Introduction to statistical inference

In this chapter, we introduce two techniques of statistical inference: hypothesis testing and computing confidence intervals. Following this chapter, we will frequently use these techniques to answer questions about linear models.

We begin the chapter by exploring two types of continuous distributions commonly used in statistical inference: normal and \( t \) distributions. After that, we show how to use such distributions to conduct hypothesis tests and to compute confidence intervals.

**Normal distributions**

We begin our study of continuous distributions with normal distributions, by far the most important continuous distributions in probability and statistics.

**Defining normal distributions**

We now define normal distributions precisely. We give the formula for a normal probability function for completeness, but for our purposes you do not need to know this formula. You should, however, pay particular attention to the many new terminology and notation introduced in the following definition.

**Definition: Normal Distribution.** A **normal distribution** is a continuous distribution whose probability density function is of the form

\[
N^{pdf}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]
for some real number $\mu$ and some positive real number $\sigma$. The real number $\mu$ is called the **mean** of the distribution, and the positive real number $\sigma$ is called the **standard deviation**. The mean $\mu$ and standard deviation $\sigma$ are referred to as the **parameters** of a normal distribution.

We denote the normal distribution with mean $\mu$ and standard deviation $\sigma$ by $N(\mu, \sigma)$. If a random variable $X$ has a normal distribution with mean $\mu$ and standard deviation $\sigma$, we write:

$$X \sim N(\mu, \sigma),$$

read “$X$ is normally distributed with mean $\mu$ and standard deviation $\sigma$”. A random variable $X$ is said to be **normally distributed** if the distribution of $X$ is a normal distribution.

One particular normal distribution is noteworthy enough to receive its own name.

**Definition:** **Standard Normal Distribution.** The **standard normal distribution** is defined to be $N(0, 1)$, the normal distribution with mean 0 and standard deviation 1.

The parameter names **mean** and **standard deviation** are not accidental. The random variable mean and random variable standard deviation of a normally distributed random variable actually do equal the correspondingly named parameters. This helps us understand what these parameters tell us about the distribution, and it also brings up another term.

**Definition.** The **variance** $\sigma^2$ of a normal distribution is defined to be the square of the standard deviation $\sigma$ of the distribution.

Since the random variable standard deviation of a normally distributed random variable equals its standard deviation parameter, the random variable variance of a normally distributed random variable also equals the variance parameter.

The standard deviation and the variance of a normal distribution carry exactly the same information, so using both as parameters would be redundant. In this text, we have chosen the standard deviation as the parameter, in keeping with most of the statistical computer packages. However, in the mathematical and theoretical statistics literature, the variance is usually taken to be the parameter instead. This discrepancy means that wherever normal distributions are discussed in the literature, you should be sure to know which alternative is being used. In this text, we will consistently use $\sigma$ as the parameter, not $\sigma^2$. 
Basic properties of normal distributions

The definition of a normal distribution is simple enough, but it begs a number of questions. The most immediate is a good first question to ask of any distribution: what does the graph of a normal distribution look like? A particular good way to answer this question is to search the internet for keywords such as "normal distribution" applet (with quotation marks as indicated) to find interactive demonstrations. Such demonstrations not only show graphs of individual normal distributions but also help develop a solid intuition for the roles played by the parameters.

Since a book is necessarily static though, we merely show the graphs of some individual normal distributions here to illustrate the basic properties of normal distributions.

The figure on page 39 depicts the graph of the normal distribution with mean $\mu$ and standard deviation $\sigma$. This picture shows a few of the fundamental properties of normal distributions:
Properties of normal distributions

1. The overall shape of a normal distribution is approximately that of a side view of a bell.
2. A normal distribution is symmetric about its mean $\mu$.
3. The high point of the graph of a normal distribution has $x$ coordinate equal to $\mu$.
4. The width of the bell shape of a normal distribution is related to the standard deviation $\sigma$.

As mentioned above, to see how the parameters $\mu$ and $\sigma$ relate to the overall shape of the normal distribution, we highly recommend using an interactive demonstration on a computer. However, since this is a book, we illustrate the role the parameters play with some pictures instead in the figure on page 41.

We now elaborate on the relationship between the standard deviation of a normal distribution and the width of its bell shape, as shown in the figure on page 41. The spread of a distribution can be described in terms of how much of its area lies close to its mean. For normal distributions, this is usually stated in terms of the fraction of the area that lies within 1, 2, and 3 standard deviations of the mean, which leads us to the following rule.
Changing the parameters of a normal distribution. Increasing the mean $\mu$ shifts the whole distribution to the right, and increasing the standard deviation $\sigma$ widens the distribution.
Figure. Graphical depiction of the 68-95-99.7 Rule. Since the total area of the distribution equals 1, multiplying each shaded area by 100 gives the percentage of the total area within each shaded region. This then gives the rule its name.
THE 68-95-99.7 RULE

As shown in the figure on page 42:

1. Approximately 68% of the area of a normal distribution lies within 1 standard deviation of the mean.

2. Approximately 95% of the area of a normal distribution lies within 2 standard deviations of the mean.

3. Approximately 99.7% of the area of a normal distribution lies within 3 standard deviations of the mean.

This rule concisely describes the relationship between the standard deviation $\sigma$ and the width of the bell shape of the normal distribution. This rule helps develop a good intuition for normal distributions, so you should memorize it.

Standardized units

From the definition of a normal distribution, we know that there are infinitely many normal distributions, one for each pair of parameter values $\mu$ and $\sigma$. However, these infinitely many normal distributions really only differ by a change of units. We can use the concept of standardized units to see this.

**Definition: Standardized units.** Let $X$ be a normally distributed random variable. A value is expressed in the standardized units for $X$ by giving the signed number of standard deviations that it is from the mean.

We use the term signed number to indicate that values smaller (larger) than the mean are a negative (positive) signed number of standard deviations from it.

The relationship between the original units and standardized units is shown in the figure on page 44, where the original units are shown on the top $x$ axis and standardized units on the bottom $x$-axis. Notice how much more relevant the standardized units are to the curve being depicted, compared to the original units.
Figure. A normal distribution with mean 13 and standard deviation 5. The upper $x$ axis uses the original units of the random variable, while the lower $x$ axis is in standardized units.
If a normally distributed random variable \( X \sim N(\mu, \sigma) \) is re-expressed in its standardized units, then the resulting random variable has a standard normal distribution \( N(0, 1) \). This follows since the mean of \( X \) is 0 standardized units from the mean, and the mean plus the standard deviation of \( X \) is 1 standardized unit from the mean.

More generally, the formula in the following definition shows how to convert a random variable \( X \) expressed in arbitrary units to a random variable \( Z \) expressed in standardized units.

**Definition: Standardizing Random Variables.** If \( X \sim N(\mu, \sigma) \), then the random variable \( Z = \frac{X - \mu}{\sigma} \) is called the standardized version of \( X \). This random variable \( Z \) gives the values of \( X \) expressed in standardized units. For any value \( x \) assumed by \( X \), the corresponding value \( z = \frac{x - \mu}{\sigma} \) assumed by \( Z \) is called the z-score of \( x \).

We won’t give a proof here that formula above expresses values of \( X \) in standardized units, but we offer some justification. The quantity \( x - \mu \) in the numerator tells us how far (and in which direction, positive or negative) \( x \) is from the mean \( \mu \), still measured in the original units. Dividing by the standard deviation \( \sigma \) then gives us the signed number of standard deviations \( x \) is from \( \mu \).

Notice that because the units for \( \sigma \) are the original units of the random variable, those original units cancel out of the expression for \( z \). Therefore standardized random variables are unitless.

Standardized units are really the right units to use when working with normally distributed random variables. Knowing a value in non-standardized units tells us almost nothing of interest about how where that value falls in the distribution. By the 68-95-99.7 Rule though, knowing a value in standardized units gives us an immediate idea of the probability of observing values closer to (or further from) the mean. This sort of probability estimate turns out to be very handy in statistics.
Computing areas under graphs of normal distributions

Since probabilities associated with normally distributed random variables correspond to areas under normal curves, we frequently need to compute such areas. There is no simple closed formula for these areas, so we use the computer to approximate them numerically.

Statistical computer packages do not have built-in functions specifically to compute the area under a normal curve between any two values. Instead, almost all statistical computer packages include normal cumulative distribution functions to compute areas under normal curves. We will generally be using standardized units, so the cumulative distribution function of the standard normal distribution, for which we now introduce some notation, is of primary interest to us.

**Definition: Standard Normal Cumulative Distribution Function.** We denote the cumulative distribution function of a standard normally distributed random variable by \( N_{\text{cdf}}(z) \), where \( z \) can be any real number.

The standard normal cumulative distribution function can be used to calculate all of the commonly needed areas under normal curves, as we now show. We focus on standard normal distributions here because any normal distribution can be transformed into a standard normal distribution by a change of units.

As pictured in the figure on page 47, there are four main types of areas to compute when we work with standard normally distributed random variables:

1. The area to the left of a given value \( z \), known as the left tail of \( z \) (or with endpoint \( z \)).
2. The area to the right of a given value \( z \), known as the right tail of \( z \) (or with endpoint \( z \)).
3. The area between two given values \( z_{\text{low}} \) and \( z_{\text{high}} \) (where \( z_{\text{low}} < z_{\text{high}} \)), with no particular name.
4. The area to the left of a given value \(-z\) together with the area to the right of \( z \) (where \( z > 0 \)), known as symmetric tails of \( z \) (or of \(-z\)).

We now show how to compute each of these in terms of the standard normal cumulative distribution function \( N_{\text{cdf}}(z) \).
**Figure.** The four main types of areas under normal curves to be computed. Upper left: A left tail. Upper right: A right tail. Lower left: The area between two values (which need not be negatives of each other, although they are here). Lower right: Symmetric tails.
Chapter: Introduction to statistical inference

As the upper left picture in the figure on page 47 indicates, the area of the left tail of a value $z$ is given directly by the cumulative distribution function:

\[
\text{area of the left tail of } z = N_{\text{cdf}}(z).
\]

As the upper right picture in the figure on page 47 shows, the area of a right tail of $z$ is simply the area under the entire curve (which equals 1) minus the area of the left tail of $z$:

\[
\text{area of the right tail of } z = 1 - N_{\text{cdf}}(z).
\]

As depicted in the lower left picture in the figure on page 47, the area under the standard normal curve between $z_{\text{low}}$ and $z_{\text{high}}$ can be computed as the area of the left tail of $z_{\text{high}}$ minus the area of the left tail of $z_{\text{low}}$.

\[
\text{area between } z_{\text{low}} \text{ and } z_{\text{high}} = N_{\text{cdf}}(z_{\text{high}}) - N_{\text{cdf}}(z_{\text{low}}).
\]

As in the lower right picture in the figure on page 47, the area of the symmetric tails of a value $z$ can be expressed either as twice the area of the left tail of $z$ or as twice the area of the right tail of $z$:

\[
\text{area of symmetric tails of } z = \text{area to the left of } -z \text{ plus to the right of } z
= N_{\text{cdf}}(-z) + (1 - N_{\text{cdf}}(z))
= 2N_{\text{cdf}}(-z) \text{ or } 2(1 - N_{\text{cdf}}(z)).
\]

In the later chapters of this text, we will compute such areas by the appropriate method from above without further explanation, so you should become comfortable with these methods now.

Quantiles of normal distributions

Now that we know how to compute the area of a left tail of a standard normal curve given its endpoint, we consider the reverse problem: given an area (between 0 and 1), find the endpoint of the left tail having that area. The solution to this problem is given by the quantile function. We now introduce our notation for the standard normal quantile function.

**Definition:** Standard normal quantile function. The quantile function of a standard normally distributed random variable is denoted by $N_{\text{qf}}(p)$, where $0 < p < 1$. 

As mentioned earlier, the standard normal cumulative distribution function \( N_{\text{cdf}}(z) \) and the standard normal quantile function \( N_{\text{qf}}(p) \) are inverses of each other in the sense that

\[
N_{\text{cdf}}(z) = p \quad \text{is the same as} \quad N_{\text{qf}}(p) = z.
\]

As you might imagine, computing values of the standard normal quantile function by hand is not feasible. However, almost all statistical computer packages have the standard normal quantile function built in.

While the standard normal quantile function gives endpoints of left tails, it can also be used to find endpoints of a slightly different kind of region, as we now describe.

**Definition: Standard Normal Central Quantiles.** Let \( c \) be a real number \( c \) with \( 0 < c < 1 \). The standard normal central \( c \)-quantile interval is the interval symmetric about 0 above which the area under a standard normal curve equals \( c \). The endpoints of the standard normal central \( c \)-quantile interval are denoted by \(-z^*(c)\) and \(z^*(c)\), and the nonnegative endpoint \( z^*(c) \) is called the standard normal central \( c \)-quantile.

The figure on page 50 illustrates the standard normal central quantile interval of \( c = 0.75 \), including its right endpoint \( z^*(0.75) \), the standard normal central 0.75-quantile.

Standard normal central quantiles are used commonly in statistical inference, so we should know how to compute them. Most statistical computer packages do not have built-in functions to compute them directly, but they can be computed readily in terms of the standard normal quantile function. We now explain two ways to do so.

First, the lower endpoint of the standard normal central quantile interval of \( c \) has the property that its left tail has area \((1 - c)/2\), as indicated in the figure on page 50. This means that the lower endpoint \(-z^*(c)\) of the central quantile interval of \( c \) is given by:

\[
-z^*(c) = N_{\text{qf}} \left( \frac{(1 - c)}{2} \right).
\]

Taking the negative of both sides gives us that

\[
z^*(c) = -N_{\text{qf}} \left( \frac{(1 - c)}{2} \right),
\]

\(^{10}\)These are not standard terms but are merely convenient for this text.
Figure. How the central quantile value $z^*(c)$ (for $c = 0.75$ in this case) is determined.

Figure. Computing a central quantile value using the quantile function.
which is the first way to compute $z^*(c)$ in terms of the standard normal quantile function.

The second way to compute $z^*(c)$ is to note that the left tail of the upper endpoint $z^*(c)$ has area

$$\frac{1 - c}{2} + c = \frac{1 + c}{2}.$$ 

This means that the upper endpoint $z^*(c)$ is given by:

$$z^*(c) = N_{\text{qf}} \left( \frac{(1 + c)/2}{2} \right),$$

which gives us the second way to compute $z^*(c)$ in terms of the standard normal quantile function. To speed mental calculation, note that $(1 + c)/2$ is simply halfway between $c$ and 1.

In summary, to compute a standard normal central quantile using the quantile function, we use either of the following two formulas.

**Formulas for Computing Standard Normal Central Quantiles**

If $N_{\text{qf}} (p)$ is the quantile function for the standard normal distribution and $0 < c < 1$, then the standard normal central $c$-quantile $z^*(c)$ can be computed by either of the following:

$$z^*(c) = -N_{\text{qf}} \left( (1 - c)/2 \right) = N_{\text{qf}} \left( (1 + c)/2 \right).$$

Now that we have sufficient familiarity with normal distributions for the present, we proceed to explore $t$ distributions.

**$t$ distributions**

Although they pale in comparison to normal distributions, $t$ distributions comprise the second most commonly encountered type of continuous distribution in statistics. Since $t$ distributions are similar in many ways to normal distributions, our introduction to $t$ distributions here will roughly parallel our introduction to normal distributions.
Defining \( t \) distributions

We begin with the definition of a \( t \) distribution. As with normal distributions, the formula for the probability density function is not important to know at this point; the basic terminology and properties of \( t \) distributions are more important for our purposes here.

**Definition: \( t \) Distribution.** A \( t \) distribution is a continuous distribution whose probability density function \( f(x) \) is given by

\[
f(x) = C \left(1 + \frac{x^2}{d}\right)^{-\frac{(d+1)/2}{d}}.
\]

for some positive real number \( d \). The constant \( C \) depends only on \( d \) and is included simply to ensure that the total area of the distribution equals 1.\(^{11}\)

The positive real number \( d \) is called the (number of) **degrees of freedom** of the distribution. The number of degrees of freedom \( d \) is also referred to as the **parameter** of a \( t \) distribution.

To denote a random variable \( X \) has a \( t \) distribution with \( d \) degrees of freedom, we write:

\[X \sim t_d.\]

A random variable \( X \) is said to be **\( t \) distributed** if the distribution of \( X \) is a \( t \) distribution.

Note that \( t \) distributions have only one parameter, while normal distributions have two. There is no parameter in a \( t \) distribution to shift the center of the distribution, as there is with normal distributions (the mean).\(^{12}\)

The random variable mean and standard deviation are common measures of center and spread for normal distributions, and even coincide with the parameters of normal distributions. However, they are rarely used for \( t \) distributions.\(^{13}\) The random variable median of any \( t \) distribution is 0, as is the random variable mean when \( d > 1 \). The random variable

\(^{11}\)We omit the complicated formula for \( C \) here; suffice it to say that \( C \) is not a parameter but rather a positive real number depending only on \( d \).

\(^{12}\)There is actually a generalization of a \( t \) distribution called a **non-central \( t \) distribution** that has a second parameter to shift the center, but we will not use it in this text. We refer the interested reader to a more advanced text on mathematical statistics for a discussion of such distributions.

\(^{13}\)Except perhaps to note that a \( t \) distributed random variable with 1 degree of freedom has no random variable mean, since the defining integral diverges.
standard deviation can be computed but is not a particularly informative quantity for $t$ distributions.

**Basic properties of $t$ distributions**

As with normal distributions, we begin our study of $t$ distributions by graphing their probability density functions. Again we highly recommend searching the internet for keywords such as "$t$ distribution" applet (with quotation marks as indicated) to find interactive demonstrations of $t$ distributions. But since a book cannot be interactive the way that a website can, we simply give some pictures here to illustrate the basic properties of $t$ distributions.

The figure on page 53 depicts the graph of several $t$ distributions with different numbers of degrees of freedom. This picture shows some of the basic properties of $t$ distributions:
Chapter: Introduction to statistical inference

Properties of $t$ distributions

1. The overall shape of a $t$ distribution is very similar to that of a standard normal distribution. However, $t$ distributions have thicker tails and are shorter near 0.

2. A $t$ distribution is symmetric about 0, which is its highest point.

3. The higher the number of degrees of freedom, the closer a $t$ distribution is to a standard normal distribution.

It is important to note that the 68-95-99.7 rule applies only to normal distributions and not to $t$ distributions (or any other distributions). However, because $t$ distributions have a shape similar to the standard normal distribution, the 68-95-99.7 rule can be used to give a ballpark intuition for probabilities related to $t$ distributions. For example, while it is not true that 95% of the area under a $t$ distribution lies between $-2$ and 2, we should expect somewhere around that percentage, with the approximation being better for higher numbers of degrees of freedom. Checking this on the computer for a couple of values, we find that with $d = 1$, the approximation isn’t very good: the actual percentage of the total area between $-2$ and 2 is about 70%. However, with $d = 20$, the actual percentage is about 94%, so a ballpark estimate based on the standard normal distribution is pretty good.

Because $t$ distributions are all centered at 0 and because their random variable standard deviations are not particularly informative, $t$ distributions don’t have standardized units as normal distributions do.

Computing with $t$ distributions

We compute with $t$ distributions almost exactly as with normal distributions. For example, we compute areas under the graphs of $t$ distributions with $t$ cumulative distribution functions.

**Definition: $t$ cumulative distribution functions.** We denote the cumulative distribution function of a random variable with a $t_d$ distribu-
tion by \( t_{df}^c (x) \), where \( x \) can be any real number.

As with standard normal cumulative distribution functions, \( t \) cumulative distribution functions can be used to calculate all of the commonly needed areas under \( t \) distributions: left tails, right tails, areas between two values, and symmetric tails. The methods are exactly the same as with standard normal distributions, except that we replace the standard normal cumulative distribution function with the \( t \) cumulative distribution function with the appropriate number of degrees of freedom.

Quantile functions of \( t \) distributions work just as quantile functions for standard normal distributions do.

**Definition: \( t \) Quantile Functions.** The quantile function random variable with a \( t_d \) distribution is denoted by \( t_{df}^q (p) \), where \( 0 < p < 1 \).

As with standard normal distributions, \( t \) cumulative distribution functions and \( t \) quantile functions with the same number of degrees of freedom are inverses of each other:

\[
 t_{df}^c (x) = p \quad \text{is the same as} \quad t_{df}^q (p) = x.
\]

We define and compute \( t \) central quantiles just as you might expect from standard normal central quantiles.

**Definition: \( t \) Central Quantiles.** Let \( c \) be a real number \( 0 < c < 1 \). The \( t_d \) central \( c \)-quantile interval is the interval symmetric about 0 above which the area under a \( t_d \) distribution equals \( c \). The endpoints of the \( t_d \) central \( c \)-quantile interval are denoted by \( -t_d^c (c) \) and \( t_d^c (c) \), and the nonnegative endpoint \( t_d^c (c) \) is called the \( t_d \) central \( c \)-quantile.

We compute \( t_d \) central quantiles exactly as their standard normal counterparts, except that we substitute \( t_{df}^q (p) \) in for \( N_{df}^q (p) \). \( t_d \) central quantiles pose no additional complications.

**Hypothesis tests**

Hypothesis testing is a broad subject, and we will only scratch its surface in this text. We begin our study here by describing how hypothesis tests work in general, after which we explore a particular type of test, namely \( t \) tests.

\[\text{Footnote 14: As with standard normal distributions, these are not standard terms.}\]
The idea of a hypothesis test

To avoid getting lost in the technical details of individual hypothesis tests, it is worthwhile to have a general understanding of what hypothesis tests are and how they work.

A hypothesis test is a method of seeking evidence against a particular hypothesis. To conduct a hypothesis test, we must formulate a both null hypothesis that we seek evidence against and an alternative hypothesis that we use to determine what form such evidence must take. When we conduct hypothesis tests, we often compute the probability of an event if the null hypothesis is true. We refer to this as the event’s probability under the null hypothesis. Similarly, we refer to the probability of an event if the alternative hypothesis is true as the event’s probability under the alternative hypothesis.

The result of a hypothesis test is called a p-value, which answers the question:

**Question that the p-value answers**

What is the probability that an arbitrary sample would be at least as extreme as our sample?

The term extreme here means, roughly speaking, unusual under the null hypothesis and commonplace under the alternative hypothesis. The precise meaning of the term is therefore determined by the specific null and alternative hypotheses. The term sample here refers to a sample that has the same size as our sample.

A low p-value (near 0) says that the probability of obtaining a sample at least as extreme as ours under the null hypothesis is low. In other words, our sample is pretty extreme if the null hypothesis is true but commonplace if the alternative hypothesis is true. This provides evidence against the null hypothesis.

On the other hand, a high p-value (near 1) tells us that the probability of obtaining a sample at least as extreme as ours under the null hypothesis is high. This does not provide evidence against the null hypothesis, since it effectively states that our sample is commonplace if the null hypothesis is true.
Notice that in neither case do we prove anything. Instead, the hypothesis test gives us a way to quantify the strength of the evidence against the null hypothesis. The individual who receives the results must understand how this quantification works in order to decide how to use the results.

We illustrate this process with an example. This example is somewhat different from the other hypothesis tests that we will conduct in this text, but it shows the overall process clearly.

**Example** Suppose we have a coin and a particular repeatable way of flipping the coin. We would like to know if this way of flipping the coin is fair, in the sense that the probability that of heads is 0.5.

To test this, we conduct a hypothesis test with null hypothesis that the flipping method is fair and alternative hypothesis that the flipping method is not fair. We collect a sample of 100 flips. Without going into the details, we simply state that the term extreme (in the context of the p-value) means far from 50 heads, which would be the mean number of heads if the coin is fair.

Suppose there were only 3 heads out of the 100 flips. Then the phrase at least as extreme as 3 would mean: 0, 1, 2, 3, 97, 98, 99, or 100 heads out of the 100 flips. The p-value would be the probability of obtaining one of these values. It would be very small, which would be strong evidence that this method of flipping the coin was not fair.

On the other hand, suppose there were only 57 heads out of the 100 flips. Then the phrase at least as extreme as 57 would mean: 0-43 or 57-100 heads out of the 100 flips. The p-value would be the probability of obtaining one of these values. It would be very large, which would indicate a lack of evidence that this method of flipping the coin was not fair.

Note the double negatives that occur here; these are common in hypothesis testing because we are looking for evidence against the null hypothesis. If we don’t find such evidence, we will have a double negative: a lack of evidence against the null hypothesis. Do not confuse this double negative with a positive. A lack of evidence against the null hypothesis is not the same thing as evidence for the null hypothesis.

For example, suppose we had flipped the coin only 3 times instead of 100 times above. Then no matter how many heads occurred we wouldn’t find strong evidence that the flipping method was unfair because the test wasn’t powerful enough. Flipping the coin more times might reveal such evidence, and we should not interpret our lack of evidence of unfairness as evidence of fairness. So for hypothesis tests in general, we will never find evidence for the null hypothesis. We either find evidence against the
null hypothesis or we don’t.

This raises a serious difficulty with hypothesis testing. It doesn’t give us a definitive answer, but we often need to act based on evidence we obtain. In cases where we require a yes/no answer to the question of whether we found evidence against the null hypothesis, we use what is called a significance test.

The idea of a significance test

To conduct a significance test, we first choose a significance level, which is a $p$-value cutoff level. We then conduct a hypothesis test. If the resulting $p$-value is less than or equal to the significance level, we have statistically significant evidence against the null hypothesis. If the $p$-value is greater than the significance level, we do not have statistically significant evidence against the null hypothesis.

The phrase statistically significant is technical terminology with a precise definition.

**Definition.** The result of a hypothesis test is called statistically significant if its $p$-value is less than or equal to the pre-selected significance level. If the null hypothesis is that a particular parameter equals zero, we say that our estimate of that parameter is statistically significant to mean that the result of the hypothesis test is statistically significant.

So why would we ever conduct a significance test instead of stopping with the $p$-value of a hypothesis test? After all, a significance test is less informative. However, we generally conduct a significance test when we need to choose a further course of action based on whether or not we think the null hypothesis is true. In this case, we set a cutoff for the $p$-value before conducting the test so that we won’t be tempted to determine our course of action based on what you prefer rather than what the data indicates.

The traditional value for the significance level is 0.05. This goes back to Fisher in the middle of the 20th century, but his suggestion of this value was not intended to carry the rigidity that it does today. The traditional significance level of 0.05 has become so engrained that almost everybody uses it by default; if we depart from it, we should explain why we are doing so, since people will be wondering why.

Personally, I find it difficult to defend the practice of significance test-
ing since hypothesis testing alone carries much more information. However, significance testing is so prevalent in statistics and scientific research that we must not ignore it here.

Part of the problem with a significance test is that many non-statisticians misunderstand its results. There are only two possible outcomes of a significance test: we did find statistically significant evidence or we didn’t. People often misinterpret these two outcomes as: we proved that the alternative hypothesis is true or we proved that the null hypothesis is true. This gross misinterpretation is fraught with errors and is often used to deceive people, intentionally or unintentionally. So when you conduct or encounter significance tests, be sure that you understand the two possible outcomes and their implications, and do not ignore the p-value. The extra information that it conveys can help you analyze the strength of your results beyond the two possible outcomes of the significance test.

**Interpreting statistical significance**

Three things are important to keep in mind when we interpret statistical significance:

1. Statistical significance is not the same as practical significance.
2. Statistical significance is not proof of the alternative hypothesis.
3. Lack of statistical significance is not proof of the null hypothesis.

For the first, we must remember that statistically significant evidence against the null hypothesis suggests that the null hypothesis may not be true. But even if it isn’t, this may have no practical implications. Consider the fair coin flipping example above. If the probability that the coin lands heads is 0.500000000001, then the coin isn’t fair. If we collect enough data (zillions of flips no doubt), then we should be able to find statistically significant evidence against the null hypothesis that the coin is fair. However, such a tiny difference is of no practical consequence for anything that we are likely to use such a coin flip for. In short, just because we detect evidence of something (statistical significance) doesn’t mean that that something makes any difference in a real-world setting (practical significance), even if it is true.

For the second, suppose for simplicity that we use the traditional significance level of 0.05. This means that if the null hypothesis is true, we
will flag the most extreme 5% of the samples as statistically significant. So if the null hypothesis is true, the probability that we will find statistically significant evidence against it anyway is 0.05 or 1/20. This has the important implication that we expect about 1 in every 20 hypothesis tests of a true null hypothesis to result in statistically significant evidence against the null hypothesis! In short, statistically significant evidence is decidedly not proof.

For the third, we have already discussed it above when we introduced the \( p \)-value.

\( t \) tests

We now introduce a particular type of hypothesis test that we will use throughout this text, the \( t \) test.

To conduct a \( t \) test, we need three things:

1. A \textit{null hypothesis}, which can be any hypothesis as long as the last two conditions are fulfilled.

2. A \textit{test statistic}, which is a random variable whose distribution under the null hypothesis is a \( t \) distribution.

3. An \textit{alternative hypothesis}, which specifies or implies which regions of the test statistic's distribution under the null hypothesis are considered \textit{extreme}.

4. An \textit{observation} of the test statistic.

With these three elements, it is simple to conduct a \( t \) test.

\textbf{Definition: T Test.} Suppose that we have a null hypothesis, test statistic, alternative hypothesis, and observation as described above. To conduct a \( t \) test, we compute the probability of obtaining a value at least as extreme as our observation under the null hypothesis. This probability is the \( p \)-value. To conduct a \( t \) \textit{significance test} instead, we decide on a significance level, conduct a \( t \) test, and then compare the \( p \)-value to the significance level.

We illustrate this definition with an example, one that serves as a prototype for all \( t \) tests.
EXAMPLE Suppose we have a null hypothesis and a test statistic whose distribution under the null hypothesis is $t_{10}$, and suppose that we have an alternative hypothesis which implies that extreme means far from 0. This means that for any real number $x$, the symmetric tails at $x$ form the set of points that are at least as extreme as $x$.

Also suppose that we have observed the test statistic and obtained the value $-1.2$. To conduct a $t$ test, we compute the area of the symmetric tails at $-1.2$ for $t_{10}$. This yields 0.258, which is therefore the $p$-value of this $t$ test.

If we had instead wanted to conduct a significance test with significance level 0.05, we conduct the above $t$ test and compare the resulting $p$-value to 0.05. Since $0.258 > 0.05$, we conclude that we do not have statistically significant evidence against the null hypothesis. ♦

Most $t$ tests are very similar to the one in this example, although sometimes the definition of extreme is different. For example, the left tail or the right tail alone might form the set of points that are at least as extreme as a value.

Now that we have some familiarity with hypothesis testing, we proceed to another type of statistical inference: computing confidence intervals.

Confidence intervals

While hypothesis tests occur in a wide variety of statistical settings, confidence intervals are specific to parameter estimation. As with hypothesis tests, it is worthwhile to have an understanding of the general process of computing a confidence interval before dealing with some of the technical details.

The idea of a confidence interval

Confidence intervals allow us to estimate a parameter with a range of values rather than with a single value. This is important because a range of values includes information about the variability of our estimates, whereas a single value does not.

To compute a confidence interval, we first need to decide which parameter we are estimating. To avoid getting into a technical description of what exactly constitutes a parameter, for the present we will think of a parameter as an unknown quantity relating to a family of distributions. The
prototypical example of a parameter is the mean of a normally distributed random variable. The standard deviation of a normally distributed random variable is also a parameter, although we will compute confidence intervals for means more often than for standard deviations.

Once we have settled on a parameter to estimate, we choose a confidence level, a term that we will define in a moment. We then choose a confidence interval algorithm to produce a confidence interval at the chosen confidence level.

**Definition: Confidence interval algorithm.** Let $c$ be a real number with $0 < c < 1$, and suppose that we have a parameter that we would like to estimate based on a sample of size $n$. For this sample size, a $c$ confidence interval algorithm is an algorithm with the following properties:

1. Its input is a sample of size $n$.
2. Its output is an interval of real numbers called a $c$ confidence interval.
3. For a sample selected at random, the algorithm has probability $c$ of producing a confidence interval that contains the true parameter value being estimated.

In this context, the real number $c$ is called the **confidence level** of the confidence interval.

For any given parameter estimation and confidence level, there are many different confidence interval algorithms and correspondingly many different confidence intervals. We will focus our attention on standard, widely-used confidence interval algorithms here, but you should be aware that the confidence intervals we produce are not the only possible confidence intervals. Consequently, we always refer to a confidence interval at confidence level $c$ rather than *the* confidence interval at confidence level $c$.

The overall process of computing a confidence interval works as follows. We choose a parameter to estimate, a confidence level, a sample size, and a confidence interval algorithm. We then collect a sample of the appropriate size and use as input for the confidence interval algorithm, which then produces the confidence interval. The algorithm has probability equal to the confidence interval of producing a confidence interval that contains the true value of the parameter.
Now that we know how to compute a confidence interval in general, we describe how to compute a particular type of confidence interval.

**t confidence intervals**

One particularly common type of confidence interval is a *t* confidence interval, so named for the distribution used in computing it. It arises as follows. We have a normally distributed random variable $X$ with unknown mean $\mu(X)$ and standard deviation $\sigma(X)$:

$$X \sim N(\mu(X), \sigma(X)).$$

We also have a sample $x$ of size $n$ of this random variable.

Because $X$ has a normal distribution, the standardized version

$$Z = \frac{X - \mu(X)}{\sigma(X)}$$

has a standard normal distribution. However, since we don’t know $\sigma(X)$, we can’t use this standardization to help us estimate the other parameter $\mu(X)$. If we approximate $\sigma(X)$ by the sample standard deviation $s(X)$ though, we arrive at

$$T = \frac{X - \mu(X)}{s(X)}.$$ 

If we knew the distribution of this, then we could use it to help us estimate $\mu(X)$. Fortunately, there is a key mathematical theorem that tells us the distribution. Before we state it though, we must introduce an important preliminary definition.

**Definition: Standard error of the sample mean.** Suppose that $X \sim N(\mu(X), \sigma(X))$, and let $n$ be a positive integer. Let $\bar{X}$ denote the random variable whose random process is drawing a sample of $X$ of size $n$ and whose value for any such sample $x$ is $\bar{x}$. Then the random variable

$$se(\bar{X}) := \frac{s(X)}{\sqrt{n}}$$

gives an approximation of the standard deviation of $\bar{X}$ and is called the **standard error of $\bar{X}$**.
We now give the key theorem that allows us to compute $t$ confidence intervals.

**Theorem: $t$ Distributions.** Let $X$ be normally distributed with mean $\mu(X)$ (and unknown standard deviation), and let $x$ be a size $n$ sample of $X$. Then the random variable

$$T = \frac{\bar{X} - \mu(X)}{se(\bar{X})}$$

has a $t$ distribution. The number of degrees of freedom of this distribution equals $n$ minus the number $p$ of parameters that $\mu(X)$ depends on:

$$T \sim t_{n-p}.$$

In the simplest case, there are no restrictions on the mean, so it depends on only a single parameter (itself) and $p = 1$. However, when we explore linear models, we will encounter many cases in which the mean depends on a greater number of parameters.

With this theorem, we can now explain the traditional $t$ confidence interval algorithm. This algorithm produces a confidence interval from the values of:

$$\bar{X} - se(\bar{X}) t^*_{n-p}(c) \text{ to } \bar{X} + se(\bar{X}) t^*_{n-p}(c),$$

where $t^*_{n-p}(c)$ is the central $c$-quantile for a $t_{n-p}$ distribution. To memorize the interval produced by this algorithm, we often shorten it to:

$$\bar{x} \pm se t^*,$$

supplying the sample $x$, degrees of freedom $n - p$, and confidence level $c$ from the context as appropriate.

So why does this algorithm have a 95% probability of producing a confidence interval that contains the true mean $\mu(X)$? Well, no matter what $\mu(X)$ is, the theorem on page 64 tells us that the distribution of $T$ is $t_{n-p}$. By the definition of a central quantile interval, the probability is $c$ that $T$ will fall in the central quantile interval at $t^*_{n-p}(c)$:

$$\text{Prob} \left( -t^*_{n-p}(c) \leq T \leq t^*_{n-p}(c) \right) = c.$$

Using the definition of $T$, we can expand the expression inside the probability parentheses to:

$$-t^*_{n-p}(c) \leq \frac{\bar{X} - \mu(X)}{se(\bar{X})} \leq t^*_{n-p}(c).$$
The denominator $s(x)$ is nonnegative, so we can multiply through by it without changing the inequalities. After that, we can subtract $X$ through the inequalities and then multiply through by $-1$, which reverses the inequalities. From this, we arrive at:

$$\bar{X} - \text{se}(\bar{X})t^*_{n-p}(c) \leq \mu(X) \leq \bar{X} + \text{se}(\bar{X})t^*_{n-p}(c).$$

Since this is the same as the original expression inside the probability parentheses, we can substitute it back in:

$$\text{Prob}\left(\bar{X} - \text{se}(\bar{X})t^*_{n-p}(c) \leq \mu(X) \leq \bar{X} + \text{se}(\bar{X})t^*_{n-p}(c)\right) = c.$$  

Therefore no matter what the true value of $\mu(X)$ is, the probability is $c$ that it is contained in the confidence interval given by the traditional $t$ confidence interval algorithm:

$$\bar{X} - \text{se}(\bar{X}) t^*_{n-p}(c) \text{ to } \bar{X} + \text{se}(\bar{X}) t^*_{n-p}(c).$$

In short, we have shown that the algorithm works as advertised. We now give an example to illustrate how to compute $t$ confidence intervals.

**Example** Let $X$ be a normally distributed random variable with unknown mean $\mu(X)$ and standard deviation $\sigma(X)$. Suppose that $x$ is a size 11 sample of $X$ with sample mean $\bar{x} = 2$ and sample standard deviation $s(x) = 3$. To compute a 95% confidence interval for $\mu(X)$, we use the traditional $t$ confidence interval algorithm. For this, we use a computer to find the standard error of $\bar{X}$ and the central 0.95-quantile of a $t_{10}$ distribution:

$$\text{se}(\bar{X}) = \frac{s(x)}{\sqrt{n}} = \frac{3}{\sqrt{11}} = 0.905 \quad \text{and} \quad t^*_{10}(0.95) = 2.228.$$  

We obtained the number of degrees of freedom as the difference between the sample size (11 here) and the number of parameters that the mean depends on (1, since it depends only on itself and no other parameters here).

These allow us to compute the desired confidence interval. It is from:

$$\bar{X} - \text{se}(\bar{X}) t^*_{n-p}(c) \text{ to } \bar{X} + \text{se}(\bar{X}) t^*_{n-p}(c)$$

which equals

$$2 - (0.905)(2.228) \text{ to } 2 + (0.905)(2.228),$$

which gives us a 95% confidence interval for $\mu(X)$ from $-0.015$ to $4.015$. ♦

We can use this example as a model for any $t$ confidence interval computation. The process is always the same; only the confidence level and the values of $\bar{x}, s(x), n,$ and $p$ change.
Problems

1. Suppose you have a test statistic that is standard normally distributed under the null hypothesis $H_{null}$, and that you compute the test statistic’s value to be $-2.43$ for your data.

   (a) Compute the 2-sided $p$-value of the data.
   (b) Compute the 1-sided to the right $p$-value of the data.
   (c) Compute the 1-sided to the left $p$-value of the data.
   (d) Interpret your result in terms of evidence about the null hypothesis, using the traditional significance level $\alpha = 0.05$.

2. Repeat Problem 1 only for a test statistic whose distribution under the null hypothesis $H_{null}$ is a $t$ distribution with 4 degrees of freedom. Continue to assume that the test statistic’s value is $-2.43$.

3. Suppose you have a test statistic that is standard normally distributed under the null hypothesis $H_{null}$, and that you compute the test statistic’s value to be $1.98$ for your data.

   (a) Compute the 2-sided $p$-value of the data.
   (b) Compute the 1-sided to the right $p$-value of the data.
   (c) Compute the 1-sided to the left $p$-value of the data.
   (d) Interpret your result in terms of evidence about the null hypothesis, using the traditional significance level $\alpha = 0.05$.

4. Repeat Problem 3 only for a test statistic whose distribution under the null hypothesis $H_{null}$ is a $t$ distribution with 11 degrees of freedom. Continue to assume that the test statistic’s value is $1.98$.

5. Suppose you have a test statistic $X$ that is normally distributed with unknown mean $\mu$ and standard deviation $\sigma = 0.74$, and that you have computed the value of $X$ to be $-2.43$ for your data.

   (a) Find a level 0.95 confidence interval for $\mu$.
   (b) Find a level 0.90 confidence interval for $\mu$. 


6. Repeat both parts of Problem 5, only now assume that $\sigma$ is not known, but that the sample standard deviation is $s = 0.74$. The number of observations in the sample is 12 and the number of parameters used in describing $\mu$ is 1, so the number of degrees of freedom for the problem equals $12 - 1 = 11$. Continue to assume that the test statistic’s value is $-2.43$. 