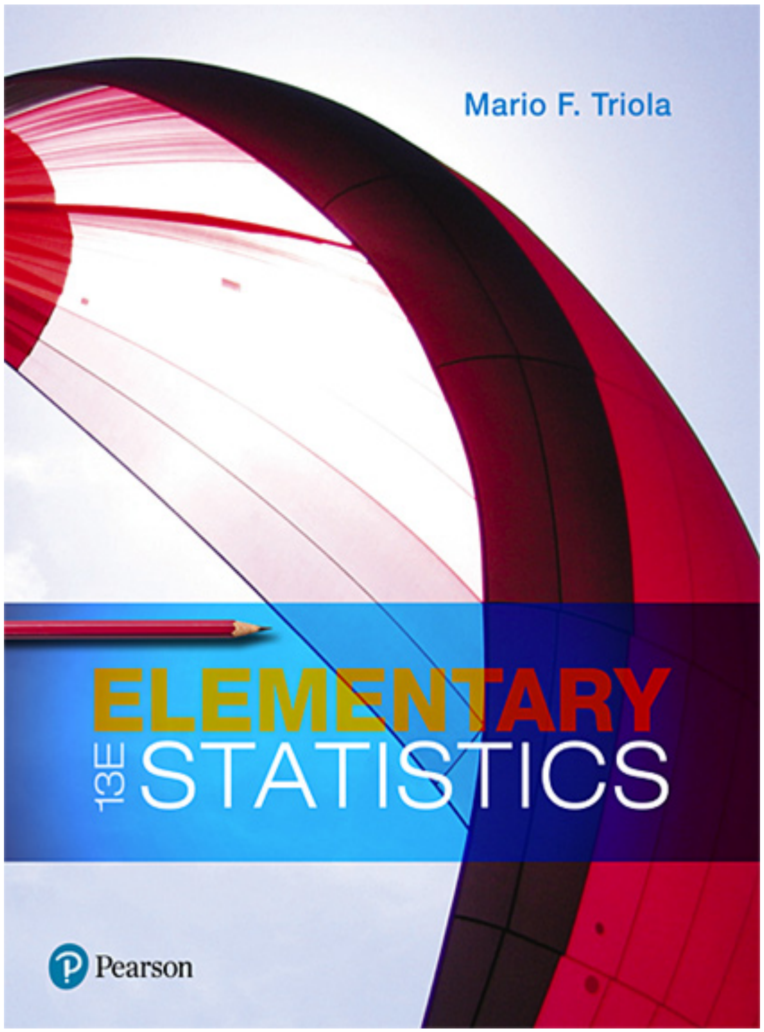
Elementary Statistics

Thirteenth Edition



Chapter 8

Hypothesis Testing

Hypothesis Testing

8-1 Basics of Hypothesis Testing

8-2 Testing a Claim about a Proportion

8-3 Testing a Claim About a Mean

8-4 Testing a Claim About a Standard Deviation or Variance

Key Concept

In this section we present key components of a formal hypothesis test. The concepts in this section are general and apply to hypothesis tests involving proportions, means, or standard deviations or variances.

Hypothesis and Hypothesis Test

- Hypothesis
  - In statistics, a **hypothesis** is a claim or statement about a property of a population.
- Hypothesis Test
  - A **hypothesis test** (or **test of significance**) is a procedure for testing a claim about a property of a population.

Example: Majority of Consumers are not Comfortable with Drone Deliveries (1 of 6)

1009 consumers were asked if they are comfortable with having drones deliver their purchases, and 54% (or 545) of them responded with "no." Using p to denote the proportion of consumers not comfortable with drone deliveries, the "majority" claim is equivalent to the claim that the proportion is greater than half, or $p > 0.5$. The expression $p > 0.5$ is the symbolic form of the original claim.

Example: Majority of Consumers are not Comfortable with Drone Deliveries (2 of 6)

The Big Picture We have the claim that the population proportion p is such that $p > 0.5$. Among 1009 consumers, how many do we need to get a **significantly high** number who are not comfortable with drone delivery?

- A result of 506 (or 50.1%) is just barely more than half, so 506 is clearly **not significantly high.**
- A result of 1006 (or 99.7%) is clearly **significantly high.** But what about the result of 545 (or 54.0%) that was actually obtained in the Pitney Bowes survey?
- Is 545 (or 54.0%) **significantly high?** The method of hypothesis testing allows us to answer that key question.

Example: Majority of Consumers are not Comfortable with Drone Deliveries (3 of 6)

Using Technology It is easy to obtain hypothesis-testing results using technology. The accompanying screen displays show results from four different technologies, **so we can use computers or calculators to do all of the computational heavy lifting.**

Example: Majority of Consumers are not Comfortable with Drone Deliveries (4 of 6)

Using Technology

Statdisk

Alternative Hypothesis:
p > p(hyp)

Sample proportion: 0.5401388
Test Statistic, z: 2.5500
Critical z: 1.6449
P-Value: 0.0054

90% Confidence interval:
0.5143312 < p < 0.5659463

Minitab

```
Test of p = 0.5 vs p > 0.5

Sample    X     N  Sample p  95% Lower Bound  Z-Value  P-Value
1       545  1009  0.540139         0.514331     2.55    0.005

Using the normal approximation.
```

TI-83/84 Plus

```
NORMAL FLOAT AUTO REAL RADIAN MP
        1-PropZTest
prop>.5
z=2.549995628
p=.005386238
p̂=.5401387512
n=1009
```

StatCrunch

Hypothesis test results:
p : Proportion of successes
$H_0 : p = 0.5$
$H_A : p > 0.5$

Proportion	Count	Total	Sample Prop.	Std. Err.	Z-Stat	P-value
p	545	1009	0.54013875	0.015740714	2.5499956	0.0054

Example: Majority of Consumers are not Comfortable with Drone Deliveries (5 of 6)

Using Technology

Examining the four screen displays, we see some common elements. They all display a "test statistic" of $z = 2.55$ (rounded), and they all include a "*P*-value" of 0.005 (rounded).

Focus on **understanding** how the hypothesis-testing procedure works and learn the associated terminology. Only then will results from technology make sense.

Example: Majority of Consumers are not Comfortable with Drone Deliveries (6 of 6)

Significance

Hypothesis tests are also called **tests of significance**. In Section 4-1 we used probabilities to determine when sample results are **significantly low** or **significantly high**. This chapter formalizes those concepts in a unified procedure that is used often throughout many different fields of application.

Null Hypothesis

- Null Hypothesis
  - The **null hypothesis** (denoted by H_0) is a statement that the value of a population parameter (such as proportion, mean, or standard deviation) is **equal to** some claimed value.

Alternative Hypothesis

- Alternative Hypothesis
  - The **alternative hypothesis** (denoted by H_1 or H_a or H_A) is a statement that the parameter has a value that somehow differs from the null hypothesis. For the methods of this chapter, the symbolic form of the alternative hypothesis must use one of these symbols: $<$, $>$, $\neq$.

Procedure for Hypothesis Tests

1. Identify the Claim
Identify the claim to be tested and express it in symbolic form.

↓

2. Give Symbolic Form
Give the symbolic form that must be true when the original claim is false.

↓

3. Identify Null and Alternative Hypothesis
Consider the two symbolic expressions obtained so far:

- **Alternative hypothesis H_1** is the one *NOT* containing equality, so H_1 uses the symbol $>$ or $<$ or $\neq$.
- **Null hypothesis H_0** is the symbolic expression that the parameter equals the fixed value being considered.

↓

4. Select Significance Level
Select the **significance level** α based on the seriousness of a type I error. Make α small if the consequences of rejecting a true H_0 are severe.

- The values of 0.05 and 0.01 are very common.

↓

5. Identify the Test Statistic
Identify the test statistic that is relevant to the test and determine its sampling distribution (such as normal, *t*, chi-square).

↓

***P*-Value Method**

6. Find Values
Find the value of the **test statistic** and the ***P*-value** (see Figure 8-3). Draw a graph and show the test statistic and *P*-value.

↓

7. Make a Decision

- **Reject H_0** if *P*-value $\leq \alpha$.
- **Fail to reject H_0** if *P*-value $> \alpha$.

↓

Critical Value Method

6. Find Values
Find the value of the **test statistic** and the **critical values**. Draw a graph showing the test statistic, critical value(s) and critical region.

↓

7. Make a Decision

- **Reject H_0** if the test statistic is in the critical region.
- **Fail to reject H_0** if the test statistic is not in the critical region.

↓

8. Restate Decision in Nontechnical Terms
Restate this previous decision in simple nontechnical terms, and address the original claim.

Confidence Interval Method

Confidence Interval Method

Construct a confidence interval with a confidence level selected as in Table 8-1.

Table 8-1 Confidence Level for Confidence Interval

Significance Level for Hypothesis Test	Two-Tailed Test	One-Tailed Test
0.01	99%	98%
0.05	95%	90%
0.10	90%	80%

Because a confidence interval estimate of a population parameter contains the likely values of that parameter, reject a claim that the population parameter has a value that is not included in the confidence interval.

Use the Original Claim to Create a Null Hypothesis H_0 and an Alternative Hypothesis H_1

Step 1. Identify the claim to be tested and express it in symbolic form.

Step 2. Give the symbolic form that must be true when the original claim is false.

Step 3. Consider the two symbolic expressions obtained so far:

- **Alternative hypothesis H_1** is the one **NOT** containing equality, so H_1 uses the symbol $<$ or $>$ or $\neq$.
- **Null hypothesis H_0** is the symbolic expression that the parameter **equals** the fixed value being considered.

Example: Drone Delivery (1 of 8)

Given the claim that "the majority of consumers are uncomfortable with drone delivery," we can apply Steps 1, 2, and 3 as follows.

Example: Drone Delivery (2 of 8)

Step 1: Identify the claim to be tested and express it in symbolic form. Using p to denote the probability of selecting a consumer uncomfortable with drone delivery, the claim that “the majority is uncomfortable with drone delivery” can be expressed in symbolic form as $p > 0.5$.

Step 2: Give the symbolic form that must be true when the original claim is false. If the original claim of $p > 0.5$ is false, then $p \leq 0.5$ must be true.

Example: Drone Delivery (3 of 8)

Step 3: This step is in two parts: Identify the alternative hypothesis H_1 and identify the null hypothesis H_0.

- Identify H_1: Using the two symbolic expressions $p > 0.5$ and $p \leq 0.5$, the alternative hypothesis H_1 is the one that does not contain equality. Of those two expressions, $p > 0.5$ does not contain equality, so we get H_1: $p > 0.5$
- Identify H_0: The null hypothesis H_0 is the symbolic expression that the parameter **equals** the fixed value being considered, so we get H_0: $p = 0.5$

Step 4: Significance Level α

- Significance Level
  - The **significance level** α for a hypothesis test is the probability value used as the cutoff for determining when the sample evidence constitutes **significant** evidence against the null hypothesis. By its nature, the significance level α is the probability of mistakenly rejecting the null hypothesis when it is true:

Significance level $\alpha = P$ (rejecting H_0 when H_0 is true)

Select the Significance Level *α*

Step 4. The **significance level *α*** is the same *α* introduced in sections 7-1, where we defined "critical value". Common choices for *α* are 0.05, 0.01, and 0.10; 0.05 is most common.

Identify the test statistic that is relevant to the test and determine its sampling distribution (such as normal, *t*, χ^2) (1 of 2)

Step 5. Identify the test statistic that is relevant to the test and determine its sampling distribution (such as normal, *t*, χ^2). The table on the following slide lists parameters along with the corresponding sampling distributions.

Identify the test statistic that is relevant to the test and determine its sampling distribution (such as normal, *t*, χ^2) (2 of 2)

Parameter	**Sampling Distribution**	**Requirements**	**Test Statistic**
Proportion *p*	Normal (*z*)	$np \geq 5$ and $nq \geq 5$	$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$
Mean μ	*t*	σ not known and normally distributed population or σ not known and $n > 30$	$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
Mean μ	Normal (*z*)	σ known and normally distributed population or σ known and $n > 30$	$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
St. dev. σ or variance σ^2	χ^2	Strict requirement: normally distributed population	$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

Example: Drone Delivery (4 of 8)

The claim $p > 0.5$ is a claim about the population proportion p, so use the normal distribution, provided that the requirements are satisfied. (With $n = 1009$, $p = 0.5$, and $q = 0.5$ from the ongoing example, $np \geq 5$ and $nq \geq 5$ are both true.)

Find the Value of the Test Statistic, Then Find Either the *P*-Value or the Critical Values(s)

Step 6. Find the value of the **test statistic** and the ***P*-value** or **critical value(s)**.

Test Statistic

- Test Statistic
  - The **test statistic** is a value used in making a decision about the null hypothesis. It is found by converting the sample statistic (such as $\hat{p}$, $\bar{x}$, or s) to a score (such as z, t, or χ^2) with the assumption that the null hypothesis is true.

Example: Drone Delivery (5 of 8)

We have a claim made about the population proportion p, we have n = 1009 and x = 545, so $\hat{p} = \frac{x}{n} = 0.540.$

With the null hypothesis of H_0: p = 0.5, we are working with the assumption that p = 0.5, and it follows that $q = 1 - p = 0.5$. We can evaluate the test statistic as shown below. The test statistic of z = 2.55 from each of the previous technology displays is more accurate than the result of z = 2.54 shown below.

(If we replace 0.540 with $\frac{545}{1009} = 0.54013875$, we get z = 2.55.)

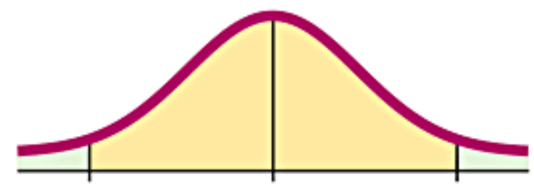
$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.540 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{1009}}} = 2.54$$

Critical Region

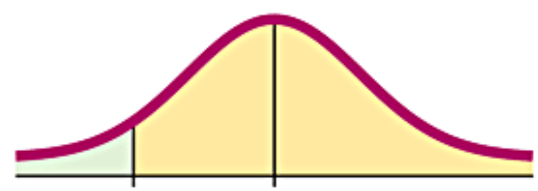
- Critical Region
  - The **critical region** (or **rejection region**) is the area corresponding to all values of the test statistic that cause us to reject the null hypothesis.

Two-Tailed, Left-Tailed, Right-Tailed

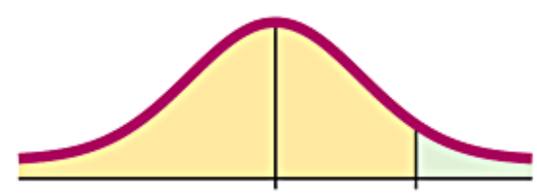
- **Two-tailed test:** The critical region is in the two extreme regions (tails) under the curve.
- **Left-tailed test:** The critical region is in the extreme left region (tail) under the curve.
- **Right-tailed test:** The critical region is in the extreme right region (tail) under the curve.



Sign used in H_1: $\neq$
Two-tailed test



Sign used in H_1: $<$
Left-tailed test



Sign used in H_1: $>$
Right-tailed test

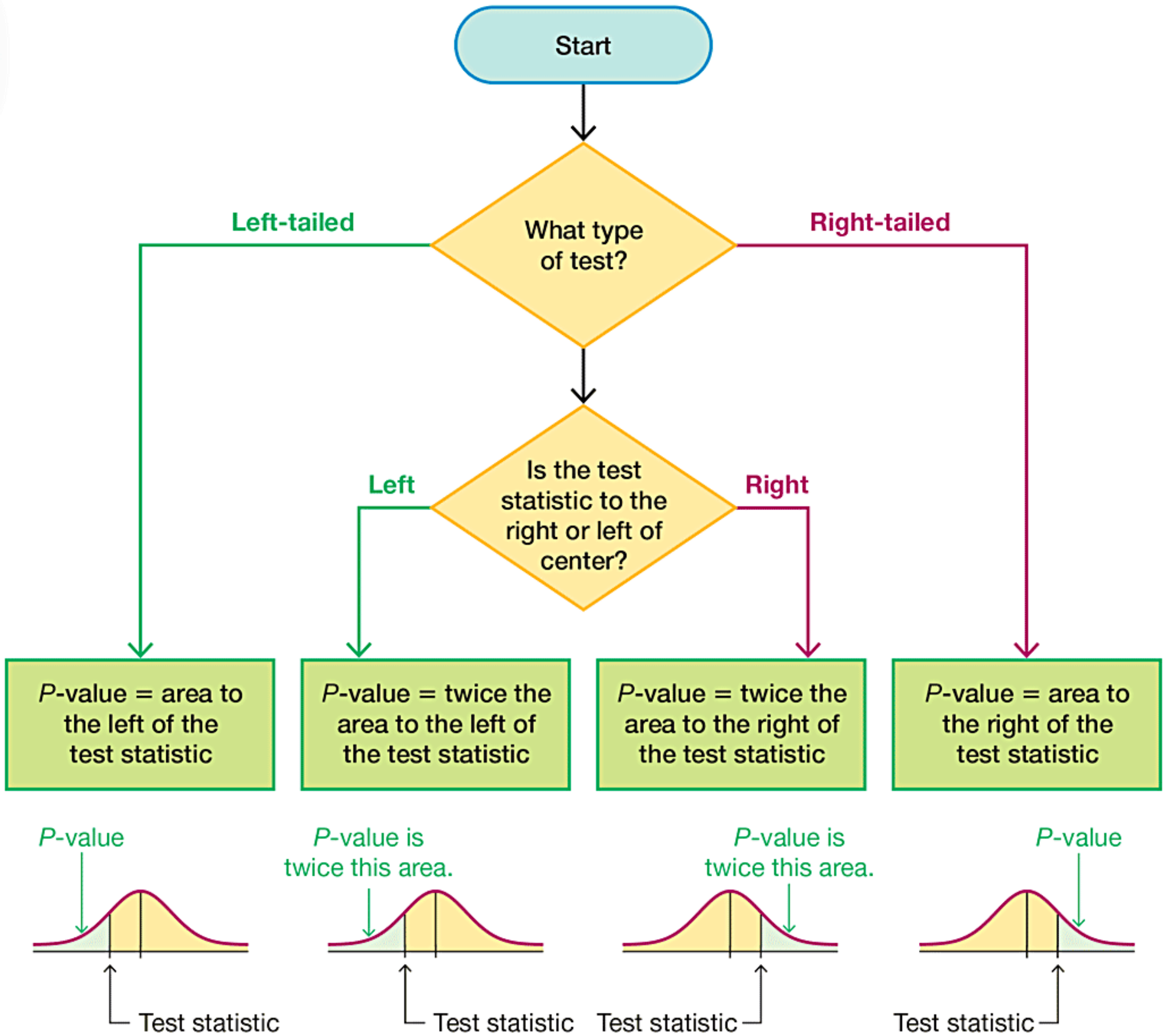
P-Value Method

- *P*-Value Method
  - In a hypothesis test, the ***P*-value** is the probability of getting a value of the test statistic that is **at least as extreme** as the test statistic obtained from the sample data, assuming that the null hypothesis is true.

Example: Drone Delivery (6 of 8)

Using the data from the previous problem, the test statistic is $z = 2.55$, and it has a normal distribution area of 0.0054 to its right, so a right-tailed test with test statistic $z = 2.55$ has a P-value of 0.0054. See the different technology displays given earlier, and note that each of them provides the same P-value of 0.005 after rounding.

Finding *P*-Values



Caution

Don't confuse a *P*-value with the parameter *p* or the statistic $\hat{p}$. Know the following notation:

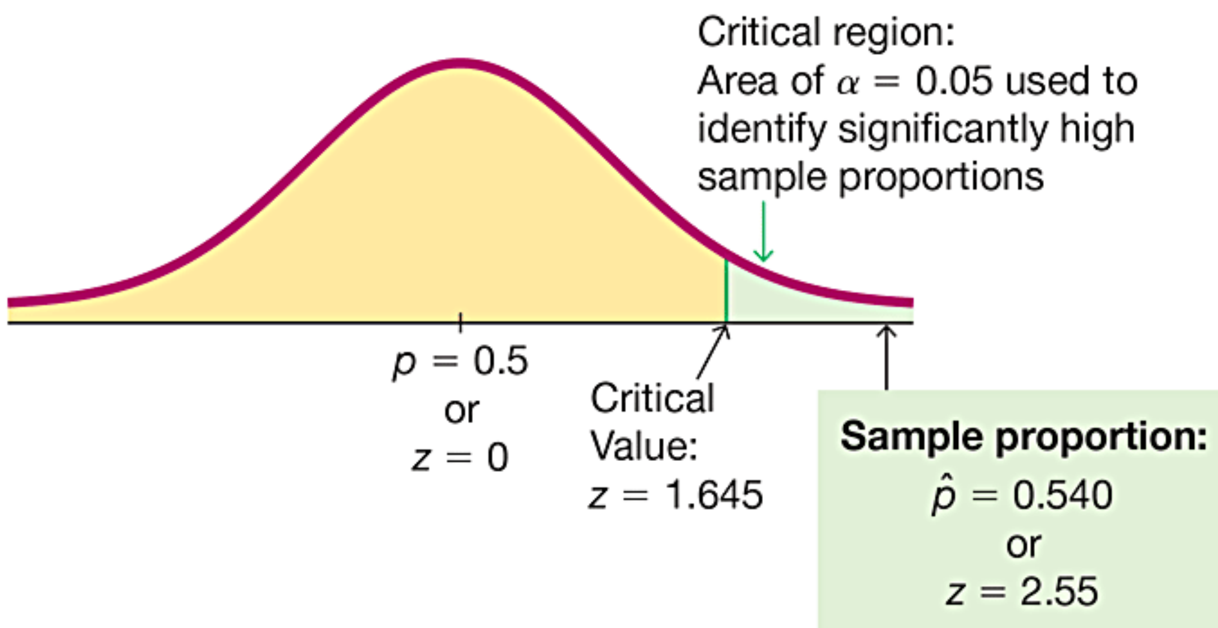
- ***P*-value** = probability of a test statistic at least as extreme as the one obtained
- ***p*** = population proportion
- $\hat{p}$ = sample proportion

Critical Value Method

- Critical Values
  – In a hypothesis test, the **critical value(s)** separates the critical region (where we reject the null hypothesis) from the values of the test statistic that do not lead to rejection of the null hypothesis.

Example: Drone Delivery

The critical region is shaded in green. The figure shows that with a significance level of $\alpha = 0.05$, the critical value is $z = 1.645$.



Make a decision to Either Reject H_0 or Fail to Reject H_0

Step 7. Make a decision to either reject H_0 or fail to reject H_0.

Decision Criteria for the *P*-Value Method:

- If *P*-value ≤ α, reject H_0 ("If the *P* is low, the null must go.")
- If *P*-value > α, fail to reject H_0.

Restate the Decision Using Simple and Nontechnical Terms

Step 8. Restate the decision using simple and nontechnical terms.

Without using technical terms not understood by most people, state a final conclusion that addresses the original claim with wording that can be understood by those without knowledge of statistical procedures.

Example: Drone Delivery (8 of 8)

There is sufficient evidence to support the claim that the majority of consumers are uncomfortable with drone deliveries.

Restate the Decision Using Simple and Nontechnical Terms (1 of 4)

Wording the Final Conclusion

For help in wording the final conclusion, refer to the table on the next slide, which lists the four possible circumstances and their corresponding conclusions.

Note that only the first case leads to wording indicating **support** for the original conclusion. If you want to support some claim, state it in such a way that it becomes the alternative hypothesis, and then hope that the null hypothesis gets rejected.

Restate the Decision Using Simple and Nontechnical Terms (2 of 4)

Wording the Final Conclusion

Condition	Conclusion
Original claim does not include equality, and you reject H_0.	"There is sufficient evidence to **support** the claim that … (original claim)"
Original claim does not include equality, and you fail to reject H_0.	"There is not sufficient evidence to **support** the claim that … (original claim)"
Original claim includes equality, and you reject H_0.	"There is sufficient evidence to warrant **rejection** of the claim that … (original claim)."
Original claim includes equality, and you fail to reject H_0.	"There is not sufficient evidence to warrant **rejection** of the claim that … (original claim)."

Restate the Decision Using Simple and Nontechnical Terms (3 of 4)

Accept or Fail to Reject?

We should say that we "fail to reject the null hypothesis" instead of saying that we "accept the null hypothesis." The term **accept** is misleading, because it implies incorrectly that the null hypothesis has been proved, but we can never prove a null hypothesis. The phrase **fail to reject** says more correctly that the available evidence isn't strong enough to warrant rejection of the null hypothesis.

Restate the Decision Using Simple and Nontechnical Terms (4 of 4)

Multiple Negatives

Final conclusions can include as many as three negative terms. For such confusing conclusions, it is better to restate them to be understandable. Instead of saying that "there is not sufficient evidence to warrant rejection of the claim of no difference between 0.5 and the population proportion," a better statement would be this: "Until stronger evidence is obtained, continue to assume that the population proportion is equal to 0.5."

Confidence Intervals for Hypothesis Tests (1 of 2)

A confidence interval estimate of a population parameter contains the likely values of that parameter. We should therefore reject a claim that the population parameter has a value that is not included in the confidence interval.

Confidence Intervals for Hypothesis Tests (2 of 2)

Equivalent Methods

A confidence interval estimate of a proportion might lead to a conclusion different from that of a hypothesis test.

Parameter	Is a confidence interval equivalent to a hypothesis test in the sense that they always lead to the same conclusion?
Proportion	No
Mean	Yes
Standard deviation or variance	Yes

Type I and Type II Errors (1 of 2)

- **Type I error:** The mistake of rejecting the null hypothesis when it is actually true. The symbol α (alpha) is used to represent the probability of a type I error.

  $\alpha = P(\textbf{type I error}) = P(\textbf{rejecting } H_0 \textbf{ when } H_0 \textbf{ is true})$

- **Type II error:** The mistake of failing to reject the null hypothesis when it is actually false. The symbol β(beta) is used to represent the probability of a type II error.

  $\beta = P(\textbf{type II error}) = P(\textbf{failing to reject } H_0 \textbf{ when } H_0 \textbf{ is false})$

Type I and Type II Errors (2 of 2)

Preliminary Conclusion	True State of Nature Null hypothesis is true	True State of Nature Null hypothesis is false
Reject H_0	**Type I error:** Reject a true H_0. P(type I error) = α	Correct decision
Fail to reject H_0	Correct decision	**Type II error:** Fail to reject a false H_0. P(type II error) = β

HINT FOR DESCRIBING TYPE I AND TYPE II ERRORS

Descriptions of a type I error and a type II error refer to the **null hypothesis** being true or false, but when wording a statement representing a type I error or a type II error, **be sure that the conclusion addresses the original claim** (which may or may not be the null hypothesis).

Example: Describing Type I and Type II Errors (1 of 3)

Consider the claim that a medical procedure designed to increase the likelihood of a baby girl is effective, so that the probability of a baby girl is $p > 0.5$. Given the following null and alternative hypotheses, write statements describing (a) a type I error, and (b) a type II error.

H_0: $p = 0.5$

H_1: $p > 0.5$ (original claim that will be addressed in the final conclusion)

Example: Describing Type I and Type II Errors (2 of 3)

Solution

a. **Type I Error:** A type I error is the mistake of rejecting a true null hypothesis, so the following is a type I error: In reality $p = 0.5$, but sample evidence leads us to conclude that $p > 0.5$. (In this case, a type I error is to conclude that the medical procedure is effective when in reality it has no effect.)

Example: Describing Type I and Type II Errors (3 of 3)

Solution

b. **Type II Error:** A type II error is the mistake of failing to reject the null hypothesis when it is false, so the following is a type II error: In reality $p > 0.5$, but we fail to support that conclusion. (In this case, a type II error is to conclude that the medical procedure has no effect, when it really is effective in increasing the likelihood of a baby girl.)

Power of a Hypothesis

- Power
  - The **power** of a hypothesis test is the probability $1 - \beta$ of rejecting a false null hypothesis. The value of the power is computed by using a particular significance level α and a **particular** value of the population parameter that is an alternative to the value assumed true in the null hypothesis.

Example: Power of a Hypothesis (1 of 5)

Consider these preliminary results from the XSORT method of gender selection: There were 13 girls among the 14 babies born to couples using the XSORT method. If we want to test the claim that girls are more likely ($p > 0.5$) with the XSORT method, we have the following null and alternative hypotheses:

$$H_0\colon p = 0.5 \quad H_1\colon p > 0.5$$

Example: Power of a Hypothesis (2 of 5)

Let's use a significance level of $\alpha = 0.05$. In addition to all of the given test components, finding power requires that we select a particular value of p that is an alternative to the value assumed in the null hypothesis H_0: $p = 0.5$. Find the values of power corresponding to these alternative values of p: 0.6, 0.7, 0.8, and 0.9.

Example: Power of a Hypothesis (3 of 5)

Solution

The values of power in the following table were found by using Minitab, and exact calculations are used instead of a normal approximation to the binomial distribution.

Specific Alternative Value of p	β	Power of Test = $1 - \beta$
0.6	0.820	0.180
0.7	0.564	0.436
0.8	0.227	0.773
0.9	0.012	0.988

Example: Power of a Hypothesis (4 of 5)

Interpretation

We see that this hypothesis test has power of 0.180 (or 18.0%) of rejecting H_0: $p = 0.5$ when the population proportion p is actually 0.6. That is, if the true population proportion is actually equal to 0.6, there is an 18.0% chance of making the correct conclusion of rejecting the false null hypothesis that $p = 0.5$. That low power of 18.0% is not so good.

Example: Power of a Hypothesis (5 of 5)

Interpretation

There is a 0.436 probability of rejecting $p = 0.5$ when the true value of p is actually 0.7. It makes sense that this test is more effective in rejecting the claim of $p = 0.5$ when the population proportion is actually 0.7 than when the population proportion is actually 0.6.

In general, increasing the difference between the assumed parameter value and the actual parameter value results in an increase in power.

Power and the Design of Experiments

Just as 0.05 is a common choice for a significance level, a power of at least 0.80 is a common requirement for determining that a hypothesis test is effective. (Some statisticians argue that the power should be higher, such as 0.85 or 0.90.) When designing an experiment, we might consider how much of a difference between the claimed value of a parameter and its true value is an important amount of difference.

When designing an experiment, a goal of having a power value of at least 0.80 can often be used to determine the minimum required sample size, as in the following example.

Example: Finding the Sample Size Required to Achieve 80% Power (1 of 2)

Here is a statement similar to one in an article from the **Journal of the American Medical Association:** "The trial design assumed that with a 0.05 significance level, 153 randomly selected subjects would be needed to achieve 80% power to detect a reduction in the coronary heart disease rate from 0.5 to 0.4." From that statement, we know the following:

- Before conducting the experiment, the researchers selected a significance level of 0.05 and a power of at least 0.80.

Example: Finding the Sample Size Required to Achieve 80% Power (2 of 2)

- The researchers decided that a reduction in the proportion of coronary heart disease from 0.5 to 0.4 is an important difference that they wanted to detect.
- Using a significance level of 0.05, power of 0.80, and the alternative proportion of 0.4, technology such as Minitab is used to find that the required minimum sample size is 153.

The researchers can then proceed by obtaining a sample of at least 153 randomly selected subjects. Because of factors such as dropout rates, the researchers are likely to need somewhat more than 153 subjects.