Large Sample Size Isn't Good Enough

Biased sample data should not be used for inferences, no matter how large the sample is. For example, in *Women and Love: A Cultural Revolution in Progress,* Shere Hite bases her conclusions on 4500 replies that she received after mailing 100,000 questionnaires to various women's groups. A *random* sample of 4500 subjects would usually provide good results, but Hite's sample is biased. It is criticized for over-representing women who join groups and women who feel strongly about the issues addressed. Because Hite's sample is biased, her inferences are not valid, even though the sample size of 4500 might seem to be sufficiently large.

Multiple Negatives When stating the final conclusion in nontechnical terms, it is possible to get correct statements with up to three negative terms. (Example: "There is *not* sufficient evidence to warrant *rejection* of the claim of *no* difference between 0.5 and the population proportion.") Such conclusions are confusing, so it is good to restate them in a way that makes them understandable, but care must be taken to not change the meaning. For example, 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," better statements would be these:

- Fail to reject the claim that the population proportion is equal to 0.5.
- Unless stronger evidence is obtained, continue to assume that the population proportion is equal to 0.5.

EXAMPLE 9 **Stating the Final Conclusion** Suppose a geneticist claims that the XSORT method of gender selection increases the likelihood of a baby girl. This claim of $p > 0.5$ becomes the alternative hypothesis, while the null hypothesis becomes $p = 0.5$. Further suppose that the sample evidence causes us to reject the null hypothesis of $p = 0.5$. State the conclusion in simple, nontechnical terms.

SOLUTION Refer to Figure 8-7. Because the original claim does not contain equality, it becomes the alternative hypothesis. Because we reject the null hypothesis, the wording of the final conclusion should be as follows: "There is sufficient evidence to support the claim that the XSORT method of gender selection increases the likelihood of a baby girl."

Errors in Hypothesis Tests

When testing a null hypothesis, we arrive at a conclusion of rejecting it or failing to reject it. Such conclusions are sometimes correct and sometimes wrong (even if we do everything correctly). Table 8-1 summarizes the two different types of errors that can be made, along with the two different types of correct decisions. We distinguish between the two types of errors by calling them type I and type II errors.

- **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.
- **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.

Because it can be difficult to remember which error is type I and which is type II, we recommend a mnemonic device, such as "routine for fun." Using only the consonants from those words (**R**ou**T**i**N**e **F**o**R** **F**u**N**), we can easily remember that a type I error is RTN: Reject True Null (hypothesis), whereas a type II error is FRFN: Fail to Reject a False Null (hypothesis).

Notation

α (alpha) = probability of a type I error (the probability of rejecting the null hypothesis when it is true)

β (beta) = probability of a type II error (the probability of failing to reject a null hypothesis when it is false)

Table 8-1 Type I and Type II Errors

		True State of Nature	
		The null hypothesis is true	The null hypothesis is false
Decision	We decide to reject the null hypothesis	**Type I error** (rejecting a true null hypothesis) $P(\text{type I error}) = \alpha$	Correct decision
	We fail to reject the null hypothesis	Correct decision	**Type II error** (failing to reject a false null hypothesis) $P(\text{type II error}) = \beta$

EXAMPLE 10 **Identifying Type I and Type II Errors** Assume that we are conducting a hypothesis test of the claim that a method of gender selection increases the likelihood of a baby girl, so that the probability of a baby girl is $p > 0.5$. Here are the null and alternative hypotheses:

$$H_0: p = 0.5$$
$$H_1: p > 0.5$$

Give statements identifying the following.

a. Type I error **b.** Type II error

SOLUTION

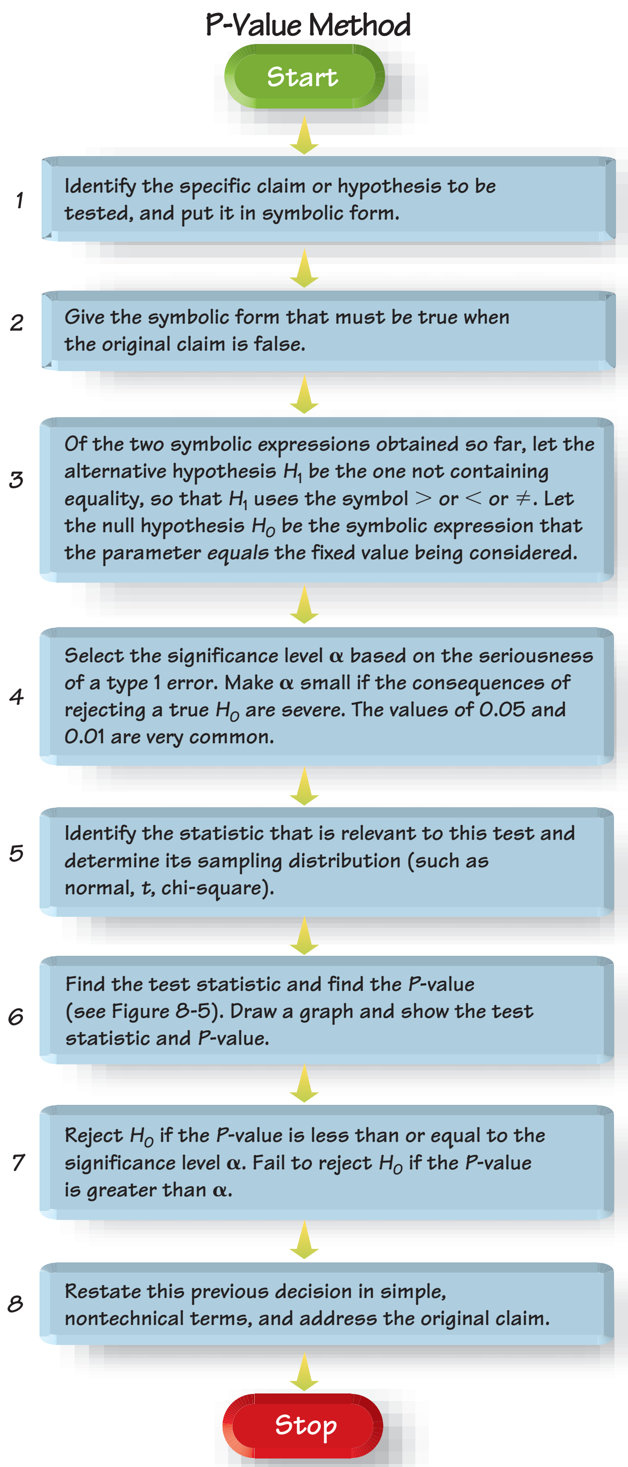
a. A type I error is the mistake of rejecting a true null hypothesis, so this is a type I error: Conclude that there is sufficient evidence to support $p > 0.5$, when in reality $p = 0.5$. That is, a type I error is made when we conclude that the gender selection method is effective when in reality it has no effect.

b. A type II error is the mistake of failing to reject the null hypothesis when it is false, so this is a type II error: Fail to reject $p = 0.5$ (and therefore fail to support $p > 0.5$) when in reality $p > 0.5$. That is, a type II error is made if we conclude that the gender selection method has no effect, when it really is effective in increasing the likelihood of a baby girl.

Controlling Type I and Type II Errors: One step in our standard procedure for testing hypotheses involves the selection of the significance level α (such as 0.05), which is the probability of a type I error. The values of α, β, and the sample size n are all related, so when you choose or determine any two of them, the third is automatically determined. One common practice is to select the significance level α, then select a sample size that is practical, so the value of β is determined. Generally try to use the largest α that you can tolerate, but for type I errors with more serious consequences, select smaller values of α. Then choose a sample size n as large as is reasonable, based on considerations of time, cost, and other relevant factors. Another common practice is to select α and β, so the required sample size n is automatically determined. (See Example 12 in Part 2 of this section.)

Comprehensive Hypothesis Test In this section we describe the individual components used in a hypothesis test, but the following sections will combine those components in comprehensive procedures. We can test claims about population parameters by using the *P*-value method summarized in Figure 8-8, the traditional method summarized in Figure 8-9, or we can use a confidence interval, as described on page 407.

Figure 8-8 ***P*-Value Method**



P-Value Method

Start

1 Identify the specific claim or hypothesis to be tested, and put it in symbolic form.

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

3 Of the two symbolic expressions obtained so far, let the alternative hypothesis H_1 be the one not containing equality, so that H_1 uses the symbol $>$ or $<$ or $\neq$. Let the null hypothesis H_0 be the symbolic expression that the parameter *equals* the fixed value being considered.

4 Select the significance level α based on the seriousness of a type 1 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 statistic that is relevant to this test and determine its sampling distribution (such as normal, *t*, chi-square).

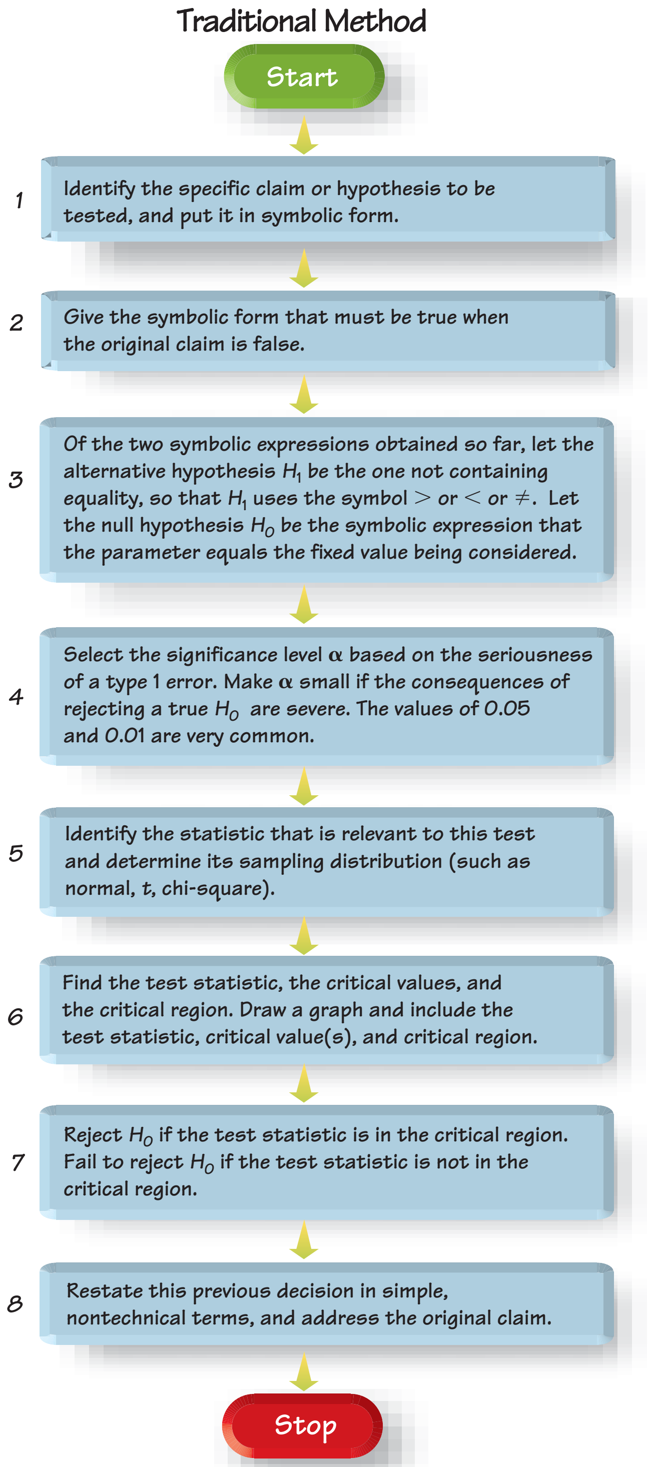
6 Find the test statistic and find the *P*-value (see Figure 8-5). Draw a graph and show the test statistic and *P*-value.

7 Reject H_0 if the *P*-value is less than or equal to the significance level α. Fail to reject H_0 if the *P*-value is greater than α.

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

Stop

Figure 8-9 **Traditional Method**



Traditional Method

Start

1 Identify the specific claim or hypothesis to be tested, and put it in symbolic form.

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

3 Of the two symbolic expressions obtained so far, let the alternative hypothesis H_1 be the one not containing equality, so that H_1 uses the symbol $>$ or $<$ or $\neq$. Let the null hypothesis H_0 be the symbolic expression that the parameter *equals* the fixed value being considered.

4 Select the significance level α based on the seriousness of a type 1 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 statistic that is relevant to this test and determine its sampling distribution (such as normal, *t*, chi-square).

6 Find the test statistic, the critical values, and the critical region. Draw a graph and include the test statistic, critical value(s), and critical region.

7 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 this previous decision in simple, nontechnical terms, and address the original claim.

Stop

Confidence Interval Method

Construct a confidence interval with a confidence level selected as in Table 8-2.

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.

Table 8-2 Confidence Level for Confidence Interval

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

Confidence Interval Method For two-tailed hypothesis tests construct a confidence interval with a confidence level of $1 - \alpha$; but for a one-tailed hypothesis test with significance level α, construct a confidence interval with a confidence level of $1 - 2\alpha$. (See Table 8-2 for common cases.) After constructing the confidence interval, use this criterion:

> **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.**

CAUTION

In some cases, a conclusion based on a confidence interval may be different from a conclusion based on a hypothesis test. See the comments in the individual sections that follow.

The exercises for this section involve isolated components of hypothesis tests, but the following sections will involve complete and comprehensive hypothesis tests.

Part 2: Beyond the Basics of Hypothesis Testing: The *Power* of a Test

We use β to denote the probability of failing to reject a false null hypothesis, so $P(\text{type II error}) = \beta$. It follows that $1 - \beta$ is the probability of rejecting a false null hypothesis, and statisticians refer to this probability as the *power* of a test, and they often use it to gauge the effectiveness of a hypothesis test in allowing us to recognize that a null hypothesis is false.

DEFINITION

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.

Note that in the above definition, determination of power requires a particular value that is an alternative to the value assumed in the null hypothesis. Consequently, a hypothesis test can have many different values of power, depending on the particular values of the population parameter chosen as alternatives to the null hypothesis.

EXAMPLE 11

Power of a Hypothesis Test Let's again 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:

$$H_0: p = 0.5 \qquad H_1: p > 0.5$$

Let's use $\alpha = 0.05$. In addition to all of the given test components, we need a particular value of p that is an alternative to the value assumed in the null hypothesis $H_0: p = 0.5$. Using the given test components along with different alternative values of p, we get the following examples of power values. These values of power were found by using Minitab, and exact calculations are used instead of a normal approximation to the binomial distribution.

continued

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

INTERPRETATION Based on the above list of power values, 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 good. There is a 0.564 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. (When identifying animals assumed to be horses, there's a better chance of rejecting an elephant as a horse (because of the greater difference) than rejecting a mule as a horse.) In general, increasing the difference between the assumed parameter value and the actual parameter value results in an increase in power, as shown in the above table.

Because the calculations of power are quite complicated, the use of technology is strongly recommended. (In this section, only Exercises 46–48 involve 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. If testing the effectiveness of the XSORT gender-selection method, a change in the proportion of girls from 0.5 to 0.501 is not very important. A change in the proportion of girls from 0.5 to 0.6 might be important. Such magnitudes of differences affect power. 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 12 **Finding Sample Size Required to Achieve 80% Power**

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." Before conducting the experiment, the researchers selected a significance level of 0.05 and a power of at least 0.80. They also 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 (by correctly rejecting the false null hypothesis). 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. Due to factors such as dropout rates, the researchers are likely to need somewhat more than 153 subjects. (See Exercise 48.)