* \mathrm{SE}_{\tilde{D}} = (1.96)\ (0.1760) = 0.345$$

The confidence interval is

$$\begin{aligned} \tilde{D} \pm m &= 0.071 \pm 0.345 \\ &= (-0.274, 0.416) \end{aligned}$$

With 95% confidence, we can say that the difference in the proportions is between −0.274 and 0.416. Alternatively, we can report that the difference in the proportions of boys and girls with high Tanner scores in this population is 7.1% with a 95% margin of error of 34.5%.

Significance test for a difference in proportions

Although we prefer to compare two proportions by giving a confidence interval for the difference between the two population proportions, it is sometimes useful to test the null hypothesis that the two population proportions are the same.

We standardize $D = \hat{p}_1 - \hat{p}_2$ by subtracting its mean $p_1 - p_2$ and then dividing by its standard deviation

$$\sigma_D = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

If n_1 and n_2 are large, the standardized difference is approximately $N(0, 1)$. For the large-sample confidence interval we used sample estimates in place of the unknown population values in the expression for σ_D. Although this approach would lead to a valid significance test, we instead adopt the more common practice of replacing the unknown

σ_D with an estimate that takes into account our null hypothesis H_0: $p_1 = p_2$. If these two proportions are equal, then we can view all the data as coming from a single population. Let p denote the common value of p_1 and p_2; then the standard deviation of $D = \hat{p}_1 - \hat{p}_2$ is

$$\begin{aligned} \sigma_D &= \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} \\ &= \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \end{aligned}$$

We estimate the common value of p by the overall proportion of successes in the two samples:

$$\hat{p} = \frac{\text{number of successes in both samples}}{\text{number of observations in both samples}} = \frac{X_1 + X_2}{n_1 + n_2}$$

pooled estimate of *p*

This estimate of p is called the **pooled estimate** because it combines, or pools, the information from both samples.

To estimate σ_D under the null hypothesis, we substitute $\hat{p}$ for p in the expression for σ_D. The result is a standard error for D that assumes $H_0: p_1 = p_2$:

$$SE_{Dp} = \sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

The subscript on SE_{Dp} reminds us that we pooled data from the two samples to construct the estimate.

SIGNIFICANCE TEST FOR COMPARING TWO PROPORTIONS

To test the hypothesis

$$H_0: p_1 = p_2$$

compute the ***z* statistic**

$$z = \frac{p_1 - p_2}{SE_{D_p}}$$

where the **pooled standard error** is

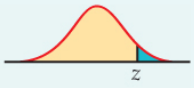
$$SE_{D_p} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

and where the **pooled estimate** of the common value of p_1 and p_2 is

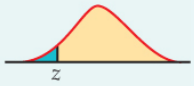
$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

In terms of a standard Normal random variable Z, the approximate P-value for a test of H_0 against

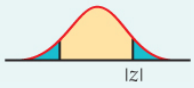
$H_a: p_1 > p_2$ is $P(Z \geq z)$



$H_a: p_1 < p_2$ is $P(Z \leq z)$



$H_a: p_1 \neq p_2$ is $2P(Z \geq |z|)$



This z test is based on the Normal approximation to the binomial distribution. As a general rule, we will use it when the number of successes and the number of failures in each of the samples are at least 5.

Sex and Instagram use: The *z* test. Are young women and men equally likely to say they use Instagram? We examine the data in Example 8.11 (page 507) to answer this question. Here is the data summary:

Sex	*n*	*X*	$\hat{p} = X/n$
Women	537	328	0.6108
Men	532	234	0.4398
Total	1069	562	0.5257

The sample proportions are certainly quite different, but we will perform a significance test to see if the difference is large enough to lead us to believe that the population proportions are not equal. Formally, we test the hypotheses

$$H_0: p_1 = p_2$$

$$H_a: p_1 \neq p_2$$

The pooled estimate of the common value of p is

$$\hat{p} = \frac{328 + 234}{537 + 532} = \frac{562}{1069} = 0.5257$$

Note that this is the estimate on the bottom line of the preceding data summary. The test statistic is calculated as follows:

$$SE_{Dp} = \sqrt{(0.5257)\,(1 - 0.5257)\left(\frac{1}{537} + \frac{1}{532}\right)} = 0.03055$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{Dp}} = \frac{0.6108 - 0.4398}{0.03055} = 5.60$$

The P-value is $2P(Z \geq 5.60)$. Note that the largest value for z in Table A is 3.49. Therefore, from Table A, we can conclude that $P < 2(1 - 0.9998) = 0.0004$, although we know that the true P value is smaller.

Here is our summary: 61% of the women and 44% of the men are Instagram users; the difference is statistically significant ($z = 5.60$, $P < 0.0004$).

SAMPLE SIZE FOR DESIRED MARGIN OF ERROR

The level C confidence interval for a difference in two proportions will have a margin of error approximately equal to a specified value m when the sample size for each of the two proportions is

$$n = \left(\frac{z^*}{m}\right)^2 (p_1^*(1-p_1^*) + p_2^*(1-p_2^*))$$

Here, z^* is the critical value for confidence C, and p_1^* and p_2^* are guessed values for p_1 and p_2, the proportions of successes in the future sample.

The margin of error will be less than or equal to m if p_1^* and p_2^* are chosen to be 0.5. The common sample size required is then given by

$$n = \left(\frac{1}{2}\right)\left(\frac{z^*}{m}\right)^2$$

Note that to use the confidence interval, which is based on the Normal approximation, we still require that the number of successes and the number of failures in each of the samples are at least 10.

Confidence interval–based sample sizes for preferences of women and men. Consider the setting in Exercise 8.50, where we compared the preferences of women and men for two commercials. Suppose we want to do a study in which we perform a similar comparison using a 95% confidence interval that will have a margin of error of 0.1 or less. What should we choose for our sample size? Using $m = 0.1$ and z^* in our formula, we have

$$n = \left(\frac{1}{2}\right)\left(\frac{z^*}{m}\right)^2 = \left(\frac{1}{2}\right)\left(\frac{1.96}{0.1}\right)^2 = 192.08$$

We would include 192 women and 192 men in our study.

Note that we have rounded the calculated value, 192.08, down because it is very close to 192. The normal procedure would be to round the calculated value up to the next larger integer.

Aspirin and blood clots: Relative risk. A study of patients who had blood clots (venous thromboembolism) and had completed the standard treatment were randomly assigned to receive a low-dose aspirin or a placebo treatment. The 822 patients in the study were randomized to the treatments, 411 to each. Patients were monitored for several years for the occurrence of several related medical conditions. Counts of patients who experienced one or more of these conditions were reported for each year after the study began.[18] The following table gives the data for a composite of events, termed "major vascular events." Here, X is the number of patients who had a major event.

Population	n	X	$\hat{p} = X/n$
1 (aspirin)	411	45	0.1095
2 (placebo)	411	73	0.1776
Total	822	118	0.1436

The relative risk is

$$\text{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{45/411}{73/411} = 0.6164$$

Software gives the 95% confidence interval as 0.4364 to 0.8707. Taking aspirin has reduced the occurrence of major events to 62% of what it is for patients taking the placebo. The 95% confidence interval is 44% to 87%.

SECTION 8.2 SUMMARY

- The **large-sample estimate of the difference in two population proportions** is

$$D = \hat{p}_1 - \hat{p}_2$$

where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions:

$$\hat{p}_1 = \frac{X_1}{n_1} \text{ and } \hat{p}_2 = \frac{X_2}{n_2}$$

- The **standard error of the difference *D*** is

$$\text{SE}_D = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- The **margin of error for confidence level *C*** is

$$m = z^* \text{SE}_D$$

where z^* is the value for the standard Normal density curve with area C between $-z^*$ and z^*. The **large-sample level *C* confidence interval** is

$$D \pm m$$

We recommend using this interval for 90%, 95%, or 99% confidence when the number of successes and the number of failures in both samples are all at least 10. When sample sizes are smaller, alternative procedures such as the **plus four estimate of the difference in two population proportions** are recommended.

- Significance tests of H_0: $p_1 = p_2$ use the ***z* statistic**

$$z = \frac{p_1 - p_2}{\text{SE}_{Dp}}$$

with P-values from the $N(0, 1)$ distribution. In this statistic,

$$\text{SE}_{Dp} = \sqrt{\hat{p}\,(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

and $\hat{p}$ is the **pooled estimate** of the common value of p_1 and p_2:

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Use this test when the number of successes and the number of failures in each of the samples are at least 5.

- **Relative risk** is the ratio of two sample proportions:

$$\text{RR} = \frac{p_1}{p_2}$$

Confidence intervals for relative risk are often used to summarize the comparison of two proportions.