+

Chapter 9: Testing a Claim

Section 9.1

Significance Tests: The Basics

The Practice of Statistics, 4th edition – For AP*

STARNES, YATES, MOORE

Chapter 9
Testing a Claim

- **9.1 Significance Tests: The Basics**
- 9.2 Tests about a Population Proportion
- 9.3 Tests about a Population Mean

Section 9.1
Significance Tests: The Basics

Learning Objectives

After this section, you should be able to…

- ✓ STATE correct hypotheses for a significance test about a population proportion or mean.
- ✓ INTERPRET *P*-values in context.
- ✓ INTERPRET a Type I error and a Type II error in context, and give the consequences of each.
- ✓ DESCRIBE the relationship between the significance level of a test, *P*(Type II error), and power.

Introduction

Confidence intervals are one of the two most common types of statistical inference. Use a confidence interval when your goal is to estimate a population parameter. The second common type of inference, called *significance tests*, has a different goal: to assess the evidence provided by data about some claim concerning a population.

A **significance test** is a formal procedure for comparing observed data with a claim (also called a hypothesis) whose truth we want to assess. The claim is a statement about a parameter, like the population proportion p or the population mean μ. We express the results of a significance test in terms of a probability that measures how well the data and the claim agree.

In this chapter, we'll learn the underlying logic of statistical tests, how to perform tests about population proportions and population means, and how tests are connected to confidence intervals.

The Reasoning of Significance Tests

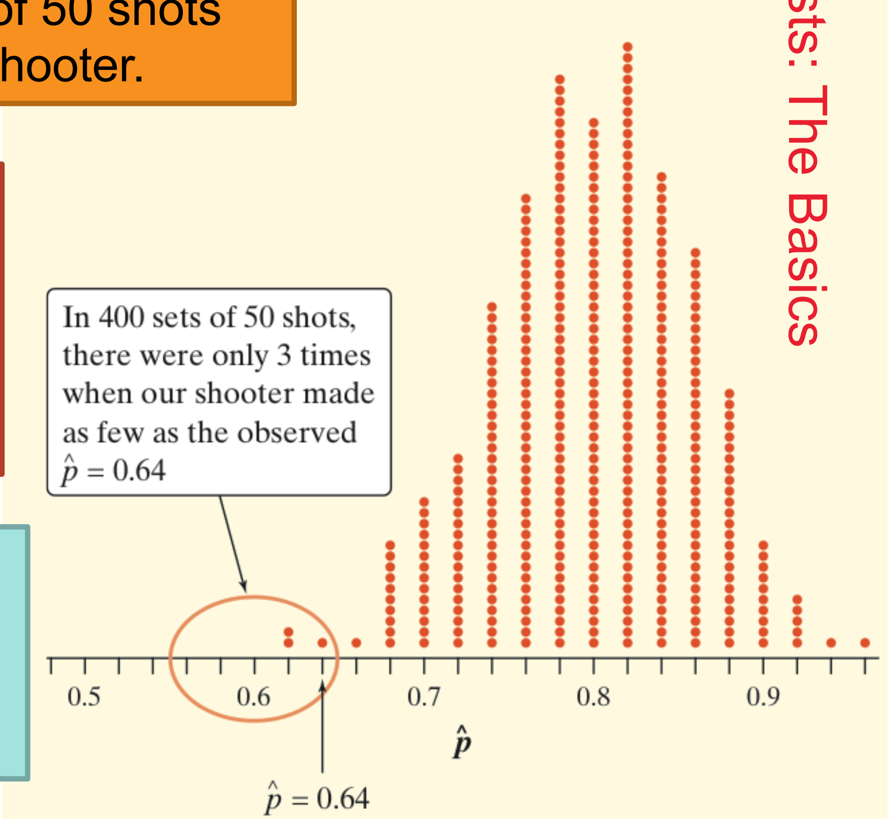
Suppose a basketball player claimed to be an 80% free-throw shooter. To test this claim, we have him attempt 50 free-throws. He makes 32 of them. His sample proportion of made shots is 32/50 = 0.64.

What can we conclude about the claim based on this sample data?

We can use software to simulate 400 sets of 50 shots assuming that the player is really an 80% shooter.

You can say how strong the evidence against the player's claim is by giving the probability that he would make as few as 32 out of 50 free throws if he really makes 80% in the long run.

The observed statistic is so unlikely if the actual parameter value is p = 0.80 that it gives convincing evidence that the player's claim is not true.



The Reasoning of Significance Tests

Based on the evidence, we might conclude the player's claim is incorrect.

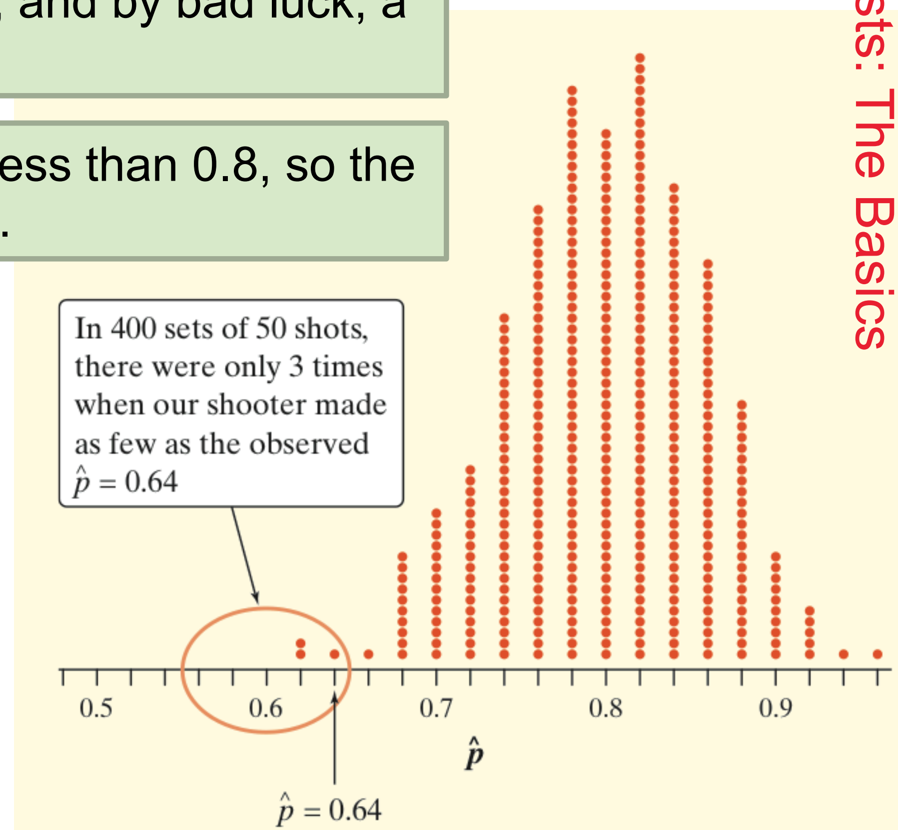
In reality, there are two possible explanations for the fact that he made only 64% of his free throws.

1) The player's claim is correct ($p = 0.8$), and by bad luck, a very unlikely outcome occurred.

2) The population proportion is actually less than 0.8, so the sample result is not an unlikely outcome.

Basic Idea

An outcome that would rarely happen if a claim were true is good evidence that the claim is not true.



Stating Hypotheses

A significance test starts with a careful statement of the claims we want to compare. The first claim is called the **null hypothesis**. Usually, the null hypothesis is a statement of "no difference." The claim we hope or suspect to be true instead of the null hypothesis is called the **alternative hypothesis**.

Definition:

The claim tested by a statistical test is called the **null hypothesis** (H_0). The test is designed to assess the strength of the evidence against the null hypothesis. Often the null hypothesis is a statement of "no difference."

The claim about the population that we are trying to find evidence for is the **alternative hypothesis** (H_a).

In the free-throw shooter example, our hypotheses are

$$H_0 : p = 0.80$$
$$H_a : p < 0.80$$

where p is the long-run proportion of made free throws.

Stating Hypotheses

In any significance test, the null hypothesis has the form

$$H_0 : \textbf{parameter} = \textbf{value}$$

The alternative hypothesis has one of the forms

$$H_a : \textbf{parameter} < \textbf{value}$$
$$H_a : \textbf{parameter} > \textbf{value}$$
$$H_a : \textbf{parameter} \neq \textbf{value}$$

To determine the correct form of H_a, read the problem carefully.

Definition:

The alternative hypothesis is **one-sided** if it states that a parameter is *larger than* the null hypothesis value or if it states that the parameter is *smaller than* the null value.

It is **two-sided** if it states that the parameter is *different* from the null hypothesis value (it could be either larger or smaller).

✓ Hypotheses always refer to a *population*, not to a sample. Be sure to state H_0 and H_a in terms of *population parameters*.

✓ It is *never* correct to write a hypothesis about a sample statistic, such as

Example: Studying Job Satisfaction

Does the job satisfaction of assembly-line workers differ when their work is machine-paced rather than self-paced? One study chose 18 subjects at random from a company with over 200 workers who assembled electronic devices. Half of the workers were assigned at random to each of two groups. Both groups did similar assembly work, but one group was allowed to pace themselves while the other group used an assembly line that moved at a fixed pace. After two weeks, all the workers took a test of job satisfaction. Then they switched work setups and took the test again after two more weeks. The response variable is the difference in satisfaction scores, self-paced minus machine-paced.

a) Describe the parameter of interest in this setting.

The parameter of interest is the mean μ of the differences (*self-paced minus machine-paced*) in job satisfaction scores in the population of all assembly-line workers at this company.

b) State appropriate hypotheses for performing a significance test.

Because the initial question asked whether job satisfaction differs, the alternative hypothesis is two-sided; that is, either $\mu < 0$ or $\mu > 0$. For simplicity, we write this as $\mu \neq 0$. That is,

$$H_0: \mu = 0$$
$$H_a: \mu \neq 0$$

Interpreting *P*-Values

The null hypothesis H_0 states the claim that we are seeking evidence against. The probability that measures the strength of the evidence against a null hypothesis is called a ***P*-value**.

Definition:

The probability, computed assuming H_0 is true, that the statistic would take a value as extreme as or more extreme than the one 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.

✓ Small *P*-values are evidence against H_0 because they say that the observed result is unlikely to occur when H_0 is true.

✓ Large *P*-values fail to give convincing evidence against H_0 because they say that the observed result is likely to occur by chance when H_0 is true.

Example: Studying Job Satisfaction

For the job satisfaction study, the hypotheses are

$$H_0: \mu = 0$$
$$H_a: \mu \neq 0$$

a) Explain what it means for the null hypothesis to be true in this setting.

In this setting, H_0: $\mu = 0$ says that the mean difference in satisfaction scores (*self-paced - machine-paced*) for the entire population of assembly-line workers at the company is 0. If H_0 is true, then the workers don't favor one work environment over the other, on average.

b) Interpret the *P*-value in context.

An outcome that would occur so often just by chance (almost 1 in every 4 random samples of 18 workers) when H_0 is true is not convincing evidence against H_0.
We fail to reject H_0: $\mu = 0$.

Statistical Significance

The final step in performing a significance test is to draw a conclusion about the competing claims you were testing. We will make one of two decisions based on the strength of the evidence against the null hypothesis (and in favor of the alternative hypothesis) -- **reject H_0 or fail to reject H_0**.

- ✓ If our sample result is too unlikely to have happened by chance assuming H_0 is true, then we'll reject H_0.
- ✓ Otherwise, we will fail to reject H_0.

Note: A fail-to-reject H_0 decision in a significance test doesn't mean that H_0 is true. For that reason, you should never "*accept* H_0" or use language implying that you believe H_0 is true.

In a nutshell, our conclusion in a significance test comes down to

P-value small → reject H_0 → conclude H_a (in context)

P-value large → fail to reject H_0 → cannot conclude H_a (in context)

Statistical Significance

There is no rule for how small a *P*-value we should require in order to reject H_0 — it's a matter of judgment and depends on the specific circumstances. But we can compare the *P*-value with a fixed value that we regard as decisive, called the **significance level**. We write it as α, the Greek letter alpha. When our *P*-value is less than the chosen α, we say that the result is **statistically significant**.

Definition:

If the *P*-value is smaller than alpha, we say that the data are **statistically significant at level** $\boldsymbol{\alpha}$. In that case, we reject the null hypothesis H_0 and conclude that there is convincing evidence in favor of the alternative hypothesis H_a.

When we use a fixed level of significance to draw a conclusion in a significance test,

P-value < α → reject H_0 → conclude H_a (in context)

P-value ≥ α → fail to reject H_0 → cannot conclude H_a (in context)

Example: Better Batteries

A company has developed a new deluxe AAA battery that is supposed to last longer than its regular AAA battery. However, these new batteries are more expensive to produce, so the company would like to be convinced that they really do last longer. Based on years of experience, the company knows that its regular AAA batteries last for 30 hours of continuous use, on average. The company selects an SRS of 15 new batteries and uses them continuously until they are completely drained. A significance test is performed using the hypotheses

$$H_0 : \mu = 30 \text{ hours}$$
$$H_a : \mu > 30 \text{ hours}$$

where μ is the true mean lifetime of the new deluxe AAA batteries. The resulting *P*-value is 0.0276.

a) What conclusion can you make for the significance level $\alpha = 0.05$?

Since the *P*-value, 0.0276, is less than $\alpha = 0.05$, the sample result is statistically significant at the 5% level. We have sufficient evidence to reject H_0 and conclude that the company's deluxe AAA batteries last longer than 30 hours, on average.

b) What conclusion can you make for the significance level $\alpha = 0.01$?

Since the *P*-value, 0.0276, is greater than $\alpha = 0.01$, the sample result is not statistically significant at the 1% level. We do not have enough evidence to reject H_0 in this case. therefore, we cannot conclude that the deluxe AAA batteries last longer than 30 hours, on average.

Type I and Type II Errors

When we draw a conclusion from a significance test, we hope our conclusion will be correct. But sometimes it will be wrong. There are two types of mistakes we can make. We can reject the null hypothesis when it's actually true, known as a **Type I error**, or we can fail to reject a false null hypothesis, which is a **Type II error**.

Definition:

If we reject H_0 when H_0 is true, we have committed a **Type I error**.

If we fail to reject H_0 when H_0 is false, we have committed a **Type II error**.

		Truth about the population: H_0 true	Truth about the population: H_0 false (H_a true)
Conclusion based on sample	Reject H_0	**Type I error**	*Correct conclusion*
	Fail to reject H_0	*Correct conclusion*	**Type II error**

Example: Perfect Potatoes

A potato chip producer and its main supplier agree that each shipment of potatoes must meet certain quality standards. If the producer determines that more than 8% of the potatoes in the shipment have "blemishes," the truck will be sent away to get another load of potatoes from the supplier. Otherwise, the entire truckload will be used to make potato chips. To make the decision, a supervisor will inspect a random sample of potatoes from the shipment. The producer will then perform a significance test using the hypotheses

$$H_0 : p = 0.08$$
$$H_a : p > 0.08$$

where p is the actual proportion of potatoes with blemishes in a given truckload.

Describe a Type I and a Type II error in this setting, and explain the consequences of each.

- A Type I error would occur if the producer concludes that the proportion of potatoes with blemishes is greater than 0.08 when the actual proportion is 0.08 (or less). *Consequence*: The potato-chip producer sends the truckload of acceptable potatoes away, which may result in lost revenue for the supplier.
- A Type II error would occur if the producer does not send the truck away when more than 8% of the potatoes in the shipment have blemishes. *Consequence*: The producer uses the truckload of potatoes to make potato chips. More chips will be made with blemished potatoes, which may upset consumers.

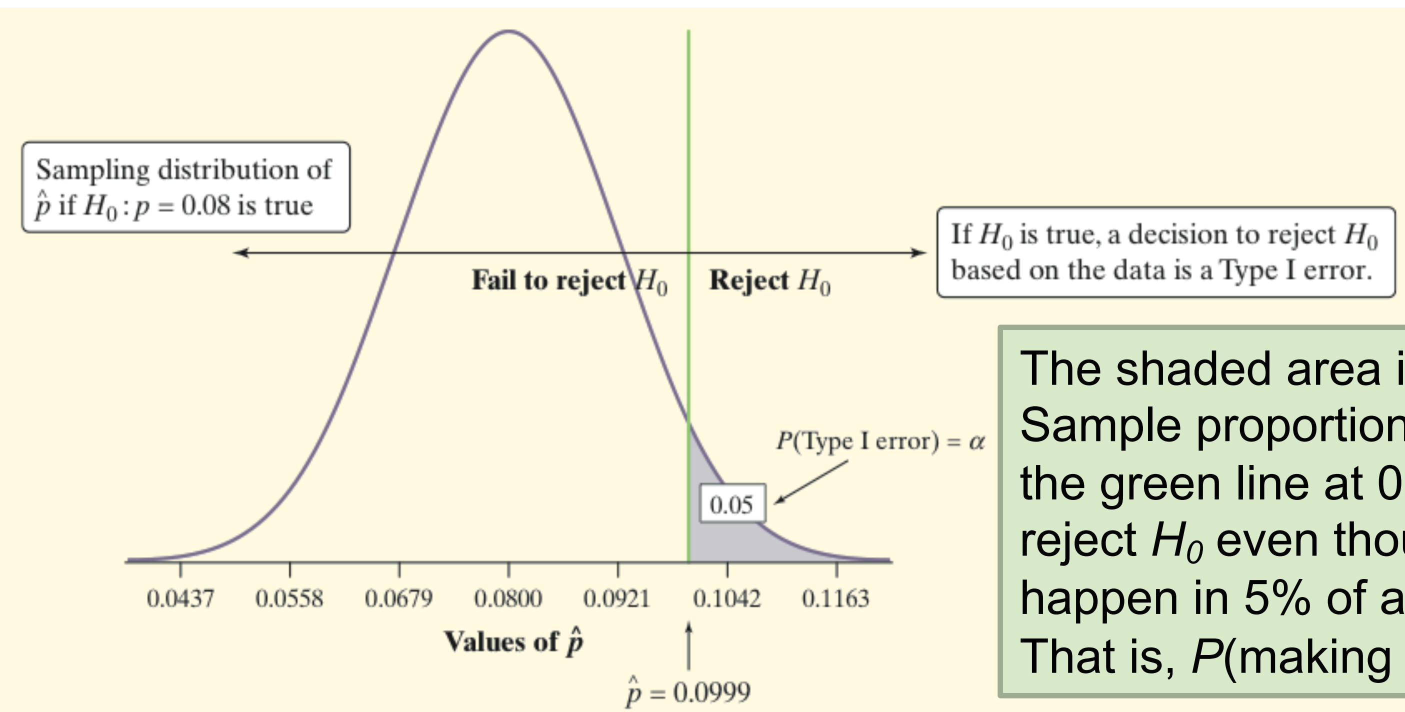
Error Probabilities

We can assess the performance of a significance test by looking at the probabilities of the two types of error. That's because statistical inference is based on asking, "What would happen if I did this many times?"

For the truckload of potatoes in the previous example, we were testing

$$H_0 : p = 0.08$$
$$H_a : p > 0.08$$

where p is the actual proportion of potatoes with blemishes. Suppose that the potato-chip producer decides to carry out this test based on a random sample of 500 potatoes using a 5% significance level ($\alpha = 0.05$).



The shaded area in the right tail is 5%. Sample proportion values to the right of the green line at 0.0999 will cause us to reject H_0 even though H_0 is true. This will happen in 5% of all possible samples. That is, P(making a Type I error) = 0.05.

Error Probabilities

The probability of a Type I error is the probability of rejecting H_0 when it is really true. As we can see from the previous example, this is exactly the significance level of the test.

Significance and Type I Error

The significance level α of any fixed level test is the probability of a Type I error. That is, α is the probability that the test will reject the null hypothesis H_0 when H_0 is in fact true. Consider the consequences of a Type I error before choosing a significance level.

What about Type II errors? A significance test makes a Type II error when it fails to reject a null hypothesis that really is false. There are many values of the parameter that satisfy the alternative hypothesis, so we concentrate on one value. We can calculate the probability that a test *does* reject H_0 when an alternative is true. This probability is called the **power** of the test against that specific alternative.

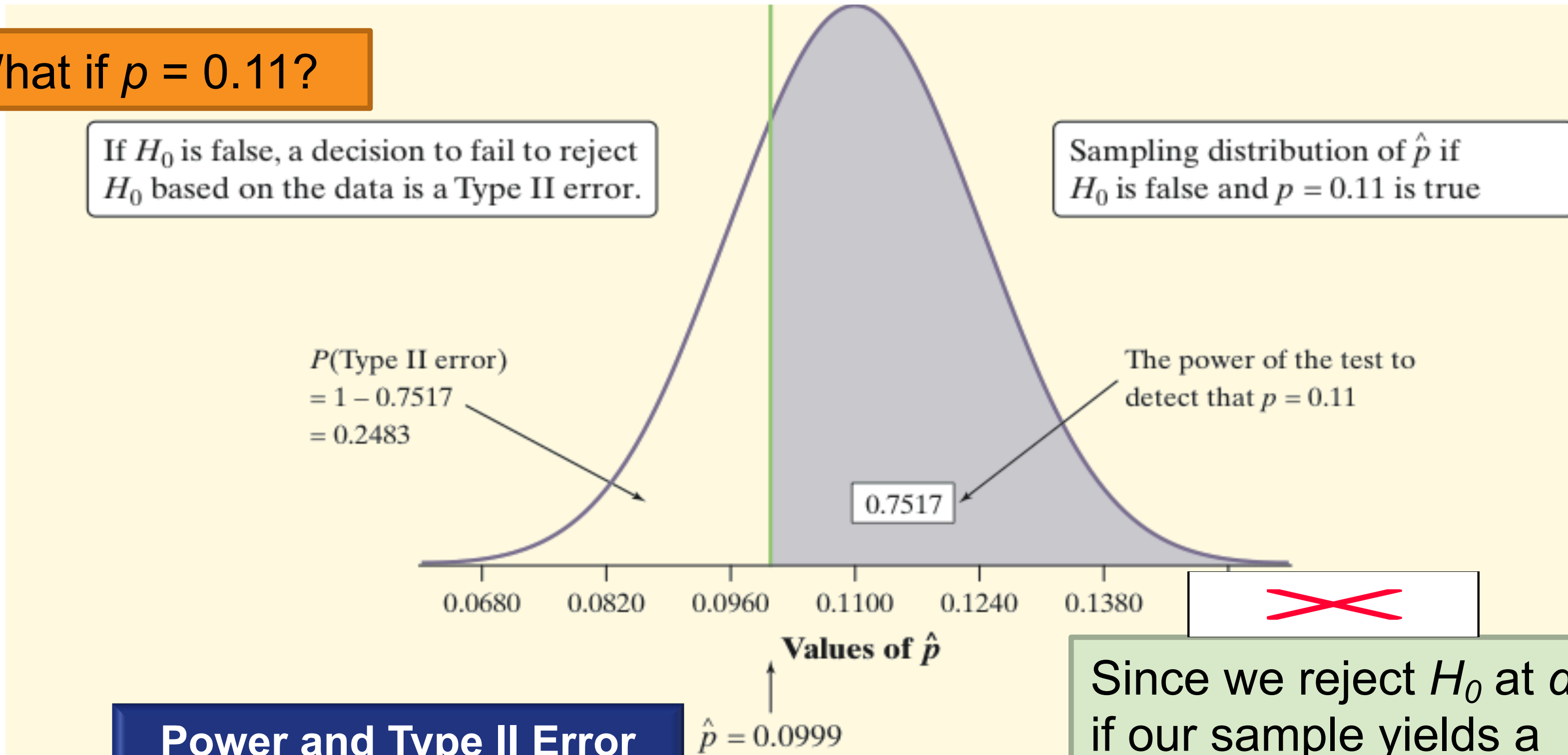
Definition:

The **power** of a test against a specific alternative is the probability that the test will reject H_0 at a chosen significance level α when the specified alternative value of the parameter is true.

Error Probabilities

The potato-chip producer wonders whether the significance test of $H_0 : p = 0.08$ versus $H_a : p > 0.08$ based on a random sample of 500 potatoes has enough power to detect a shipment with, say, 11% blemished potatoes. In this case, a particular Type II error is to fail to reject $H_0 : p = 0.08$ when $p = 0.11$.

What if $p = 0.11$?



Power and Type II Error

The power of a test against any alternative is 1 minus the probability of a Type II error for that alternative; that is, power = 1 - β.

Since we reject H_0 at α= 0.05 if our sample yields a proportion > 0.0999, we'd correctly reject the shipment about 75% of the time.

Planning Studies: The Power of a Statistical Test

How large a sample should we take when we plan to carry out a significance test? The answer depends on what alternative values of the parameter are important to detect.

Summary of influences on the question "How many observations do I need?"

- If you insist on a smaller significance level (such as 1% rather than 5%), you have to take a larger sample. A smaller significance level requires stronger evidence to reject the null hypothesis.
- If you insist on higher power (such as 99% rather than 90%), you will need a larger sample. Higher power gives a better chance of detecting a difference when it is really there.
- At any significance level and desired power, detecting a small difference requires a larger sample than detecting a large difference.

Section 9.1
Significance Tests: The Basics

Summary

In this section, we learned that…

- ✓ A **significance test** assesses the evidence provided by data against a **null hypothesis** H_0 in favor of an **alternative hypothesis** H_a.
- ✓ The hypotheses are stated in terms of population parameters. Often, H_0 is a statement of no change or no difference. H_a says that a parameter differs from its null hypothesis value in a specific direction (**one-sided alternative**) or in either direction (**two-sided alternative**).
- ✓ The reasoning of a significance test is as follows. Suppose that the null hypothesis is true. If we repeated our data production many times, would we often get data as inconsistent with H_0 as the data we actually have? If the data are unlikely when H_0 is true, they provide evidence against H_0 .
- ✓ The ***P*-value** of a test is the probability, computed supposing H_0 to be true, that the statistic will take a value at least as extreme as that actually observed in the direction specified by H_a .

+

Section 9.1
Significance Tests: The Basics

Summary

- ✓ Small *P*-values indicate strong evidence against H_0 . To calculate a *P*-value, we must know the sampling distribution of the test statistic when H_0 is true. There is no universal rule for how small a *P*-value in a significance test provides convincing evidence against the null hypothesis.

- ✓ If the *P*-value is smaller than a specified value α (called the **significance level**), the data are **statistically significant** at level α. In that case, we can reject H_0 . If the *P*-value is greater than or equal to α, we fail to reject H_0 .

- ✓ A **Type I error** occurs if we reject H_0 when it is in fact true. **A Type II error** occurs if we fail to reject H_0 when it is actually false. In a fixed level α significance test, the probability of a Type I error is the significance level α.

- ✓ The power of a significance test against a specific alternative is the probability that the test will reject H_0 when the alternative is true. **Power** measures the ability of the test to detect an alternative value of the parameter. For a specific alternative, *P*(Type II error) = 1 - power.

Looking Ahead…

<u>In the next Section…</u>

We'll learn how to test a claim about a population proportion.

We'll learn about

- ✓ **Carrying out a significance test**
- ✓ **The one-sample *z* test for a proportion**
- ✓ **Two-sided tests**
- ✓ **Why confidence intervals give more information than significance tests**