A confidence interval is a useful way to report the result of an analysis because it sets limits on the expected result. In the absence of determinate error, a confidence interval indicates the range of values in which we expect to find the population’s expected mean. When we report a 95% confidence interval for the mass of a penny as 3.117 g ± 0.047 g, for example, we are claiming that there is only a 5% probability that the expected mass of penny is less than 3.070 g or more than 3.164 g.

Because a confidence interval is a statement of probability, it allows us to answer questions such as “Are the results for a newly developed method for determining cholesterol in blood significantly different from those obtained using a standard method?” or “Is there a significant variation in the composition of rainwater collected at different sites downwind from a coal-burning utility plant?”. In this section we introduce a general approach to the statistical analysis of data. Specific statistical tests are presented in Section 4.6.

### 4.5.1 Significance Testing

Let’s consider the following problem. To determine if a medication is effective in lowering blood glucose concentrations, we collect two sets of blood samples from a patient. We collect one set of samples immediately before administering the medication, and collect the second set of samples several hours later. After analyzing the samples, we report their respective means and variances. How do we decide if the medication was successful in lowering the patient’s concentration of blood glucose?

One way to answer this question is to construct normal distribution curves for each sample, and to compare them to each other. Three possible outcomes are shown in Figure 4.12. In Figure 4.12a, there is a complete separation of the normal distribution curves, strongly suggesting that the samples are significantly different. In Figure 4.12b, the normal distributions for the two samples almost completely overlap each other, suggesting that any difference between the samples is insignificant. Figure 4.12c, however, presents a dilemma. Although the means for the two samples appear to be different, there is sufficient overlap of the normal distributions that a significant number of possible outcomes could belong to either distribution. In this case the best we can do is to make a statement about the probability that the samples are significantly different.

![Figure 4.12](image)

**Figure 4.12** Three examples showing possible relationships between the normal distribution curves for two samples. In (a) the curves are completely separate, suggesting that the samples are significantly different from each other. In (b) the two curves are almost identical, suggesting that the samples are indistinguishable. The partial overlap of the curves in (c) means that the best we can do is to indicate the probability that there is a difference between the samples.

The process by which we determine the probability that there is a significant difference between two samples is called significance testing or hypothesis testing. Before discussing specific examples we will first establish a general approach
to conducting and interpreting significance tests.

### 4.5.2 Constructing a Significance Test

The purpose of a **significance test** is to determine whether the difference between two or more values is too large to be explained by indeterminate error. The first step in constructing a significance test is to state the problem as a yes or no question, such as “Is this medication effective at lowering a patient’s blood glucose levels?” A null hypothesis and an alternative hypothesis provide answers to the question. The **null hypothesis**, $H_0$, is that indeterminate error is sufficient to explain any differences in our data. The **alternative hypothesis**, $H_A$, is that the differences are too great to be explained by random error and, therefore, must be determinate. We test the null hypothesis, which we either retain or reject. If we reject the null hypothesis, then we must accept the alternative hypothesis, concluding that the difference is significant and that it cannot be explained by random error.

Failing to reject a null hypothesis is not the same as accepting it. We retain a null hypothesis because there is insufficient evidence to prove it incorrect. It is impossible to prove that a null hypothesis is true. This is an important point that is easy to forget. To appreciate this point let’s return to our sample of 100 pennies in [*Table 4.13*](#). After looking at the data you might propose the following null and alternative hypotheses.

- $H_0$: The mass of any U.S. penny in circulation is in the range of 2.900 g–3.200 g.
- $H_A$: A U.S. penny in circulation may have a mass less than 2.900 g or a mass of more than 3.200 g.

To test the null hypothesis you reach into your pocket, retrieve a penny, and determine its mass. If the penny’s mass is 2.512 g then you reject the null hypothesis, and accept the alternative hypothesis. Suppose that the penny’s mass is 3.162 g. Although this result increases your confidence in the null hypothesis, it does not prove it is correct because the next penny you pull from your pocket might weigh less than 2.900 g or more than 3.200 g.

After stating the null and alternative hypotheses, the second step is to choose a confidence level for the analysis. The confidence level defines the probability that we will reject the null hypothesis when it is, in fact, true. We can express this as our confidence in correctly rejecting the null hypothesis (e.g. 95%), or as the probability that we are incorrectly rejecting the null hypothesis. For the latter, the confidence level is given as $\alpha$, where

\[
\alpha = 1 - \frac{\text{confidence level (\%)} \times 100}{100}
\]

For a 95% confidence level, $\alpha$ is 0.05.

In this textbook, we use $\alpha$ to represent the probability of incorrectly rejecting the null hypothesis. In other textbooks this probability is given as $p$ (often read as “p-value”). Although the symbols differ, the meaning is the same.

The third step is to calculate an appropriate test statistic and to compare it to a critical value. The test statistic’s critical value defines a breakpoint between values that lead us to reject or to retain the null hypothesis. How we calculate the test statistic depends on what we are comparing, a topic we cover in section 4.6. The last step is to either retain the null hypothesis, or to reject it and accept the alternative hypothesis.

Note
The four steps for a statistical analysis of data:
1. Pose a question, and state the null hypothesis and the alternative hypothesis.
2. Choose a confidence level for the statistical analysis.
3. Calculate an appropriate test statistic and compare it to a critical value.
4. Either retain the null hypothesis, or reject it and accept the alternative hypothesis.

4.5.3 One-Tailed and Two-Tailed Significance Tests

Suppose we want to evaluate the accuracy of a new analytical method. We might use the method to analyze a Standard Reference Material containing a known concentration of analyte, \( \mu \). We analyze the standard several times, obtaining an mean value, \( \bar{X} \), for the analyte’s concentration. Our null hypothesis is that there is no difference between \( \bar{X} \) and \( \mu \)

\[ H_0: \bar{X} = \mu \]

If we conduct the significance test at \( \alpha = 0.05 \), then we retain the null hypothesis if a 95% confidence interval around \( \bar{X} \) contains \( \mu \). If the alternative hypothesis is

\[ H_A: \bar{X} \neq \mu \]

then we reject the null hypothesis and accept the alternative hypothesis if \( \mu \) lies in the shaded areas at either end of the sample’s probability distribution curve (Figure 4.13a). Each of the shaded areas accounts for 2.5% of the area under the probability distribution curve. This is a two-tailed significance test because we reject the null hypothesis for values of \( \mu \) at either extreme of the sample’s probability distribution curve.

We also can write the alternative hypothesis in two additional ways

\[ H_A: \bar{X} > \mu \]

\[ H_A: \bar{X} < \mu \]

rejecting the null hypothesis if \( \mu \) falls within the shaded areas shown in Figure 4.13b or Figure 4.13c, respectively. In each case the shaded area represents 5% of the area under the probability distribution curve. These are examples of a one-tailed significance test.

**Figure 4.13** Examples of a (a) two-tailed, and a (b, c) one-tailed, significance test of \( \bar{X} \) and \( \mu \). The normal distribution curves are drawn using the sample’s mean and standard deviation. For \( \alpha = 0.05 \), the blue areas account for 5% of the area under the curve. If the value of \( \mu \) is within the blue areas, then we reject the null hypothesis and accept the alternative hypothesis. We retain the null hypothesis if the value of \( \mu \) is within the unshaded area of the curve.
For a fixed confidence level, a two-tailed significance test is always a more conservative test because rejecting the null hypothesis requires a larger difference between the parameters we are comparing. In most situations we have no particular reason to expect that one parameter must be larger (or smaller) than the other parameter. This is the case, for example, in evaluating the accuracy of a new analytical method. A two-tailed significance test, therefore, is usually the appropriate choice.

We reserve a one-tailed significance test for a situation where we are specifically interested in whether one parameter is larger (or smaller) than the other parameter. For example, a one-tailed significance test is appropriate if we are evaluating a medication’s ability to lower blood glucose levels. In this case we are interested only in whether the result after administering the medication is less than the result before beginning treatment. If the patient’s blood glucose level is greater after administering the medication, then we know the answer—the medication did not work—without conducting a statistical analysis.

### 4.5.4 Errors in Significance Testing

Because a significance test relies on probability, its interpretation is naturally subject to error. In a significance test, $\alpha$ defines the probability of rejecting a null hypothesis that is true. When we conduct a significance test at $\alpha = 0.05$, there is a 5% probability that we will incorrectly reject the null hypothesis. This is known as a **type 1 error**, and its risk is always equivalent to $\alpha$. Type 1 errors in two-tailed and one-tailed significance tests correspond to the shaded areas under the probability distribution curves in Figure 4.13.

A second type of error occurs when we retain a null hypothesis even though it is false. This is known as a **type 2 error**, and its probability of occurrence is $\beta$. Unfortunately, in most cases we cannot calculate or estimate the value for $\beta$. The probability of a type 2 error, however, is inversely proportional to the probability of a type 1 error.

Minimizing a type 1 error by decreasing $\alpha$ increases the likelihood of a type 2 error. When we choose a value for $\alpha$ we are making a compromise between these two types of error. Most of the examples in this text use a 95% confidence level ($\alpha = 0.05$) because this provides a reasonable compromise between type 1 and type 2 errors. It is not unusual, however, to use more stringent (e.g. $\alpha = 0.01$) or more lenient (e.g. $\alpha = 0.10$) confidence levels.

### Contributors

- David Harvey (DePauw University)