Do men take more risks than women? Psychologists have documented that in many situations, men choose riskier behavior than women do. But what is the effect of having a woman by their side? A recent seatbelt observation study in Massachusetts\(^1\) found that, not surprisingly, male drivers wear seatbelts less often than women do. The study also noted that men’s belt-wearing jumped more than 16 percentage points when they had a female passenger. Seatbelt use was recorded at 161 locations in Massachusetts, using random-sampling methods developed by the National Highway Traffic Safety Administration (NHTSA). Female drivers wore belts more than 70% of the time, regardless of the sex of their passengers. Of 4208 male drivers with female passengers, 2777 (66.0%) were belted. But among 2763 male drivers with male passengers only, 1363 (49.3%) wore seatbelts. This was only a random sample, but it suggests there may be a shift in men’s risk-taking behavior when women are present. What would we estimate the true size of that gap to be?

Comparisons between two percentages are much more common than questions about isolated percentages. And they are more interesting. We often want to know how two groups differ, whether a treatment is better than a placebo control, or whether this year’s results are better than last year’s.

Another Ruler

We know the difference between the proportions of men wearing seatbelts seen in the sample. It’s 16.7%. But what’s the true difference for all men? We know that our estimate probably isn’t exactly right. To say more, we need a new ruler—the standard deviation of the sampling distribution model for the difference in the proportions. Now we have two proportions, and each will vary from sample to sample. We are interested in the difference between them. So what is the correct standard deviation?

\(^1\)Massachusetts Traffic Safety Research Program [June 2007].
The Standard Deviation of the Difference Between Two Proportions

For independent random variables, variances add.

The answer comes to us from Chapter 16. Remember the Pythagorean Theorem of Statistics?

\[ \text{The variance of the sum or difference of two independent random variables is the sum of their variances.} \]

This is such an important (and powerful) idea in Statistics that it’s worth pausing a moment to review the reasoning. Here’s some intuition about why variation increases even when we subtract two random quantities.

Grab a full box of cereal. The box claims to contain 16 ounces of cereal. We know that’s not exact: There’s some small variation from box to box. Now pour a bowl of cereal. Of course, your 2-ounce serving will not be exactly 2 ounces. There’ll be some variation there, too. How much cereal would you guess was left in the box? Do you think your guess will be as close as your guess for the full box?

After you pour your bowl, the amount of cereal in the box is still a random quantity (with a smaller mean than before), but it is even more variable because of the additional variation in the amount you poured.

According to our rule, the variance of the amount of cereal left in the box would now be the sum of the two variances.

We want a standard deviation, not a variance, but that’s just a square root away. We can write symbolically what we’ve just said:

\[ \text{Be careful, though—this simple formula applies only when } X \text{ and } Y \text{ are independent. Just as the Pythagorean Theorem works only for right triangles, our formula works only for independent random variables. Always check for independence before using it.} \]

The Standard Deviation of the Difference Between Two Proportions

Fortunately, proportions observed in independent random samples are independent, so we can put the two proportions in for X and Y and add their variances. We just need to use careful notation to keep things straight.

When we have two samples, each can have a different size and proportion value, so we keep them straight with subscripts. Often we choose subscripts that remind us of the groups. For our example, we might use “M” and “F”, but generally we’ll just use “1” and “2”. We will represent the two sample proportions as \( \hat{p}_1 \) and \( \hat{p}_2 \), and the two sample sizes as \( n_1 \) and \( n_2 \).

The standard deviations of the sample proportions are \( SD(\hat{p}_1) = \sqrt{\frac{p_1 q_1}{n_1}} \) and \( SD(\hat{p}_2) = \sqrt{\frac{p_2 q_2}{n_2}} \), so the variance of the difference in the proportions is

\[ \text{Var}(\hat{p}_1 - \hat{p}_2) = \left( \frac{p_1 q_1}{n_1} \right)^2 + \left( \frac{p_2 q_2}{n_2} \right)^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \]

The standard deviation is the square root of that variance:

\[ SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \]

---

\(^2\) If you don’t remember the formula, don’t rely on the Scarecrow’s version from The Wizard of Oz. He may have a brain and have been awarded his Th.D. (Doctor of Thinkology), but he gets the formula wrong.
We usually don’t know the true values of $p_1$ and $p_2$. When we have the sample proportions in hand from the data, we use them to estimate the variances. So the standard error is

$$SE(p_1 - p_2) = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

### FOR EXAMPLE

**Finding the standard error of a difference in proportions**

A recent survey of 886 randomly selected teenagers (aged 12–17) found that more than half of them had online profiles. Some researchers and privacy advocates are concerned about the possible access to personal information about teens in public places on the Internet. There appear to be differences between boys and girls in their online behavior. Among teens aged 15–17, 57% of the 248 boys had posted profiles, compared to 70% of the 256 girls. Let’s start the process of estimating how large the true gender gap might be.

**Question:** What’s the standard error of the difference in sample proportions?

Because the boys and girls were selected at random, it’s reasonable to assume their behaviors are independent, so it’s okay to use the Pythagorean Theorem of Statistics and add the variances:

$$SE(\hat{p}_{\text{girls}}) = \sqrt{\frac{0.57 \times 0.43}{248}} = 0.0314$$

$$SE(\hat{p}_{\text{boys}}) = \sqrt{\frac{0.70 \times 0.30}{256}} = 0.0286$$

$$SE(\hat{p}_{\text{girls}} - \hat{p}_{\text{boys}}) = \sqrt{0.0314^2 + 0.0286^2} = 0.0425$$

### Assumptions and Conditions

Before we look at our example, we need to check assumptions and conditions.

#### INDEPENDENCE ASSUMPTIONS

**Independence Assumption:** Within each group, the data should be based on results for independent individuals. We can’t check that for certain, but we can check the following:

- **Randomization Condition:** The data in each group should be drawn independently and at random from a homogeneous population or generated by a randomized comparative experiment.

- **The 10% Condition:** If the data are sampled without replacement, the sample should not exceed 10% of the population.

Because we are comparing two groups in this way, we need an additional Independence Assumption. In fact, this is the most important of these assumptions. If it is violated, these methods just won’t work.

**Independent Groups Assumption:** The two groups we’re comparing must also be independent of each other. Usually, the independence of the groups from each other is evident from the way the data were collected.

Why is the Independent Groups Assumption so important? If we compare husbands with their wives, or a group of subjects before and after some treatment, we can’t just add the variances. Subjects’ performance before a treatment might very well be related to their performance after the treatment. So the proportions are not independent and the Pythagorean-style variance formula does not hold. We’ll see a way to compare a common kind of nonindependent samples in a later chapter.

---

Sample Size Condition

Each of the groups must be big enough. As with individual proportions, we need larger groups to estimate proportions that are near 0% or 100%. We usually check the Success/Failure Condition for each group.

Success/Failure Condition: Both groups are big enough that at least 10 successes and at least 10 failures have been observed in each.

For Example: Checking assumptions and conditions

Recap: Among randomly sampled teens aged 15–17, 57% of the 248 boys had posted online profiles, compared to 70% of the 256 girls.

Question: Can we use these results to make inferences about all 15–17-year-olds?

✓ Randomization Condition: The sample of boys and the sample of girls were both chosen randomly.
✓ 10% Condition: 248 boys and 256 girls are each less than 10% of all teenage boys and girls.
✓ Independent Groups Assumption: Because the samples were selected at random, it’s reasonable to believe the boys’ online behaviors are independent of the girls’ online behaviors.
✓ Success/Failure Condition: Among the boys, 248(0.57) = 141 had online profiles and the other 248(0.43) = 107 did not. For the girls, 256(0.70) = 179 successes and 256(0.30) = 77 failures. All counts are at least 10.

Because all the assumptions and conditions are satisfied, it’s okay to proceed with inference for the difference in proportions.

(Note that when we find the observed counts of successes and failures, we round off to whole numbers. We’re using the reported percentages to recover the actual counts.)

The Sampling Distribution

We’re almost there. We just need one more fact about proportions. We already know that for large enough samples, each of our proportions has an approximately Normal sampling distribution. The same is true of their difference.

Why Normal?
In Chapter 16 we learned that sums and differences of independent Normal random variables also follow a Normal model. That’s the reason we use a Normal model for the difference of two independent proportions.

The Sampling Distribution Model for a Difference Between Two Independent Proportions

Provided that the sampled values are independent, the sample are independent, and the sample sizes are large enough, the sampling distribution of \( \hat{p}_1 - \hat{p}_2 \) is modeled by a Normal model with mean \( \mu = p_1 - p_2 \) and standard deviation

\[
SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}.
\]

The sampling distribution model and the standard deviation give us all we need to find a margin of error for the difference in proportions—or at least they would if we knew the true proportions, \( p_1 \) and \( p_2 \). However, we don’t know the true values, so we’ll work with the observed proportions, \( \hat{p}_1 \) and \( \hat{p}_2 \), and use \( SE(\hat{p}_1 - \hat{p}_2) \) to estimate the standard deviation. The rest is just like a one-proportion z-interval.
A TWO-PROPORTION z-INTERVAL
When the conditions are met, we are ready to find the confidence interval for the difference of two proportions, \( p_1 - p_2 \). The confidence interval is

\[
(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE(\hat{p}_1 - \hat{p}_2)
\]

where we find the standard error of the difference,

\[
SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}},
\]

from the observed proportions.

The critical value \( z^* \) depends on the particular confidence level, \( C \), that we specify.

FOR EXAMPLE
Finding a two-proportion z-interval

Recap: Among randomly sampled teens aged 15–17, 57% of the 248 boys had posted online profiles, compared to 70% of the 256 girls. We calculated the standard error for the difference in sample proportions to be \( SE(\hat{p}_{\text{girls}} - \hat{p}_{\text{boys}}) = 0.0425 \) and found that the assumptions and conditions required for inference checked out okay.

Question: What does a confidence interval say about the difference in online behavior?

A 95% confidence interval for \( p_{\text{girls}} - p_{\text{boys}} \) is \( (0.70 - 0.57) \pm 1.96(0.0425) \)

\[
0.13 \pm 0.083
\]

(4.7%, 21.3%)

We can be 95% confident that among teens aged 15–17, the proportion of girls who post online profiles is between 4.7 and 21.3 percentage points higher than the proportion of boys who do. It seems clear that teen girls are more likely to post profiles than are boys the same age.

STEP-BY-STEP EXAMPLE
A Two-Proportion z-Interval

Now we are ready to be more precise about the passenger-based gap in male drivers’ seatbelt use. We’ll estimate the difference with a confidence interval using a method called the two-proportion z-interval and follow the four confidence interval steps.

Question: How much difference is there in the proportion of male drivers who wear seatbelts when sitting next to a male passenger and the proportion who wear seatbelts when sitting next to a female passenger?

Plan State what you want to know. Discuss the variables and the W’s.

Identify the parameter you wish to estimate. (It usually doesn’t matter in which direction we subtract, so, for convenience, we usually choose the direction with a positive difference.)
Choose and state a confidence level.

Model Think about the assumptions and check the conditions.

The Success/Failure Condition must hold for each group.

State the sampling distribution model for the statistic.

Choose your method.

Mechanics Construct the confidence interval.

As often happens, the key step in finding the confidence interval is estimating the standard deviation of the sampling distribution model of the statistic. Here the statistic is the difference in the proportions of men who wear seatbelts when they have a female passenger and the proportion who do so with a male passenger. Substitute the data values into the formula.

The sampling distribution is Normal, so the critical value for a 95% confidence interval, \( z^* \), is 1.96. The margin of error is the critical value times the SE.

I will find a 95% confidence interval for this parameter.

- **Independence Assumption:** Driver behavior was independent from car to car.
- **Randomization Condition:** The NHTSA methods are more complex than an SRS, but they result in a suitable random sample.
- **10% Condition:** The samples include far fewer than 10% of all male drivers accompanied by male or by female passengers.
- **Independent Groups Assumption:** There's no reason to believe that seatbelt use among drivers with male passengers and those with female passengers are not independent.
- **Success Failure Condition:** Among male drivers with female passengers, 2777 wore seatbelts and 1431 did not; of those driving with male passengers, 1363 wore seatbelts and 1400 did not. Each group contained far more than 10 successes and 10 failures.

Under these conditions, the sampling distribution of the difference between the sample proportions is approximately Normal, so I'll find a two-proportion \( z \)-interval.

I know

\[
\begin{align*}
n_F &= 4208, \ n_M &= 2763. \\
\hat{p}_F &= \frac{2777}{4208} = 0.660, \ \hat{p}_M &= \frac{1363}{2763} = 0.493 \\
\end{align*}
\]

I'll estimate the SD of the difference with

\[
SE(\hat{p}_F - \hat{p}_M) = \sqrt{\frac{\hat{p}_F(1-\hat{p}_F)}{n_F} + \frac{\hat{p}_M(1-\hat{p}_M)}{n_M}}
\]

\[
= \sqrt{\left(\frac{0.660(0.340)}{4208}\right) + \left(\frac{0.493(0.507)}{2763}\right)}
\]

\[
= 0.012
\]

\[
ME = z^* \times SE(\hat{p}_F - \hat{p}_M) \\
= 1.96(0.012) = 0.024
\]
The confidence interval is the statistic ± ME.

The observed difference in proportions is \( \hat{p}_F - \hat{p}_M = 0.660 - 0.493 = 0.167 \), so the 95% confidence interval is

\[
0.167 \pm 0.024
\]

or 14.3% to 19.1%.

**Conclusion** Interpret your confidence interval in the proper context. (Remember: We’re 95% confident that our interval captured the true difference.)

I am 95% confident that the proportion of male drivers who wear seatbelts when driving next to a female passenger is between 14.3 and 19.1 percentage points higher than the proportion who wear seatbelts when driving next to a male passenger.

This is an interesting result—but be careful not to try to say too much! In Massachusetts, overall seatbelt use is lower than the national average, so we can’t be certain that these results generalize to other states. And these were two different groups of men, so we can’t say that, individually, men are more likely to buckle up when they have a woman passenger. You can probably think of several alternative explanations; we’ll suggest just a couple. Perhaps age is a lurking variable: Maybe older men are more likely to wear seatbelts and also more likely to be driving with their wives. Or maybe men who don’t wear seatbelts have trouble attracting women!

**Finding a confidence interval**

You can use a routine in the **STAT TESTS** menu to create confidence intervals for the difference of two proportions. Remember, the calculator can do only the mechanics—checking conditions and writing conclusions are still up to you.

A Gallup Poll asked whether the attribute “intelligent” described men in general. The poll revealed that 28% of 506 men thought it did, but only 14% of 520 women agreed. We want to estimate the true size of the gender gap by creating a 95% confidence interval.

- Go to the **STAT TESTS** menu. Scroll down the list and select **B:2-PropZInt**.
- Enter the observed number of males: \( 0.28 \times 506 \). Remember that the actual number of males must be a whole number, so be sure to round off.
- Enter the sample size: 506 males.
- Repeat those entries for women: \( 0.14 \times 520 \) agreed, and the sample size was 520.
- Specify the desired confidence level.
- **Calculate** the result.

And now explain what you see: We are 95% confident that the proportion of men who think the attribute “intelligent” describe males in general is between 9 and 19 percentage points higher than the proportion of women who think so.
The National Sleep Foundation asked a random sample of 1010 U.S. adults questions about their sleep habits. The sample was selected in the fall of 2001 from random telephone numbers, stratified by region and sex, guaranteeing that an equal number of men and women were interviewed (2002 Sleep in America Poll, National Sleep Foundation, Washington, DC).

One of the questions asked about snoring. Of the 995 respondents, 37% of adults reported that they snored at least a few nights a week during the past year. Would you expect that percentage to be the same for all age groups? Split into two age categories, 26% of the 184 people under 30 snored, compared with 39% of the 811 in the older group. Is this difference of 13% real, or due only to natural fluctuations in the sample we’ve chosen?

The question calls for a hypothesis test. Now the parameter of interest is the true difference between the (reported) snoring rates of the two age groups. What’s the appropriate null hypothesis? That’s easy here. We hypothesize that there is no difference in the proportions. This is such a natural null hypothesis that we rarely consider any other. But instead of writing $H_0: p_1 = p_2$, we usually express it in a slightly different way. To make it relate directly to the difference, we hypothesize that the difference in proportions is zero:

$$H_0: p_1 - p_2 = 0.$$

### JUST CHECKING

A public broadcasting station plans to launch a special appeal for additional contributions from current members. Unsure of the most effective way to contact people, they run an experiment. They randomly select two groups of current members. They send the same request for donations to everyone, but it goes to one group by e-mail and to the other group by regular mail. The station was successful in getting contributions from 26% of the members they e-mailed but only from 15% of those who received the request by regular mail. A 90% confidence interval estimated the difference in donation rates to be $11\% \pm 7\%$.

1. Interpret the confidence interval in this context.
2. Based on this confidence interval, what conclusion would we reach if we tested the hypothesis that there's no difference in the response rates to the two methods of fundraising? Explain.

Our hypothesis is about a new parameter: the difference in proportions. We’ll need a standard error for that. Wait—don’t we know that already? Yes and no. We know that the standard error of the difference in proportions is

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}},$$

and we could just plug in the numbers, but we can do even better. The secret is that proportions and their standard deviations are linked. There are two proportions in the standard error formula—but look at the null hypothesis. It says that these proportions are equal. To do a hypothesis test, we assume that the null hypothesis is true. So there should be just a single value of $\hat{p}$ in the SE formula (and, of course, $\hat{q}$ is just $1 - \hat{p}$).
How would we do this for the snoring example? If the null hypothesis is true, then, among all adults, the two groups have the same proportion. Overall, we saw $48 + 318 = 366$ snorers out of a total of $184 + 811 = 995$ adults who responded to this question. The overall proportion of snorers was $\frac{366}{995} = 0.3678$.

Combining the counts like this to get an overall proportion is called pooling. Whenever we have data from different sources or different groups but we believe that they really came from the same underlying population, we pool them to get better estimates.

When we have counts for each group, we can find the pooled proportion as

$$p_{\text{pooled}} = \frac{\text{Success}_1 + \text{Success}_2}{n_1 + n_2},$$

where $\text{Success}_1$ is the number of successes in group 1 and $\text{Success}_2$ is the number of successes in group 2. That’s the overall proportion of success.

When we have only proportions and not the counts, as in the snoring example, we have to reconstruct the number of successes by multiplying the sample sizes by the proportions:

$$\text{Success}_1 = n_1 \hat{p}_1 \quad \text{and} \quad \text{Success}_2 = n_2 \hat{p}_2.$$

If these calculations don’t come out to whole numbers, round them first. There must have been a whole number of successes, after all. (This is the only time you should round values in the middle of a calculation.)

We then put this pooled value into the formula, substituting it for both sample proportions in the standard error formula:

$$SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_{\text{pooled}} \hat{q}_{\text{pooled}}}{n_1} + \frac{\hat{p}_{\text{pooled}} \hat{q}_{\text{pooled}}}{n_2}} = \sqrt{\frac{0.3678 \times (1 - 0.3678)}{184} + \frac{0.3678 \times (1 - 0.3678)}{811}}.$$

This comes out to 0.039.

## Improving the Success/Failure Condition

The vaccine Gardasil® was introduced to prevent the strains of human papillomavirus (HPV) that are responsible for almost all cases of cervical cancer. In randomized placebo-controlled clinical trials, only 1 case of HPV was diagnosed among 7897 women who received the vaccine, compared with 91 cases diagnosed among 7899 who received a placebo. The one observed HPV case (“success”) doesn’t meet the at-least-10-successes criterion. Surely, though, we should not refuse to test the effectiveness of the vaccine just because it failed so rarely; that would be absurd.

For that reason, in a two-proportion $z$-test, the proper Success/Failure test uses the expected frequencies, which we can find from the pooled proportion. In this case,

$$\hat{p}_{\text{pooled}} = \frac{91 + 1}{7899 + 7897} = 0.0058,$$

$$n_1 \hat{p}_{\text{pooled}} = 7899(0.0058) = 46,$$

$$n_2 \hat{p}_{\text{pooled}} = 7897(0.0058) = 46,$$

so we can proceed with the hypothesis test.

---

Often it is easier just to check the observed numbers of successes and failures. If they are both greater than 10, you don’t need to look further. But keep in mind that the correct test uses the expected frequencies rather than the observed ones.

**Compared to What?**

Naturally, we’ll reject our null hypothesis if we see a large enough difference in the two proportions. How can we decide whether the difference we see, \( \hat{p}_1 - \hat{p}_2 \), is large? The answer is the same as always: We just compare it to its standard deviation.

Unlike previous hypothesis-testing situations, the null hypothesis doesn’t provide a standard deviation, so we’ll use a standard error (here, pooled). Since the sampling distribution is Normal, we can divide the observed difference by its standard error to get a \( z \)-score. The \( z \)-score will tell us how many standard errors the observed difference is away from 0. We can then use the 68–95–99.7 Rule to decide whether this is large, or some technology to get an exact P-value. The result is a two-proportion \( z \)-test.

**TWO-PROPORTION \( z \)-TEST**

The conditions for the two-proportion \( z \)-test are the same as for the two-proportion \( z \)-interval. We are testing the hypothesis

\[
H_0: p_1 - p_2 = 0.
\]

Because we hypothesize that the proportions are equal, we pool the groups to find

\[
\hat{p}_{\text{pooled}} = \frac{\text{Success}_1 + \text{Success}_2}{n_1 + n_2}
\]

and use that pooled value to estimate the standard error:

\[
SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_{\text{pooled}}(1 - \hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1 - \hat{p}_{\text{pooled}})}{n_2}}.
\]

Now we find the test statistic,

\[
z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2)}.
\]

When the conditions are met and the null hypothesis is true, this statistic follows the standard Normal model, so we can use that model to obtain a P-value.

**Activity: Test for a Difference Between Two Proportions.** Is premium-brand chicken less likely to be contaminated than store-brand chicken?

**Question:** Are the snoring rates of the two age groups really different?

**Step-by-Step Example: A Two-Proportion \( z \)-Test**

**Plan** State what you want to know. Discuss the variables and the W’s.

I want to know whether snoring rates differ for those under and over 30 years old. The data are from a random sample of 1010 U.S. adults surveyed in the 2002 Sleep in America Poll. Of these, 995 responded to the question about snoring, indicating whether or not they had snored at least a few nights a week in the past year.
CHAPTER 22  Comparing Two Proportions

Hypotheses  The study simply broke down the responses by age, so there is no sense that either alternative was preferred. A two-sided alternative hypothesis is appropriate.

Model  Think about the assumptions and check the conditions.

State the null model.

Choose your method.

Mechanics  The hypothesis is that the proportions are equal, so pool the sample data.

Use the pooled SE to estimate $SD(p_{old} - p_{young})$.

$$n_{young} = 184, y_{young} = 48, \hat{p}_{young} = 0.261$$
$$n_{old} = 811, y_{old} = 318, \hat{p}_{old} = 0.392$$

$$\hat{p}_{pooled} = \frac{y_{old} + y_{young}}{n_{old} + n_{young}} = \frac{318 + 48}{811 + 184} = 0.3678$$

$$SE_{pooled}(\hat{p}_{old} - \hat{p}_{young})$$

$$= \sqrt{\frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_{old}} + \frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_{young}}}$$

$$= \sqrt{\frac{(0.3678)(0.6322)}{811} + \frac{(0.3678)(0.6322)}{184}}$$

$$= 0.039375$$

The observed difference in sample proportions is

$$\hat{p}_{old} - \hat{p}_{young} = 0.392 - 0.261 = 0.131$$

$^5$This is one of those situations in which the traditional term “success” seems a bit weird. A success here could be that a person snores. “Success” and “failure” are arbitrary labels left over from studies of gambling games.
Compared to What?

Make a picture. Sketch a Normal model centered at the hypothesized difference of 0. Shade the region to the right of the observed difference, and because this is a two-tailed test, also shade the corresponding region in the other tail.

Find the z-score for the observed difference in proportions, 0.131.

Find the P-value using Table Z or technology. Because this is a two-tailed test, we must double the probability we find in the upper tail.

\[
\hat{p}_{\text{old}} - \hat{p}_{\text{young}} = 0.131
\]

\[
z = \frac{0.131 - 0}{0.039375} = 3.33
\]

The P-value of 0.0008 says that if there really were no difference in (reported) snoring rates between the two age groups, then the difference observed in this study would happen only 8 times in 10,000. This is so small that I reject the null hypothesis of no difference and conclude that there is a difference in the rate of snoring between older adults and younger adults. It appears that older adults are more likely to snore.

**Conclusion** Link the P-value to your decision about the null hypothesis, and state your conclusion in context.

**Testing the hypothesis**

Yes, of course, there’s a STAT TESTS routine to test a hypothesis about the difference of two proportions. Let’s do the mechanics for the test about snoring. Of 811 people over 30 years old, 318 snored, while only 48 of the 184 people under 30 did.

- In the STAT TESTS menu select 6:2-PropZTest.
- Enter the observed numbers of snorers and the sample sizes for both groups.
- Since this is a two-tailed test, indicate that you want to see if the proportions are unequal. When you choose this option, the calculator will automatically include both tails as it determines the P-value.
- Calculate the result. Don’t worry; for this procedure the calculator will pool the proportions automatically.

Now it is up to you to interpret the result and state a conclusion. We see a z-score of 3.33 and the P-value is 0.0008. Such a small P-value indicates that the observed difference is unlikely to be sampling error. What does that mean about snoring and age? Here’s a great opportunity to follow up with a confidence interval so you can Tell even more!
WHAT CAN GO WRONG?

- Don’t use two-sample proportion methods when the samples aren’t independent. These methods give wrong answers when this assumption of independence is violated. Good random sampling is usually the best insurance of independent groups. Make sure there is no relationship between the two groups. For example, you can’t compare the
proportion of respondents who own SUVs with the proportion of those same respondents who think the tax on gas should be eliminated. The responses are not independent because you’ve asked the same people. To use these methods to estimate or test the difference, you’d need to survey two different groups of people.

Alternatively, if you have a random sample, you can split your respondents according to their answers to one question and treat the two resulting groups as independent samples. So, you could test whether the proportion of SUV owners who favored eliminating the gas tax was the same as the corresponding proportion among non-SUV owners.

- Don’t apply inference methods where there was no randomization. If the data do not come from representative random samples or from a properly randomized experiment, then the inference about the differences in proportions will be wrong.
- Don’t interpret a significant difference in proportions causally. It turns out that people with higher incomes are more likely to snore. Does that mean money affects sleep patterns? Probably not. We have seen that older people are more likely to snore, and they are also likely to earn more. In a prospective or retrospective study, there is always the danger that other lurking variables not accounted for are the real reason for an observed difference. Be careful not to jump to conclusions about causality.

### CONNECTIONS

In Chapter 3 we looked at contingency tables for two categorical variables. Differences in proportions are just $2 \times 2$ contingency tables. You’ll often see data presented in this way. For example, the snoring data could be shown as

<table>
<thead>
<tr>
<th></th>
<th>18–29</th>
<th>30 and over</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snore</td>
<td>48</td>
<td>318</td>
<td>366</td>
</tr>
<tr>
<td>Don’t snore</td>
<td>136</td>
<td>493</td>
<td>629</td>
</tr>
<tr>
<td>Total</td>
<td>184</td>
<td>811</td>
<td>995</td>
</tr>
</tbody>
</table>

We tested whether the column percentages of snorers were the same for the two age groups.

This chapter gives the first examples we’ve seen of inference methods for a parameter other than a simple proportion. Although we have a different standard error, the step-by-step procedures are almost identical. In particular, once again we divide the statistic (the difference in proportions) by its standard error and get a $z$-score. You should feel right at home.

### WHAT HAVE WE LEARNED?

In the last few chapters we began our exploration of statistical inference; we learned how to create confidence intervals and test hypotheses about a proportion. Now we’ve looked at inference for the difference in two proportions. In doing so, perhaps the most important thing we’ve learned is that the concepts and interpretations are essentially the same—only the mechanics have changed slightly.

We’ve learned that hypothesis tests and confidence intervals for the difference in two proportions are based on Normal models. Both require us to find the standard error of the difference in
two proportions. We do that by adding the variances of the two sample proportions, assuming our two groups are independent. When we test a hypothesis that the two proportions are equal, we pool the sample data; for confidence intervals, we don’t pool.

Terms

- **Variances of independent random variables add**
  - The variance of a sum or difference of independent random variables is the sum of the variances of those variables.

- **Sampling distribution of the difference between two proportions**
  - The sampling distribution of \( \hat{p}_1 - \hat{p}_2 \) is, under appropriate assumptions, modeled by a Normal model with mean \( \mu = p_1 - p_2 \) and standard deviation \( SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \).

- **Two-proportion z-interval**
  - A two-proportion z-interval gives a confidence interval for the true difference in proportions, \( p_1 - p_2 \), in two independent groups.
    The confidence interval is \( (\hat{p}_1 - \hat{p}_2) \pm z^* \times SE(\hat{p}_1 - \hat{p}_2) \), where \( z^* \) is a critical value from the standard Normal model corresponding to the specified confidence level.

- **Pooling**
  - When we have data from different sources that we believe are homogeneous, we can get a better estimate of the common proportion and its standard deviation. We can combine, or pool, the data into a single group for the purpose of estimating the common proportion. The resulting pooled standard error is based on more data and is thus more reliable (if the null hypothesis is true and the groups are truly homogeneous).

- **Two-proportion z-test**
  - Test the null hypothesis \( H_0: p_1 - p_2 = 0 \) by referring the statistic
    \[
    z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{pooled}(\hat{p}_1 - \hat{p}_2)}
    \]
to a standard Normal model.

Skills

- Be able to state the null and alternative hypotheses for testing the difference between two population proportions.
- Know how to examine your data for violations of conditions that would make inference about the difference between two population proportions unwise or invalid.
- Understand that the formula for the standard error of the difference between two independent sample proportions is based on the principle that when finding the sum or difference of two independent random variables, their variances add.
- Know how to find a confidence interval for the difference between two proportions.
- Be able to perform a significance test of the natural null hypothesis that two population proportions are equal.
- Know how to write a sentence describing what is said about the difference between two population proportions by a confidence interval.
- Know how to write a sentence interpreting the results of a significance test of the null hypothesis that two population proportions are equal.
- Be able to interpret the meaning of a P-value in nontechnical language, making clear that the probability claim is made about computed values and not about the population parameter of interest.
- Know that we do not “accept” a null hypothesis if we fail to reject it.
INFERENCES FOR THE DIFFERENCE BETWEEN TWO PROPORTIONS ON THE COMPUTER

It is so common to test against the null hypothesis of no difference between the two true proportions that most statistics programs simply assume this null hypothesis. And most will automatically use the pooled standard deviation. If you wish to test a different null (say, that the true difference is 0.3), you may have to search for a way to do it.

Many statistics packages don’t offer special commands for inference for differences between proportions. As with inference for single proportions, most statistics programs want the “success” and “failure” status for each case. Usually these are given as 1 or 0, but they might be category names like “yes” and “no.” Often we just know the proportions of successes, \( \hat{p}_1 \) and \( \hat{p}_2 \), and the counts, \( n_1 \) and \( n_2 \). Computer packages don’t usually deal with summary data like these easily. Calculators typically do a better job.

---

**EXERCISES**

1. **Online social networking.** The Parents & Teens 2006 Survey of 935 12- to 17-year-olds found that, among teens aged 15–17, girls were significantly more likely to have used social networking sites and online profiles. 70% of the girls surveyed had used an online social network, compared to 54% of the boys. What does it mean to say that the difference in proportions is “significant”?

2. **Science news.** In 2007 a Pew survey asked 1447 Internet users about their sources of news and information about science. Among those who had broadband access at home, 34% said they would turn to the Internet for most of their science news. The report on this survey claims that this is not significantly different from the percentage (33%) who said they ordinarily get their science news from television. What does it mean to say that the difference is not significant?

3. **Name recognition.** A political candidate runs a week-long series of TV ads designed to attract public attention to his campaign. Polls taken before and after the ad campaign show some increase in the proportion of voters who now recognize this candidate’s name, with a P-value of 0.033. Is it reasonable to believe the ads may be effective?

4. **Origins.** In a 1993 Gallup poll, 47% of the respondents agreed with the statement “God created human beings pretty much in their present form at one time within the last 10,000 years or so.” When Gallup asked the same question in 2001, only 45% of those respondents agreed. Is it reasonable to conclude that there was a change in public opinion given that the P-value is 0.37? Explain.

5. **Revealing information.** 886 randomly sampled teens were asked which of several personal items of information they thought it okay to share with someone they had just met. 44% said it was okay to share their e-mail addresses, but only 29% said they would give out their cell phone numbers. A researcher claims that a two-proportion z-test could tell whether there was a real difference among all teens. Explain why that test would not be appropriate for these data.

6. **Regulating access.** When a random sample of 935 parents were asked about rules in their homes, 77% said they had rules about the kinds of TV shows their children could watch. Among the 790 of those parents whose teenage children had Internet access, 85% had rules about the kinds of Internet sites their teens could visit. That looks like a difference, but can we tell? Explain why a two-sample z-test would not be appropriate here.

7. **Gender gap.** A presidential candidate fears he has a problem with women voters. His campaign staff plans to run a poll to assess the situation. They’ll randomly sample 300 men and 300 women, asking if they have a favorable impression of the candidate. Obviously, the staff can’t know this, but suppose the candidate has a positive image with 59% of males but with only 53% of females.

   a) What sampling design is his staff planning to use?
   b) What difference would you expect the poll to show?
   c) Of course, sampling error means the poll won’t reflect the difference perfectly. What’s the standard deviation for the difference in the proportions?
   d) Sketch a sampling model for the size difference in proportions of men and women with favorable impressions of this candidate that might appear in a poll like this.
   e) Could the campaign be misled by the poll, concluding that there really is no gender gap? Explain.

8. **Buy it again?** A consumer magazine plans to poll car owners to see if they are happy enough with their vehicles that they would purchase the same model again. They’ll randomly select 450 owners of American-made cars and 450 owners of Japanese models. Obviously, the actual opinions of the entire population couldn’t be
known, but suppose 76% of owners of American cars and
78% of owners of Japanese cars would purchase another.
a) What sampling design is the magazine planning to use?
b) What difference would you expect their poll to show?
c) Of course, sampling error means the poll won’t reflect
the difference perfectly. What’s the standard deviation
for the difference in proportions?
d) Sketch a sampling model for the difference in propor-
tions that might appear in a poll like this.
e) Could the magazine be misled by the poll, concluding
that owners of American cars are much happier with
their vehicles than owners of Japanese cars? Explain.
9. Arthritis. The Centers for Disease Control and Preven-
tion reported on a survey of randomly selected Americans
age 65 and older, which found that 411 of 1012 men and
535 of 1062 women suffered from some form of arthritis.
a) Are the assumptions and conditions necessary for in-
ference satisfied? Explain.
b) Create a 95% confidence interval for the difference in
the proportions of senior men and women who have this
disease.
c) Interpret your interval in this context.
d) Does this confidence interval suggest that arthritis is
more likely to afflict women than men? Explain.

10. Graduation. In October 2000 the U.S. Department of
Commerce reported the results of a large-scale survey on
high school graduation. Researchers contacted more than
25,000 Americans aged 24 years to see if they had fin-
ished high school; 84.9% of the 12,460 males and 88.1% of
the 12,678 females indicated that they had high school
diplomas.
a) Are the assumptions and conditions necessary for in-
ference satisfied? Explain.
b) Create a 95% confidence interval for the difference in
graduation rates between males and females.
c) Interpret your confidence interval.
d) Does this provide strong evidence that girls are more
likely than boys to complete high school? Explain.
11. Pets. Researchers at the National Cancer Institute re-
leased the results of a study that investigated the effect of
weed-killing herbicides on house pets. They examined
827 dogs from homes where an herbicide was used on a
regular basis, diagnosing malignant lymphoma in 473 of
them. Of the 130 dogs from homes where no herbicides
were used, only 19 were found to have lymphoma.
a) What’s the standard error of the difference in the two
proportions?
b) Construct a 95% confidence interval for this difference.
c) State an appropriate conclusion.

12. Carpal tunnel. The painful wrist condition called
carpal tunnel syndrome can be treated with surgery or
less invasive wrist splints. In September 2002, Time maga-
azine reported on a study of 176 patients. Among the half
that had surgery, 80% showed improvement after three
months, but only 54% of those who used the wrist splints
improved.
a) What’s the standard error of the difference in the two
proportions?
b) Construct a 95% confidence interval for this difference.
c) State an appropriate conclusion.

13. Ear infections. A new vaccine was recently tested to
see if it could prevent the painful and recurrent ear infec-
tions that many infants suffer from. The Lancet, a medical
journal, reported a study in which babies about a year
old were randomly divided into two groups. One group
received vaccinations; the other did not. During the fol-
lowing year, only 333 of 2455 vaccinated children had ear
infections, compared to 499 of 2452 unvaccinated children
in the control group.
a) Are the conditions for inference satisfied?
b) Find a 95% confidence interval for the difference in
rates of ear infection.
c) Use your confidence interval to explain whether you
think the vaccine is effective.

14. Anorexia. The Journal of the American Medical Asso-
ciation reported on an experiment intended to see if the drug
Prozac® could be used as a treatment for the eating disor-
der anorexia nervosa. The subjects, women being treated
for anorexia, were randomly divided into two groups. Of
the 49 who received Prozac, 35 were deemed healthy a
year later, compared to 32 of the 44 who got the placebo.
a) Are the conditions for inference satisfied?
b) Find a 95% confidence interval for the difference in
outcomes.
c) Use your confidence interval to explain whether you
think Prozac is effective.

15. Another ear infection. In Exercise 13 you used a con-
fusion interval to examine the effectiveness of a vaccine
against ear infections in babies. Suppose that instead you
had conducted a hypothesis test. (Answer these questions
without actually doing the test.)
a) What hypotheses would you test?
b) State a conclusion based on your confidence interval.
c) What alpha level did your test use?
d) If that conclusion is wrong, which type of error did
you make?
e) What would be the consequences of such an error?

16. Anorexia again. In Exercise 14 you used a confidence
interval to examine the effectiveness of Prozac in treating
anorexia nervosa. Suppose that instead you had conducted
a hypothesis test. (Answer these questions without actu-
ally doing the test.)
a) What hypotheses would you test?
b) State a conclusion based on your confidence
interval.
c) What alpha level did your test use?
d) If that conclusion is wrong, which type of error did
you make?
e) What would be the consequences of such an error?

17. Teen smoking, part I. A Vermont study published in
December 2001 by the American Academy of Pediatrics
examined parental influence on teenagers’ decisions to
smoke. A group of students who had never smoked were
questioned about their parents’ attitudes toward smoking.
These students were questioned again two years later
to see if they had started smoking. The researchers found
that, among the 284 students who indicated that their
parents disapproved of kids smoking, 54 had become es-
established smokers. Among the 41 students who initially
said their parents were lenient about smoking, 11 became
smokers. Do these data provide strong evidence that parental attitude influences teenagers’ decisions about smoking?
   a) What kind of design did the researchers use?
   b) Write appropriate hypotheses.
   c) Are the assumptions and conditions necessary for inference satisfied?
   d) Test the hypothesis and state your conclusion.
   e) Explain in this context what your P-value means.
   f) If that conclusion is actually wrong, which type of error did you commit?
18. Depression. A study published in the Archives of General Psychiatry in March 2001 examined the impact of depression on a patient’s ability to survive cardiac disease. Researchers identified 450 people with cardiac disease, evaluated them for depression, and followed the group for 4 years. Of the 361 patients with no depression, 67 died. Of the 89 patients with minor or major depression, 26 died. Among people who suffer from cardiac disease, are depressed patients more likely to die than non-depressed ones?
   a) What kind of design was used to collect these data?
   b) Write appropriate hypotheses.
   c) Are the assumptions and conditions necessary for inference satisfied?
   d) Test the hypothesis and state your conclusion.
   e) Explain in this context what your P-value means.
   f) If your conclusion is actually incorrect, which type of error did you commit?
19. Teen smoking, part II. Consider again the Vermont study discussed in Exercise 17.
   a) Create a 95% confidence interval for the difference in the proportion of children who may smoke and have approving parents and those who may smoke and have disapproving parents.
   b) Interpret your interval in this context.
   c) Carefully explain what “95% confidence” means.
20. Depression revisited. Consider again the study of the association between depression and cardiac disease survivability in Exercise 18.
   a) Create a 95% confidence interval for the difference in survival rates.
   b) Interpret your interval in this context.
   c) Carefully explain what “95% confidence” means.
21. Pregnancy. In 1998, a San Diego reproductive clinic reported 42 live births to 157 women under the age of 38, but only 7 live births for 89 clients aged 38 and older. Is this strong evidence of a difference in the effectiveness of the clinic’s methods for older women?
   a) Was this an experiment? Explain.
   b) Test an appropriate hypothesis and state your conclusion in context.
   c) If you concluded there was a difference, estimate that difference with a confidence interval and interpret your interval in context.
22. Birthweight. In 2003 the Journal of the American Medical Association reported a study examining the possible impact of air pollution caused by the 9/11 attack on New York’s World Trade Center on the weight of babies. Researchers found that 8% of 182 babies born to mothers who were exposed to heavy doses of soot and ash on September 11 were classified as having low birth weight. Only 4% of 2300 babies born in another New York City hospital whose mothers had not been near the site of the disaster were similarly classified. Does this indicate a possibility that air pollution might be linked to a significantly higher proportion of low-weight babies?
   a) Was this an experiment? Explain.
   b) Test an appropriate hypothesis and state your conclusion in context.
   c) If you concluded there is a difference, estimate that difference with a confidence interval and interpret that interval in context.
23. Politics and sex. One month before the election, a poll of 630 randomly selected voters showed 54% planning to vote for a certain candidate. A week later it became known that he had had an extramarital affair, and a new poll showed only 51% of 1010 voters supporting him. Do these results indicate a decrease in voter support for his candidacy?
   a) Test an appropriate hypothesis and state your conclusion.
   b) If your conclusion turns out to be wrong, did you make a Type I or Type II error?
   c) If you concluded there was a difference, estimate that difference with a confidence interval and interpret your interval in context.
24. Shopping. A survey of 430 randomly chosen adults found that 21% of the 222 men and 18% of the 208 women had purchased books online.
   a) Is there evidence that men are more likely than women to make online purchases of books? Test an appropriate hypothesis and state your conclusion in context.
   b) If your conclusion in fact proves to be wrong, did you make a Type I or Type II error?
   c) Estimate this difference with a confidence interval.
   d) Interpret your interval in context.
25. Twins. In 2001, one county reported that, among 3132 white women who had babies, 94 were multiple births. There were also 20 multiple births to 606 black women. Does this indicate any racial difference in the likelihood of multiple births?
   a) Test an appropriate hypothesis and state your conclusion in context.
   b) If your conclusion is incorrect, which type of error did you commit?
26. Mammograms. A 9-year study in Sweden compared 21,088 women who had mammograms with 21,195 who did not. Of the women who underwent screening, 63 died of breast cancer, compared to 66 deaths among the control group. (The New York Times, Dec 9, 2001)
   a) Do these results support the effectiveness of regular mammograms in preventing deaths from breast cancer?
   b) If your conclusion is incorrect, what kind of error have you committed?
27. Pain. Researchers comparing the effectiveness of two pain medications randomly selected a group of patients...
who had been complaining of a certain kind of joint pain. They randomly divided these people into two groups, then administered the pain killers. Of the 112 people in the group who received medication A, 84 said this pain reliever was effective. Of the 108 people in the other group, 66 reported that pain reliever B was effective.

a) Write a 95% confidence interval for the percent of people who may get relief from this kind of joint pain by using medication A. Interpret your interval.

b) Write a 95% confidence interval for the percent of people who may get relief by using medication B. Interpret your interval.

c) Do the intervals for A and B overlap? What do you think this means about the comparative effectiveness of these medications?

d) Find a 95% confidence interval for the difference in the proportions of people who may find these medications effective. Interpret your interval.

e) Does this interval contain zero? What does that mean?

f) Why do the results in parts c and e seem contradictory? If we want to compare the effectiveness of these two pain relievers, which is the correct approach? Why?

28. Gender gap. Candidates for political office realize that different levels of support among men and women may be a crucial factor in determining the outcome of an election. One candidate finds that 52% of 473 men polled say they will vote for him, but only 45% of the 522 women in the poll express support.

a) Write a 95% confidence interval for the percent of male voters who may vote for this candidate. Interpret your interval.

b) Write and interpret a 95% confidence interval for the percent of female voters who may vote for him.

c) Do the intervals for males and females overlap? What do you think this means about the gender gap?

d) Find a 95% confidence interval for the difference in the proportions of males and females who will vote for this candidate. Interpret your interval.

e) Does this interval contain zero? What does that mean?

f) Why do the results in parts c and e seem contradictory? If we want to compare the effectiveness of these two pain relievers, which is the correct approach? Why?

29. Sensitive men. In August 2004, *Time* magazine, reporting on a survey of men’s attitudes, noted that “Young men are more comfortable than older men talking about their problems.” The survey reported that 80 of 129 surveyed 18- to 24-year-old men and 98 of 184 25- to 34-year-old men said they were comfortable. What do you think? Is *Time’s* interpretation justified by these numbers?

30. Retention rates. In 2004 the testing company ACT, Inc., reported on the percentage of first-year students at 4-year colleges who return for a second year. Their sample of 1139 students in private colleges showed a 74.9% retention rate, while the rate was 71.9% for the sample of 505 students at public colleges. Does this provide evidence that there’s a difference in retention rates of first-year students at public and private colleges?

31. Online activity checks. Are more parents checking up on their teen’s online activities? A Pew survey in 2004 found that 33% of 868 randomly sampled teens said that their parents checked to see what Web sites they visited. In 2006 the same question posed to 811 teens found 41% reporting such checks. Do these results provide evidence that more parents are checking?

32. Computer gaming. Who plays online or electronic games? A survey in 2006 found that 69% of 223 boys aged 12–14 said they “played computer or console games like Xbox or PlayStation . . . or games online.” Of 248 boys aged 15–17, only 62% played these games. Is this evidence of a real age-based difference?