15.5.1 Practice Hypothesis Testing

Now that Jeremy feels comfortable characterizing data from his colleagues, he wants to start practicing some basic statistics. He remembers a bit from his time in college, but it's been such a long time that Jeremy figures it's good to familiarize himself with some of the fundamentals before he tries to implement any statistical tests in R.

One of the largest and most critical concepts in statistics is hypothesis testing. In data science, we use statistical hypothesis testing to determine the probability of an event (or set of observations) under particular assumptions. In other words, we use statistical hypothesis testing to determine which of our hypotheses are most likely to be true. There are two types of statistical hypothesis:

- The **null hypothesis** is also known as H_0 and is generally the hypothesis that can be explained by random chance.
- The **alternate hypothesis** is also known as H_a and is generally the hypothesis that is influenced by non-random events.

By the end of this section, we should be able to generate a set of hypotheses and interpret the outcome of a statistical test. These concepts are universal and will apply to any statistical test, dataset, or analytical result.

The Importance of Hypothesis Testing

Although data collection and research are important, the backbone of the scientific method is **hypothesis testing**. Hypotheses are utilized by the scientific method to help narrow the scope of research and testing as well as provide a clear outcome of our results. Without generating a set of hypotheses, it becomes exponentially more difficult to quantify results and provide measurable outcomes to our analyses. As data analysts, it's our job to match a set of hypotheses to an appropriate statistical test to ensure that results are interpreted correctly.

Hypothesis Testing in Five Steps

Regardless of the complexity of the dataset or the proposed question, hypothesis testing uses the same five steps:

1. Generate a null hypothesis, its corresponding alternate hypothesis, and the significance level.
2. Identify a statistical analysis to assess the truth of the null hypothesis.
3. Compute the p-value using statistical analysis.
4. Compare p-value to the significance level.
5. Reject (or fail to reject) the null hypothesis and generate the conclusion.

Keep in mind that the null and alternate hypotheses are used to explain one of two outcomes from our proposed question, and both are mutually exclusive and exhaustive. In other words, no matter what, one of these statements must be used to explain our analysis results.

For example, perhaps we wanted to solve the question: "Is flipping a specific coin fair and balanced?" Given this question, our null hypothesis could be that the likelihood of flipping heads is the same as flipping tails. In other words, the likelihood of heads or tails can be totally explained by random chance. Our alternative hypothesis might be that the likelihood of flipping heads is not the same as flipping tails. If we were to represent our hypotheses using mathematical symbols, it would be expressed as:

$H_0 : P_H = 0.5$

$H_a : P_H \neq 0.5$

Where P_H represents the probability of flipping heads.

IMPORTANT

> Notice that our null hypothesis represents the scenario that our results can be explained by random chance without any outside influence. In contrast, our alternate hypothesis represents any other scenario that our results could yield.

Once we established our null and alternate hypothesis, we would then need to identify a statistical analysis to assess if our null hypothesis is true. In this example, we could test our null hypothesis by flipping our own coin 50 times and then by calculating the probability of flipping heads.

IMPORTANT

> Depending on the complexity of the question and null hypothesis, calculating probability might be insufficient, and we might need to rely on more traditional statistical tests. Fret not, we'll cover many of the most popular statistical tests later in this module.

After determining which statistical analysis is most appropriate and analyzing our data, we must quantify our statistical results using probability. In our example, we are calculating probability directly; however, most statistical tests will produce probability in the form of a p-value.

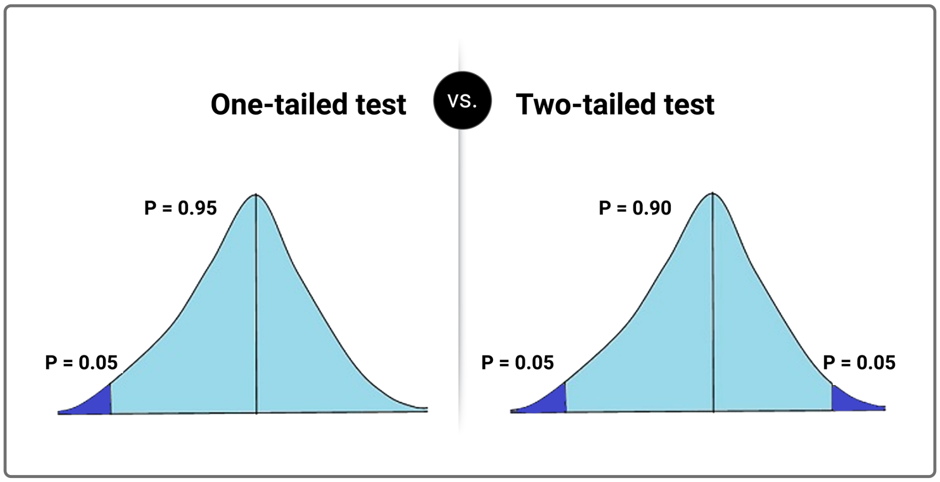
The **p-value**, or probability value, tells us the likelihood that we would see similar results if we tested our data again, if the null hypothesis is true. Therefore, we use the p-value to provide quantitative evidence as to which of our hypotheses are true.

To determine which hypothesis is most likely to be true, we compare the p-value against a significance level. A **significance level** (also denoted as alpha or α) is a predetermined cutoff for our hypothesis test. When designing our hypothesis, we would determine the significance level based on the importance of our findings.

In most cases, a significance level of 0.05 is sufficient, but if our hypotheses are being used for critical decision-making (such as the performance of a drug or the durability of a helmet), we might want to use smaller cutoffs such as 0.01 or 0.001. Regardless of what significance level we select, we want to predetermine our cutoff prior to computing the p-value as to not introduce bias into our results. Refer to the following chart:

Importance of Findings	Significance Level	Probability of Being Wrong
Low	0.1	1 in 10
Normal	0.05	5 in 100
High	0.01	1 in 100
Very high	0.001	1 in 1,000
Extreme	0.0001	1 in 10,000

In addition to determining a significance level based on the importance of our findings, we must ensure our hypotheses and statistical tests are either one-tailed or two-tailed. The tails of our hypotheses or statistical tests are referring to the distribution of measurements or observations used in the analysis.



When it comes to hypothesis testing, a **one-tailed hypothesis** is only describing one side of the distribution. One-tailed hypotheses use descriptions such as "x is greater than y" or "x is less than or equal to y." Alternatively, **two-tailed hypotheses** describe both sides of the distribution and use descriptions such as "equal to" or "not equal to."

When it comes to checking our statistical tests, the documentation will tell us if our statistical test is one-tailed or two-tailed. Once we have determined the number of tails considered for both our hypotheses and statistical test, we can determine if we need to adjust our p-value:

- If our hypotheses and statistical test are both two-tailed, use the statistical test p-value as is.

- If our hypotheses are one-tailed, but our statistical test is two-tailed, divide the statistical test p-value by 2.

Once we have determined the significance level and the calculated p-value, we can complete our statistical analysis. If our calculated p-value is smaller than our significance level, we would state that there is sufficient statistical evidence that our null hypothesis is not true, and therefore we would reject our null hypothesis. Alternatively, if our calculated p-value is larger than our significance level, we would state that we do not have sufficient evidence to reject our null hypothesis, and therefore we fail to reject our null hypothesis.

NOTE

> Outside of statistics, you may see others refer to this process as "accepting or rejecting" the null hypothesis. While this is not entirely untrue, many statisticians believe that there is no way of making a definitive choice between either hypothesis without an infinitely large dataset. Therefore, we use p-values and significance levels to determine *the probability* that our observations were obtained assuming the null hypothesis.

After we have determined which hypothesis is most likely to be true, we must conclude our statistical findings by relating our results back to the initial dataset or proposed question.