7.1: Introduction to Hypothesis Testing

Learning Objectives

- Introduce the idea of hypothesis testing
- Define null and alternative hypotheses
- Develop the logic of identifying null and alternative hypotheses
- Define the p-value
- Explain the two possible conclusions for a hypothesis test
- Introduce the α value
- Introduce type I and type II errors
- Differentiate between statistically significant and practically significant results
- Introduce one-tailed and two-tailed tests

Review and Preview

In the last chapter, we finally achieved a goal of inferential statistics: to use facts about sample data to speak confidently about the facts of the population from which it was drawn. Our method of confidence intervals provides an interval estimate of the population parameter at a certain success rate called the confidence level. When we do not know much about the population, we can utilize random sampling to build confidence intervals to learn about populations from scratch. At other times, we have claims about a certain population that we hope to test. This quest falls within the realm of inferential statistics and is the subject of this chapter.

Consider the legendary secret agent James Bond 007, the central character of a series of books and action movies, who is capable of great feats of heroism and has a penchant for martinis: shaken not stirred, never stirred. With such a strong preference, one would expect that James Bond could actually tell the difference between shaken and stirred martinis simply by taste and know which was which. Such an ability may seem unlikely. It would be natural to desire some evidence to back the claim. Taste-testing martinis could serve as a method for collecting such evidence. Suppose we settled on conducting 16 taste tests where each martini was either stirred or shaken randomly. Upon tasting each martini, James Bond claimed the martini was either shaken or stirred. After tasting all 16 martinis, suppose James Bond correctly identified 13 of the martinis. Is this sufficient evidence to back the claim that he can distinguish between shaken and stirred martinis simply by taste? If instead he got all 16, would that prove that the claim is true?

Neither result proves the claim; we cannot construct proofs of such claims. He may have been merely guessing but had great luck. Is luck a plausible explanation? If we assume that he is merely guessing, what is the probability that he actually designates 13 of the 16 martinis correctly? What is the probability that he correctly assigns at least 13 of the 16 martinis correctly? This latter question covers the case of what actually happened and anything more extreme happening and will be a frequent concern throughout this chapter. We can understand the taste testing and probability questions with binomial random variables. We have 16 trials with a success defined as correctly classifying the martini as shaken or stirred. If we assume James Bond is guessing, then the probability of success for a single tasting is $p = \frac{1}{2}$. If X is the random variable counting the number of successes in 16 trials, we are interested in computing $P(X \geq 13)$. The reader is encouraged to verify that $P(X \geq 13) \approx 0.0106$. With such a small probability, we would say that someone randomly guessing at least 13 of the 16 martinis correctly is quite unusual. We have two possible situations: James Bond was guessing and something quite rare occurred by chance, or James Bond actually has the ability to distinguish shaken and stirred martinis by taste. Given the evidence, the claim that James Bond was merely guessing seems dubious; the claim has not been proven false, but there is considerable doubt about its validity. We, therefore, say that there is considerable evidence that James Bond can distinguish between martinis shaken and stirred simply by taste.

We call the process described above as hypothesis testing. There is a claim or hypothesis about reality that needs to be tested (that James Bond can distinguish between shaken and stirred martinis simply by taste). A competing hypothesis is identified (that James Bond was merely guessing). Under the assumption that the competing hypothesis is true (James Bond is guessing), the probability that what happened or even something more extreme will happen is computed $P(X \geq 13) \approx 0.0106$. If this is a rare event, it casts doubt on the validity of the assumption that the competing hypothesis is true; thus, credence is lended to the original hypothesis (that James Bond can indeed distinguish between shaken and stirred martinis simply by taste). At no point in this process, just as in

constructing confidence intervals, is certainty achieved. Rare things do indeed happen, but that does not mean we cannot have confidence in assessing evidence. We will now begin formalizing the process of hypothesis testing.

Hypotheses

As we interact with the world around us, we begin to notice pattens in our observations and start to form hypotheses about the world based on these patterns. Before we act as if these hypotheses are true, we want to secure reasonable and sufficient evidence in support of them. So, how do we collect evidence in support of a hypothesis that arose from the claims or observations of ourselves or others? Recall that, in our example with James Bond, we identified a competing hypothesis, a hypothesis that was opposite of the one he claimed. James Bond claimed to be able to distinguish between shaken and stirred martinis by taste. The opposite claim was that he could not distinguish by taste and, therefore, was guessing. We call these two competing hypotheses the alternative hypothesis H_1 and the null hypothesis H_0. The **null hypothesis** is more practically or reasonably assumed to be true. While the **alternative hypothesis** is generally the novel or claimed hypothesis.

Within the context of the James Bond scenario, our initial disposition was that having such a refined taste palette would be abnormal. It seemed more reasonable to assume, at least initially, that one would simply be left guessing about how the martinis were mixed. Thus, the null hypothesis was that James Bond could not distinguish simply by taste and, therefore, was guessing. The alternative hypothesis was that he could distinguish simply by taste.

Consider another example: preparing to enter the shower. An important concern is the temperature of the water. Either the water is amenable to a pleasant shower or it is not. We have two hypotheses. Which of the two hypotheses do we presume to be true as we prepare to enter the shower? Not many of us turn the water on and immediately hop in the shower. Instead, we wait for it to warm up, waiting to see steam rising or periodically running a hand through the water to test the temperature. These actions speak of an initial assumption that one of the hypotheses is true. We operate with the assumption that the water is unsatisfactory until we have evidence to the contrary. Our actions reveal that the null hypothesis is that the water is not amenable to showering.

$$H_0 : \text{Water is not at a temperature suitable to showering}$$
$$H_1 : \text{Water is at a temperature suitable to showering}$$

These designations could also be thought of in terms of the potential implications of acting as if one of the hypotheses is true when indeed it is not. If we assume that the water is ready but that is not the case. What might happen? Since we assume that the water is good to go, we will hop in the shower right away. Once we are in the shower, we quickly find out that it is either too hot or too cold and immediately feel something moderately unpleasant to possibly painful. If, on the other hand, we assume that the water is not yet ready despite the water actually being perfectly suitable to us, the price is that we wait an extra minute before hopping in the shower. Which of these situations would we prefer if we were wrong with our initial assumption? We would prefer just waiting around for an extra minute over something that could potentially scald us. The hypothesis with the less drastic cost in acting as if it were true when it was false is the null hypothesis.

? Text Exercise 7.1.1

Within each context, determine the competing hypotheses and identify them as either the null hypothesis or the alternative hypothesis. Explain your reasoning.

1. In a court of law in the United States, a prosecutor (a lawyer) argues that the defendant (a citizen) is guilty of some crime. The defendant is usually represented by a criminal defense lawyer who argues against the prosecutor (that his client is not guilty). A judge and possibly a jury follow the arguments in order to draw a final conclusion called a verdict.

Answer

The two competing hypotheses are related to the defendant's innocence regarding the crime charged against him. One hypothesis is that the defendant is innocent of the charge. The other is that the defendant is guilty of the charge. In the United States, our legal system is structured with the mantra "innocent until proven guilty," with sufficient evidence described as evidence beyond a reasonable doubt.

$$H_0 : \text{The defendant is innocent}$$
$$H_1 : \text{The defendant is guilty}$$

Thinking of the potential implications of acting as if one hypothesis were true when it is not can help us understand why our legal system was set up as it is. If we act as if the defendant is guilty despite the fact that he is innocent. We could send

an innocent person to pay a fine, spend time in prison, or even be executed. Our founders experienced tyranny and sought in many ways to protect the citizens from the government. The potential price from the defendant's perspective is quite steep. Sending an innocent man to jail for years or to death is severe. On the other hand, if we act as if the defendant is innocent despite the fact that he is guilty the potential price from the defendant's perspective is the cost of playing the system. It may be that the defendant feels remorse and a desire to change his ways or it may be that he got away with breaking the law. He may be more or less prone to be a repeat offender. The local society may deem acting with more caution around the defendant a prudent decision. From the perspective of our legal system, the cost of an innocent man being wrongly persecuted is worse than a guilty man walking free.

"It is better that ten guilty persons escape than that one innocent suffer." - William Blackstone

"It is better 100 guilty Persons should escape than that one innocent Person should suffer." -Benjamin Franklin

2. A researcher at Stine, a company that develops corn and soybean seeds, identifies a new breed of sweet corn in its breeding laboratory that he thinks will produce corn that is more tolerant to the dry conditions of northwestern Kansas than the breed that most northwestern Kansas corn farmers currently use.

Answer

The two competing hypotheses are related to the drought tolerance of the new variety of corn. One hypothesis is that the new variety of corn has better drought tolerance than the commonly used variety of corn. The other is that the new variety of corn does not have better drought tolerance than the commonly used variety of corn. This could be that it is equally tolerant or that it is less tolerant. The Stine company would benefit tremendously from developing a variety of corn suitable for drier conditions. If the commonly used seed is of a business competitor, Stine can expand its market. Even if it is the producer of the current common variety, the new seed will likely be able to be sold at a higher price and bring new attention to the company. Making the assumption that the new seed is better than the common seed could be disastrous if false. The company would likely lose many clients and possibly face a lawsuit for false advertising. As such, the reasonable hypothesis to assume initially until there is evidence to the contrary is that the new variety of seed is not better suited to the dry conditions of northwestern Kansas.

$$H_0 : \text{The new variety of sweet corn seed is, at best, the same as the commonly used seed}$$
$$H_1 : \text{The new variety of sweet corn seed is better than the commonly used seed}$$

Note: Pascal's Wager

The 17^{th} century philosopher and mathematician Blaise Pascal engaged in a similar line of reasoning before any rigorous development of hypothesis testing in an argument called Pascal's Wager which can be found in his book $Pensées$. Dr. Peter Kreeft, a renowned philosopher and professor at Boston College, gives an exposition of the argument (which can be found here) which we will formulate here using the framework of hypotheses that we have been developing.

Pascal lived in a time of great religious skepticism and attempted to formulate a line of reasoning that could reach a skeptic who lacked faith and did not believe that reason was sufficient to prove that God existed. There are two competing hypotheses at play: God exists and God does not exist. The skeptic knows that only one of the hypotheses is true but cannot establish intellectual certainty as to which to adopt. Pascal, just as is done within hypothesis testing, considers the potential ramifications of living as if the hypotheses were true when, indeed, they were false. Which hypothesis is it best to live by as if it were true? Pascal argues that one cannot abstain from the wager, for we are all already playing the game of life.

In considering the ramifications, Pascal, a Christian, considers the possibility of eternal happiness because everyone seeks maximal happiness. But in considering eternal happiness, he assumes that if God exists, there is paradise (heaven), thus equivocating the existence of God such that there is no paradise (heaven) with there being no God.

Pascal continues. If one lives as if God does not exist when He really does, then that one misses out on the possibility of eternal happiness. If one lives as if God does exist when He really does not, then that one has lost nothing since there was nothing to gain or lose at the moment of death. Which of the two of these potential costs is less drastic? He argues that the cost of eternal happiness is infinitely worse than losing nothing. Then we could say, within the context of competing hypotheses, that the null hypothesis, the hypothesis initially assumed and acted upon as true, would be that God exists.

$$H_0 : \text{God exists}$$
$$H_1 : \text{God does not exist}$$

Accepting the risk assessment of the two hypotheses is a critical part of his argument, and there are reasonable grounds to object and much to consider about the context, assumptions, nuances, and ramifications of the argument. An interested reader is encouraged to ponder the argument more thoroughly and read through Kreeft's exposition linked above.

Collecting Evidence and Making Decisions

In the example about martinis and James Bond, the null hypothesis was that James Bond simply guesses whether the martinis were shaken or stirred. We collected evidence to test his claim, the alternative hypothesis (that he could distinguish between shaken and stirred martinis simply by taste) by conducting an experiment where he taste tested 16 martinis. We then analyzed that evidence under the assumption that the null hypothesis was true by computing the probability that a person simply guessing would get at least 13 identifications correct when presented with 16 martinis where the mixing method was chosen randomly. This probability is called the **p-value.**

The p-value is not the probability that James Bond was simply guessing (i.e. it is not the probability that the null hypothesis is true). Rather, it is the probability that something at least as extreme as what was observed happens assuming that the null hypothesis is true. We can understand the p-value as the probability of an event given a hypothesis. It is often misunderstood as the probability of a hypothesis given an event.

James Bond correctly identified 13 of the 16 martinis. What would be at least as extreme in the context of the taste testing? It would be that he got $13, 14, 15,$ or even 16 martinis right. So in this context, the p-value is $P(X \geq 13)$. When the p-value is quite small, either something rare happened or there was a flaw in an assumption of our analysis, namely, that the null hypothesis is true. We cannot be certain which is the case from what we know, but if the p-value is sufficiently small, we generally consider it as significant evidence that the null hypothesis is false. When this is the case, we say that we **reject the null hypothesis in favor of the alternative hypothesis**. In the case of James Bond, the p-value was 0.0106, which is rather small, so we rejected the null hypothesis (that he was simply guessing) in favor of the alternative hypothesis (that he could indeed distinguish simply by taste). If the p-value is not sufficiently small, the event that occurred was not rare enough under the assumption that the null hypothesis is true to cast doubt on the truth of the null hypothesis. This does not prove that the null hypothesis is true; rather, we simply failed to show it was likely false, so we conclude that we **fail to reject the null hypothesis** when this happens. We emphasize that failing to reject is not the same thing as accepting. Hypothesis testing can only falsify, never verify, the null hypothesis, as throughout the procedure, the null hypothesis is assumed to be true.

Reports of statistical analyses outline the logical progression for the development of hypotheses, experimental design, results of the experiment, and the p-values in order to provide the readers with the full scope of the logic and evidence. This is done because there are different approaches to what is deemed a sufficiently small p-value and the decisions need to be based on more than just whether the p-value meets some given threshold. However, deciding on a threshold which considers the context and is determined independently from the evidence is a good start to measuring the weight of the evidence. Once a threshold is set, we describe the hypothesis test as having that level of significance, which is typically referred to as the **α value** of the hypothesis test. Commonly used α values are 0.05 and 0.01.

From our discussion thus far, we note that the null hypothesis plays a pivotal role in the process of hypothesis testing. The p-value comes from a probability calculation, assuming the null hypothesis is true. The conclusions that we draw from a hypothesis test come in two forms: reject the null hypothesis or fail to reject the null hypothesis. Indeed, we can loosely understand the process of hypothesis testing as the quest for finding evidence against the null hypothesis, so that, when significant evidence (evidence that produces a p-value less than the α value, the significance level) is found, we can reject the null hypothesis and favor the alternative hypothesis.

Type I and Type II Errors

Rare events are not impossible events; they do happen from time to time. As such, it is possible to conduct a hypothesis test, collecting evidence that meets our standard of significance, which leads us to reject a true null hypothesis. Perhaps, we are partially at fault for being too easily convinced that the evidence was strong. Perhaps something extraordinarily rare occurred, but we made an error either way. In rejecting a true null hypothesis, we made a statement about reality that does not match what really is happening. We call this error a **type I error**.

Recall that when deciding which of the competing hypotheses would be the null hypothesis, we were considering the potential ramifications of acting as if one of the hypotheses were true when it really was not. We could also consider the ramifications in the other direction. What are the costs of rejecting one of the hypotheses when it was true; in terms of our current discussion, what are the costs of committing a type I error if we adopt a particular hypothesis as our null hypothesis? Consider the showering example once more. If we reject that the water is suitable, when it is not, the cost is the more drastic of the two. That was the hypothesis we set as the null hypothesis. Thus, the hypothesis with a more severe cost in rejecting it when it is true is classified as the null hypothesis. We do this because we have a certain degree of control over the occurrence of type I errors; we set the thresholds regarding sufficient evidence.

We cannot completely avoid making a type I error, but we can manage the likelihood. As we have seen, evidence is collected, the p-value is computed using the evidence, and then if the p-value is less than the α value (p-value$< \alpha$), we assert that there is sufficient evidence to reject the assumption that the null hypothesis is true. The α value is the upper bound regarding which p-values accounted for sufficient evidence, and recall that the p-value is the probability that something at least as extreme as what happened happens given the assumption that the null hypothesis is true. So, we can understand α as the probability of making a type I error. A smaller α value means a smaller rate of committing type I error.

Making a type I error is not the only way in which we may fail to recognize the reality of the world around us. It may be the case that the null hypothesis is false and would be rejected, but we fail to do so because the evidence did not meet the level of significance desired. Recall that, when the evidence gathered is not sufficient, the result of the hypothesis test is that we fail to reject the null hypothesis. Failing to reject the null hypothesis is not the same as declaring the null hypothesis is true. We are not making a strict assertion regarding which hypothesis matches reality; we are simply saying that the evidence did not cast sufficient doubt on the truthfulness of the null hypothesis. Despite the fact that, in failing to reject a false null hypothesis, we are not asserting anything false about reality, we still call it an error in the fact that we have failed to be better aligned with reality. We call this error a **type II error**.

Once again, consider the showering example. We set up the hypotheses as follows.

$$H_0 : \text{Water is not at a temperature suitable to showering}$$
$$H_1 : \text{Water is at a temperature suitable to showering}$$

A type I error would mean falsely believing that the water is suitable. A type II error would mean falsely believing that the water is not suitable. Notice that if we switched which hypothesis was the null hypothesis, the error types would also switch. The fact that our framing yields a type I error which is more severe than the type II error indicates the hypotheses were formulated correctly.

Note: Hypotheses and Errors

Our original discussion on determining which hypothesis to set as the null hypothesis centered on the ramifications of acting as if one of the hypotheses were true when that was not the case. We set the null hypothesis as the hypothesis with the less drastic costs; the alternative hypothesis would thus have the more drastic costs. This process is similar to the consideration of making a type II error. In failing to reject a null hypothesis which is false, we do not assert that the null hypothesis is true, but our initial disposition towards the hypotheses remains the same. We can connect the two ideas to formalize our hypothesis-setting process; the alternative hypothesis is set by considering the ramifications of making a type II error if the particular hypothesis is set as the alternative hypothesis. The hypothesis with the more drastic cost in making a type II error is the alternative hypothesis. We have two mechanisms for deciding how to set the hypotheses in a hypothesis test (both produce the same results).

1. When considering which hypothesis to set as the null hypothesis, consider the costs of committing a type I error if the hypotheses were adopted as the null hypothesis. Set the hypothesis with the greater cost as the null hypothesis.
2. When considering which hypothesis to set as the alternative hypothesis, consider the costs of committing a type II error if the hypotheses were adopted as the alternative hypothesis. Set the hypothesis with the greater cost as the alternative hypothesis.

Both methodologies produce hypotheses such that a type I error is worse than a type II error. This is because we can precisely identify the probability of a type I error, but we cannot control the likelihood of a type II error. One can verify that one has correctly set up hypotheses by checking that a type I error is worse than a type II error.

Conducting Hypothesis Testing

In inferential statistics, we are primarily interested in claims about populations and assertions about the values of parameters. We could test the claims with certainty if we could compute the actual parameter value, but alas, that is not the case in reality due to literal or practical impossibilities. We use random sampling to collect evidence and then use our knowledge of sampling distributions to compute the p-value.

Consider a seed company with a new sweet corn variety that is thought to produce greater yields in northwestern Kansas than what is commonly planted now. We can measure such a claim by looking at the average yield of the two varieties; these are facts about all seeds of the two varieties, i.e., parameters. After seasons of planting and studying the common sweet corn variety, researchers and farmers have a pretty good idea about its average yield; let us suppose that it is equal to 70 bushels per acre. Since the researchers at the seed company think that they have discovered a better-yielding variety of sweet corn, we can assume there is a line of scientific reasoning or some test plots that have led to such a hypothesis, but we do not know the value of the average yield for the new variety; let us simply call it μ. Let us suppose, for pedagogical purposes, that the new sweet corn varieties from the company consistently have standard deviations of 3.6 bushels per acre; so, that we feel confident the new variety has a standard deviation of 3.6 bushels per acre too.

Just as in the case with the drought-resistant new variety of seed, we do not want to assume initially that the new seed is better than the commonly used seed. So, our null hypothesis is that the average yield of the new variety is, at best, the same as the average yield of the commonly used variety, and our alternative hypothesis is that the average yield of the new variety is better than the average yield of the commonly used variety. We can denote this symbolically as follows.

$$H_0 : \mu \leq 70 \text{ bushels per acre}$$
$$H_1 : \mu > 70 \text{ bushels per acre}$$

Let us suppose that the company has set a standard α value for these sorts of tests at the 0.01 level. With the two hypotheses and α value set, we now look to gather evidence by randomly sampling the yields of the new variety of sweet corn. Perhaps we randomly sample 50 acres across northwestern Kansas where the new variety was grown under comparable conditions to the normal farming practices. From this random sample, we compute the average yield of the sample to be 71 bushels per acre. The average yield of the sample is greater than the commonly grown average yield, but is it sufficient evidence to cast doubt on the truth of the null hypothesis?

How do we assess our evidence? We compute the p-value, the probability that something more extreme happens than what we observed given that the null hypothesis is true. We are interested in computing $P(\bar{x} > 71 \text{ bushels per acre})$. Since we randomly sampled using a sample size of 50, we expect the sampling distribution of sample means to be approximately normal with a mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. From the consistency of the company's seeds, we can feel confident that $\sigma_{\bar{x}} = \frac{3.6}{\sqrt{50}}$. But, from the null hypothesis, we do not know precisely what μ is supposed to be. All we know from the assumption that the null hypothesis is true, is that $\mu \leq 70$ bushels per acre. A common adage tell us to prepare for the worst and hope for the best. We similarly want to consider the case where sufficient evidence against the null hypothesis would be hardest to obtain. Indeed, this is when we assume the new breed produces just as well as the commonly used variety, when $\mu = 70$ bushels per acre. This happens to be the mu value that produces the largest p-value (think about why this is the case). We encourage the reader to verify that the p-value is approximately 0.0248.

A p-value of 0.0248 would constitute significant evidence at the $\alpha = 0.05$ level but would not constitute significant evidence at the $\alpha = 0.01$. This seed company thus would fail to reject the null hypothesis in this situation. That is not to say that the new variety is not better than the common seed. There has not been sufficient evidence to doubt that the new yields are, on average, comparable to the common seed. If the research team is confident in the genetics of their newly developed variety, it may be prudent to conduct another experiment with a larger sample (recall that the standard deviation of the sampling distribution shrinks as the sample size increases) because rare things do occur.

? Text Exercise 7.1.2

Suppose the researchers at the seed company were very confident in the scientific reasoning behind the genetic breakthrough with this most recently developed sweet corn variety, and they decided to conduct an experiment using 10 times as much data, meaning, 500 acres were to be randomly selected across northwestern Kansas to grow this new variety. After the harvest, the researchers found that the average yield from these 500 acres was 70.4 bushels per acre.

1. Using the same hypotheses as the example above, determine the p-value from this larger experiment. State and interpret the conclusion of the hypothesis test at the $\alpha = 0.01$ level.

Answer

Since the hypotheses from the first experiment have not changed, the p-value is the probability that a random sample of 500 acres produces an average yield at least as large as the average yield that was observed. So we are looking to compute $P(\bar{x} > 70.4 \text{ bushels per acre})$. The sampling distribution is approximately normal because the sample size is so large with a standard deviation of $\sigma_{\bar{x}} = \frac{3.6}{\sqrt{500}}$. We will again assume that $\mu = 70$ bushels per acre because that is the condition within the assumption that the null hypothesis is true that will produce the largest p-value. p-value $= 1 - \text{NORM.DIST}(70.4, 70, \frac{3.6}{\sqrt{500}}, 1) \approx 0.0065$.

Since $0.0065 < 0.01$,we say that at the significance level of 0.01 there is sufficient evidence to reject the null hypothesis that the average yield of the new variety of sweet corn yields, at best, the same as the commonly used corn seed. We conclude that the new variety of sweet corn produces, on average, a greater yield than the commonly used seed.

2. At the time of the study, sweet corn was being sold at $\$4.20$ per bushel on average. If a farmer with 80 acres designated for sweet corn was seeking advice about switching to the new breed of corn, what aspects of the study would be important to consider? What would your advice to the farmer be?

Answer

Since we rejected the null hypothesis at the α level of 0.01, there is statistical evidence to say that the average yield of the new variety of sweet corn is larger than 70 bushels per acre, the average yield of the commonly used variety. The hypothesis test itself did not determine by how much the average yield would increase; it only concluded that there was an increase. If the increase was large and resulted in larger profits despite having to pay more for the seed, the farmer may be inclined to make the switch. If the increase was not large, the farmer may be less profitable despite producing more because of the increased cost associated with the new seed. If we used the average yield from the sample to make such a comparison, we would expect the farmer to have an increased yield of 0.4 bushels per acre on average. We could then estimate the increased revenue to be $0.4 \cdot 80 \cdot 4.20 = 134.4$ dollars. Not knowing the actual prices of seed, we cannot estimate the profit, but it seems doubtful that the switch will lead to any increased profits worth noting.

More technical answers for the mathematically inclined.

Rather than just using the sample mean as a point estimate for the population mean of the new variety, we can construct a confidence interval and use the boundary points as upper and lower bounds to create an interval estimation of the increased revenue. Let us construct a 95% confidence interval; $70.0845, 70.7156$So, at the 95% confidence level, we expect the population mean to fall between $70.0845$and 70.7156bushels per acre. This leads us to expect to increase the yield by something between $0.0845$and 0.7156bushels per acre, meaning the expected increase of revenue would be between $\$28.38$and $\$240.42$which again does not seem that a switch would lead to any increased profits worth noting.

As we have just seen, there may be differences that are statistically significant that do not warrant changes in our lives. When this is the case, we say that the findings are statistically significant but not practically significant. Remember that we are trying to understand the world better so that we may better live in the world and interact with the world. The p-value does not measure the size of a difference, often referred to as the size of the effect; it measures the probability that something at least as extreme happens as what was observed to happen under the assumption that the null hypothesis is true. A small p-value does not indicate a large effect. Many people have fallen prey to misunderstanding the meaning and uses of hypothesis testing, especially regarding p-values. In fact, widespread misinterpretations of the p-value prompted the American Statistical Association published an article addressing the misuses of p-values in 2016 to help remedy the issue.

Types of Hypothesis Tests

Let us return to the James Bond example. What would we have concluded if, after tasting the 16 martinis, James Bond was only correct on 3 of them? If he were truly guessing, we would expect that he would be correct about half of the time, but that is not what happened. Such a low score could indicate that James Bond could taste a difference between shaken and stirred martinis, but

he has confused the tastes. He consistently labels shaken as stirred and stirred as shaken. This would indicate that he can taste a difference, but his preference for shaken is mistaken!

If we were simply interested in whether James Bond's taste buds could tell the difference between shaken and stirred martinis, and not that he could distinguish between them due to his strong preference for shaken martinis, evidence against the null hypothesis that he was guessing would come in the form of either really low numbers of successes or really high numbers of successes. We are looking for evidence in two different directions. In this case, looking at the probability of something at least as extreme happening as what happened takes on a more complicated meaning.

With 13 out of 16 taste tests being successful, we easily understand more extreme as at least 13 martinis being correctly identified. But what would be as extreme in the other direction in which we are looking for evidence? That would be 3 out of 16 taste tests being successful. And in this direction, at least as extreme would lead us to consider at most 3 martinis being correctly identified. So, we are looking at the two tails of the binomial distribution to compute the p-value. We call such a test a **two-tailed test**. Thus the p-value is $P(X \leq 3 \text{ or } X \geq 13) \approx 0.0212$. This, again, qualifies as sufficient evidence at the α value of 0.05.

In our previous examples, we have looked for evidence only in one direction. The p-value came from only one of the tails of the probability distribution; we call such tests **one-tailed tests**. Two-tailed tests are most common in scientific research because finding any difference is generally notable, interesting, and could lead to further development. In the case of medicine and business, one-tailed tests arise with greater frequency because there are cases where there is no need to distinguish between no effect and an effect in an unpredicted or undesired direction. For example, with the varieties of corn, the seed company is interested in increasing the yield. If a new breed has the same yield or a worse yield, the new breed does not warrant further consideration.

Note: Two-Tailed Tests and Their Conclusions

Consider the example with James Bond without considering his preference for shaken martinis. Does he possess the taste buds necessary to note the difference between the mixing method? We identified this as a two-tailed test and could write the hypotheses as follows.

$$H_0 : p = 0.5$$
$$H_1 : p \neq 0.5$$

The evidence from the taste tests remains the same; he correctly identified 13 out of the 16 martinis which produces a p-value of $P(X \leq 3 \text{ or } X \geq 13) \approx 0.0212$. At the $\alpha = 0.05$ level, there is sufficient evidence to reject the null hypothesis that he is merely guessing. We could state the conclusion as two sided: he can distinguish (perhaps incorrectly) between shaken and stirred tastes; or, given that he correctly identified most of the drinks, we could state a stronger conclusion: he can correctly identify if a drink is shaken or stirred. We will argue that the latter conclusion is not the proper inference given how the alternative hypothesis was formulated.

The statistician Kaiser published a paper in 1960 arguing that we can make the claim that James Bond can correctly distinguish between shaken and stirred martinis simply by taste. Some textbooks argue this is permissible; others argue that it is not. This alternative hypothesis looks like $H_1 : p > 0.5$ rather than $H_1 : p \neq 0.5$ as originally formulated. The form of the alternative hypothesis changed to match the direction of the sample statistic. We will not adopt this practice. Since we are operating within the realm of formal hypothesis testing, we will maintain the form of the original alternative hypothesis and conclude that $p \neq 0.5$.

To understand our position, we further explain the process and purpose of hypothesis testing. Hypothesis testing is meant to test hypotheses formulated from previous observations, previous experimentation, or working theories. Once the hypotheses have been set, new experimentation is implemented, and the results are used as evidence in the hypothesis testing. The formulation of the hypotheses happens before the experimentation and is independent of the new data. That is not to say that the data from the new experiments cannot be studied to formulate further hypotheses or nuance the current hypotheses. This is to say that these new hypotheses are not to be tested using the same data precisely because the data led to them. That would be circular reasoning. Strictly speaking, new experiments need to be conducted to test the new hypotheses. This includes modifying the alternative hypothesis of a two-tailed test into a one-tailed alternative hypothesis.

This consideration is closely related to a problem plaguing much of modern academia, where publishing statistically significant results often takes precedence over honest intellectual inquiry, resulting in the temptation to conduct unplanned analyses of experimental data in search of a statistically significant result. Acting on this temptation is called p-hacking and is not an ethical research practice. When tests are done on patterns already seen in the data, the tests become meaningless. The proper

path forward is to study the experimental data, formulate new hypotheses, and then test those hypotheses via new experimentation.

Let us consider the case of James Bond once more. Suppose that James Bond could tell the difference by taste but did not have the tastes correctly aligned with the methods. But, somehow, he still managed to get 13 of the 16 martinis labeled correctly. In this case, it is even less probable that he performed as observed than when we just assumed he was guessing. The p-value using both tails was already small enough to warrant rejection at the particular significance level; so, this is evidence that we would expect the parameter p to fall on the same side of the value in the null hypothesis 0.5 as the sample statistic. We agree that it is very tempting to conclude $p > 0.5$. Perhaps in the messiness of practical application, action may to be taken based on that conclusion, but in the rigors of hypothesis testing, another experiment is warranted to test the claim on untested data that is randomly sampled.

There is contention in the textbooks; perhaps, there is less across the field of active statisticians. Interested readers are encouraged to deepen their understanding of both sides of the argument. For an argument to the contrary of our position, see Kaiser, H. F. (1960) Directional statistical decisions. *Psychological Review*, *67*, 160-167.

7.1: Introduction to Hypothesis Testing is shared under a Public Domain license and was authored, remixed, and/or curated by The Math Department at Fort Hays State University.

- **11.1: Introduction to Hypothesis Testing** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.