SHA572: Scientific Decision Making

What you'll do

- **Formulate a question as a null and alternate hypothesis**
- **Calculate a test statistic from sample data**
- **Identify the statistical test most appropriate for testing your hypothesis**
- **Determine the likelihood of finding a result at least as extreme as the test statistic assuming the null hypothesis**

Course Description

Summary statistics are one way to measure uncertain outcomes, and statistical results can be used to inform a decision or business strategy. The summary statistics are based on a data sample, and they typically contribute to a model used to inform intuitive decision-making. That is, the model requires interpretation that relies on the business intuition of the person using it.

This course begins the process of examining sample data under a more scientific regime that restricts or limits the generalizations we can make based on the data. As always, intuition and business knowledge play an important role in the process, but the methods described in this course provide a layer of scientific rigor.

Chris Anderson
Associate Professor, School of Hotel Administration, Cornell University

Chris Anderson is an associate professor at the Cornell School of Hotel Administration. Prior to his appointment in 2006, he was on the faculty at the Ivey Business School in London, Ontario, Canada. His main research focus is on revenue management and service pricing. He actively works with industry, across numerous industry types, in the application and development of revenue management, having worked with a variety of hotels, airlines, rental car and tour companies as well as numerous consumer packaged goods and financial services firms. Anderson 's research has been funded by numerous governmental agencies and industry partners, and he serves on the editorial board of the *Journal of Revenue and Pricing Management* and is the regional editor for the *International Journal of Revenue Management*. At the School of Hotel Administration, he teaches courses in revenue management and service operations management.

Table of Contents

Module 1: Define a Hypothesis

Module 2: Test the Hypothesis

Module 3: Testing and Conclusions

T.TEST function)

7. Watch: Testing Two Sample Proportions
8. Tool: Test Function Flowcharts
9. Tool: Test Templates Workbook
10. Watch: Choosing a Test Function
11. Course Project, Part Two—Test, Analyze, and Conclude
12. Consider Other Alternative Hypotheses
13. Module Wrapup: Testing and Conclusions
14. Read: Thank You and Farewell

Module Introduction: Define a Hypothesis

In this module, you will practice the first step of framing your question as a hypothesis that can be tested in terms of the probability of a potential outcome. As straightforward as this may sound, you will see that for a given set of data there are many ways to structure the hypothesis. The way you structure the hypothesis will impact the conclusions you can draw from your data.

Because the effectiveness of a decision often hinges on the clarity with which a hypothesis is framed, this first step is crucial. Confidence in decision making can come from optimism or reliance on a good track record, but many times what carries the day is a well-crafted story and solid business intelligence to back it up. In the method described in this course, business intuition is used to frame an initial question, but after this point statistical science takes over to generate mathematical, unambiguous results.

Back to Table of Contents

Watch: Using the Experimental Statistical Process

When there is even a little uncertainty, no amount of data about past results can guarantee a specific outcome or definitively support a decision. The next best thing to certainty is a statement about past results that can be tested mathematically to increase our confidence in a course of action.

In this video, Professor Anderson introduces the concept of hypothesis testing. This statistical process uses a "null hypothesis" and an "alternative hypothesis" to establish a testing framework around the question to be answered. The null hypothesis is the outcome we assume, and in our testing we try to disprove it. If we succeed in disproving the null, it lends support for our alternative hypothesis.

Transcript

Okay, so let's focus on the scientific process, right. So we have this issue or question, right. Based upon that question, we're going to form some testable statements, right. We're going to collect some data, run some experiments, and then given that data, evaluate those testable statements, you know, focusing on the significance of those. And then ultimately draw some conclusions based upon the significance of those testable statements. If needed, we might circle back, reformulate those questions, re-evaluate other testable statements, etc. and continue through this process. We're going to focus on hypothesis testing, and how it sort of fits within this scientific process, specifically focusing on how do we translate a issue into testable statements. And then, given I've collected some data, how do I evaluate the significance of that data and relate that back to those testable statements? For example, you might ask yourself, is there such a thing as extrasensory perception or sixth sense? And how do we go about evaluating someone's statement that it exists.

One classic way is looking at Zener cards, right? So we have two experimental participants, right. They're separated, maybe in different rooms or with a divider, and one individual randomly picks a card, one of five cards. Each of those five cards has a different symbol on it. The other participant is trying to guess the symbol on that card. So, because there's just five cards with five different symbols, if I was simply guessing, then I would have a one in five chance of picking the correct symbol. Now, the question might be, if our participant is guessing more than one in five correct, right, how far away from one over five do they have to be for that to provide evidence of ESP? And so we're going to use a statistical test to help us determine how far away from that one-fifth we need to be in order to evaluate the existence of ESP. And specifically, what we're going to focus on for that statistical test is what we call a hypothesis. Right, so statistical tests are framed around hypotheses, we have what's called our null hypothesis. And basically the null is the claim that there is no effect or no difference. And then the alternative hypothesis is really the one that you're seeking evidence for, right?

So, this is the tricky part about hypothesis testing, right? We're trying to find evidence for the alternative by refuting the null, right? And so there are going to be competing claims. And we're going to focus on a sample and we're making these claims about the population, right. So we're going to have an experiment or a collection of data, that sample we're going to calculate some statistics on that and then make some inference and ultimately test that inference within the context of our hypothesis testing. Our null is going to be a very specific statement, right? We're trying to find basically whether or not a specific statement is true or false. If we can reject that null, then that provides evidence that perhaps our alternative is true. Right, so typically our null is a very specific statement. Right, so in the context of our ESP exercise, our null would be the chance of picking the correct card is simply one fifth, or 20% of the time. Our alternative would be that you're able to pick more than one out of five correct, right? So our null is that basically ESP doesn't exist, right, that it's one-fifth. Our alternative is that it does exist. We're indicating that there's more than a one-and-five chance that you would pick the correct symbol. So our null usually involves an equal sign. And our alternative usually involves some form of inequality, right? And so you could think of this as the null is specific. The alternative is rather general, less than, greater than, less than or equal to, or perhaps not equal to. Right, so, it's more of an inequality versus an equality.

Another example comes from a classic study on sleep versus caffeine for recall. Right, so, basically, the experiment is structured this way; students are given a set of words to memorize and then they're randomly allocated to one of two treatments. In one treatment they have a 90-minute nap. In the other treatment they are given caffeine via a caffeine pill. Two-and-a-half hours later, after seeing those words, we're going to test the recall, right, and see if there's any difference in recall for these two different treatments. And so our null hypothesis would be that the words recalled for those who had a nap would be the same as the words recalled for those that received caffeine. And our alternative would simply be that they're not equal. Right? Now you might have different alternatives. If you thought that a nap encouraged recall, then you could have a null were they're equal. And your alternative is that words recalled under the sleep treatment are greater than the words recalled under the caffeine treatment, right? So we can have lots of different nulls and alternatives and the idea is that they're going to link back to what you're trying to investigate.

Back to Table of Contents

Watch: Chance and Statistical Significance

In order to support a decision or point of view, you want to find a pattern of evidence in data that is distinguishable from noise or random chance. If chosen properly, the null and alternative hypotheses allow you to use sample data to reveal patterns that lead you to a degree of confidence for a given position or point of view.

As Professor Anderson takes care to emphasize, statistical results cannot prove anything definitively. However, when statistical significance is established, you have justification to take one position over another. If you've heard the term statistical significance before and are not sure you've fully grasped its meaning, this video and the ones that follow will lead you to a clear understanding.

Transcript

So we have our issue, or question, we're trying to answer. We've translated that issue into testable statements, our null hypothesis and our alternative hypothesis. We've collected some data and now we're basically trying to find convincing evidence in support of that alternative hypothesis. Statistical significance is the way that we define, or seek out, that convincing evidence. And so you can think of statistical significance as a measure of how likely it is to get the outcomes from your sample, given that the null is true, right.

So I have null hypothesis, then I have the sample. If the null hypothesis was true and it was very unlikely for me to observe the characteristics of my sample, then that indicates that perhaps the null hypothesis wasn't true and that provides evidence for the alternative. Conversely, if you mean, if the structure of the sample was consistent, or highly likely, given the null hypothesis, then that doesn't provide support for our alternative hypothesis.

So let's think of this in the context of our ESP exercise. Right, so we're trying to determine whether or not ESP exists. We set up an experiment, we have participants, seeing whether or not they can pick the correct card, right. We go through a series of different participants and, on average, they pick 30% of the cards correct. So, our null hypothesis here is basically that ESP doesn't exist or the chances of picking a card correct is 20%. Our alternative is that ESP does exist and the chance of picking a card correct is greater than 20%. And so now the question is, is 30% actually statistically greater than 20%? And we're trying to evaluate that statement through statistical significance. If we determine that the chance of picking 30% correct, given that our null is true, i.e that you're only able to get 20% correct and that chance is really, really small, then that provides evidence against the null hypothesis, so we would reject the null hypothesis, and, basically, that would provide support for our alternative hypothesis.

Conversely, if, you know, basically, it is possible, just by chance alone, to get 30% correct, even though you are just guessing, then that would indicate that the null is potentially true, so we can't reject it. And then that does not provide support for our alternative hypothesis. We have to be clear here though that, you mean, this link between statistical significance and hypothesis testing doesn't prove the alternative. Right, we're going to make a definitive statement around the null hypothesis. We're either going to reject it or not. If we reject that null then that provides support for the alternative, but it doesn't prove the alternative exists, right. So we can't statistically prove any thing, but we're trying to refute this rather specific statement that we've structured around our null hypothesis.

Back to Table of Contents

Watch: Gaining Confidence in a Result (p-value and alpha)

When you hear someone state a statistical result, you should always pay close attention to the potential for error. If you've ever heard someone state a result "plus or minus" some percentage error, you've heard the expression of a confidence interval. But what does this mean, and how did the person arrive at this result? Likewise, you may have heard the terms "false positive" or "false negative," which refer to two ways that statistical results can fail to predict accurately.

In this video, Professor Anderson introduces the notion of a test statistic and formalizes the use of a confidence interval to either accept or reject the null hypothesis. He discusses Type I and Type II errors that can affect the validity of outcomes for hypothesis testing.

Transcript

So in hypothesis testing, we're focused on how unusual would it be to get results as extreme or more extreme than those observed if our null was true. If it would be very unusual, then the null hypothesis is probably not true If it would be not be very unusual, then there is not evidence against the no hypothesis. We formalized this into the concept of statistical significance. So, when results are as as extreme as those observed are unlikely to occur.

Given the null, then we say our results are statistically significant.

We have to define this level of unlikely. And so we set a critical level. This critical level we refer to as alpha. Now we have to calculate the chance of the these results happening. Right we think of that as our p value right? So p value is a very common statistical term. We think of it as probability of getting the results as extreme or more extreme than the ones observed given the null is true. If that p value is smaller Then this critical level, this alpha, right? Then we, the chance of this null being true is highly unlikely, right? And we would reject this null if that p value is below alpha. Right, so think of, you could think of it this way, we're going to have a sample, we're going to calculate the probability of that sample occurring, or the statistic we measure occurring, given the null. If that p value is less than alpha, say if alpha is 0.05, we say our results are statistically significant at the 0.05 level. We reject the null. There is very little chance of this happening Given that Alpha it true, this provides strong evidence in support of the alternative.

Now at no point in time do we say anything definitive about the alternative. So, if we set Alpha at 0.05, if the sample has a statistic that Occurs or has a P value of less than .05, that doesn't mean that there's a 95% chance that the alternative is true, right? We make definitive statements about the null and we infer or provide support for the alternative. Now given that we're making these sort of chance-based statements then we run the risk that we could be making mistakes. And we have two types of mistakes or two types of errors, right? A Type I error basically occurs where the null was actually true. But we rejected it right? We can control Type I error by setting alpha lower, right? So if we set alpha from .05 down to .01, then we have to have more extreme results in order to reject the null. So we're less likely to make this Type I error.

Our other type of error, Type II error, is basically if our null was false, but we failed to reject it. So it's a little harder for us to control for Type II error, this largely comes about by better experimental design, better sampling, better data collection. Right, a more representative sample that we're performing our analysis on. You could think of this errors as being an Aldous to many common legal systems, right. So, in many legal systems, a person is assumed innocent until proven guilty. Right, so this sort of assumption of innocence is like our null hypothesis, and being proven guilty is like our alternative, right? We must provide evidence, right? That evidence is like our p value from our data And that evidence must be beyond a reasonable doubt. That reasonable doubt is like our alpha. If we convict an innocent person then we've made a Type I error. If we release a guilty person then we've made a Type II error.

Back to Table of Contents

Watch: Practical Significance

When looking at the results of statistical tests, you need to remember that without context your results have no meaning. Finding statistical significance is a goal, but it is not the end goal. When your test generates an outcome, you should review what that outcome means in the context of your problem and decide whether there is any practical benefit from obtaining that result.

In this video, Professor Anderson cautions that statistical significance does not necessarily translate to practical significance. He illustrates this point using the example of a weight loss trial in which the evidence supports a claim of weight loss, but in which the amount of weight lost is negligible.

Transcript

Now, we must be cautious about the relevance of statistical significance. And we really want to differentiate statistical versus practical significance. So, you could think of it this way. If you looked long enough, you will eventually find something and in our statistical world looking long enough means collecting large samples. So if we have a large sample, we're bound to find something that's statistically significant. That doesn't mean it actually matters, so it's not really practical.

So for instance, looking to evaluate a weight-loss program, we recruit ten thousand individuals to participate in this year-long study. And on average, they lose half a pound. We find that because the sample's so large, that that is statistically significant. But in practical terms, a half pound, over a year, that really doesn't matter much. So we have to sort of link this testable hypothesis back to something relevant in a physical setting and make sure that physical setting make sense. So we don't end that statistical significance, we have to sort of put this in a practical context.

Back to Table of Contents

Tool: Errors and Mitigation

Errors and Mitigation Tool

When you use hypothesis testing to draw conclusions about a population and base those conclusions on testing results derived from a sample, it is possible the conclusion you reach will be in error. Errors arising from hypothesis testing can be Type I (false positive) or Type II (false negative). While there is some communication benefit in keeping straight which error type is which, the more important consideration is what you can do to mitigate each type of error.

Use this tool to remind yourself of difference between Type I and Type II errors and as a guide for how to mitigate these errors.

Back to Table of Contents

Watch: Framing the Question with Null and Alternative Hypotheses

Part of the power of hypothesis testing is that for a given set of data, you can frame the null and alternative hypotheses any way you choose. This allows you to tailor your hypothesis test to get as precise or broad results as you need. Framing the null and alternative hypotheses appropriately is one of the most important skills you should take away from this course.

In this video, Professor Anderson summarizes the four-step hypothesis testing process and illustrates framing the hypothesis test for several specific examples. For one example in particular he illustrates how a question might lead naturally to an experiment, and he shows how the experimental results can be formalized after the fact with null and alternative hypotheses.

In the Coke™ versus generic cola example, the value of the test statistic (TS) is found to be 8. It is important to understand that in this example the value of 8 is an experimental result, and that this result of this actual experiment is what the term "test statistic" refers to. A test statistic is not a theoretical result.

Transcript

So, we can think of hypothesis testing as a four-step process, right. For the first step, we're going to frame our question of interest as this null and alternative hypotheses. We're going to collect some data, and we're going to summarize that data with a single measure. That measure we're going to refer to as our test statistic. And then we assess how unlikely it is that we would've observed that test statistic if our null hypothesis were true. Then, given that, we're going to make a decision, i.e., we're going to reject or accept our null hypothesis, and link that back to our setting. So if we think about that first step, right, our null hypothesis, it's kind of like our status quo or no relationship or chance alone. And really, the alternative hypothesis is like your research question, right? What you're trying to find out, right?

So it's really our question of interest, right, that we're trying to evaluate. So let's think about a couple of examples we've seen already, right. So, ESP. Does ESP exist? Right, so, our null hypothesis here would be that it doesn't exist, and that the chance of getting a card correct when you pick one of these five Zener cards is simply just one-fifth. Our alternative hypothesis is that ESP exists, and so the chance of picking a card correct is greater than one-fifth. Or we could think of our weight loss example. So, we're trying to evaluate this weight loss program. Our null hypothesis is that the program is ineffective, that average weight loss equals zero. Our alternative is that the weight loss program is effective. And our average weight loss is greater than zero. So let's think about, let's go through an example together, right.

So we're going to focus on the difference between Coca-Cola and say, generic or store brand soda. So you're out shopping for some food items with a friend. You reach for a two-liter bottle of soda and you're basically reaching for a bottle of Coca-Cola, and your friend indicates, no, grab the store brand, it's half the price. You inform your friend that you're very particular about your soda. And you can always tell when someone has served you a generic or store brand. You're friend says prove it. So here we go. We're trying to sort of prove whether or not you can differentiate Coke from generic or store brand. Before we do that though we have to have this test statistic that we're going to measure, right. So what is going to be the thing we're going to summarize in our endeavor to prove it.

So we could think of our experiments, right? So your friend sets up 10 pairs of cups. Each pair has one cup of Coca-Cola and one cup of our generic store brand. You sample each pair and then you indicate which cup was Coca-Cola and which was the generic brand. After you've indicated so you put the cup upside down and you see whether or not it was generic or Coca-Cola. It turns out you go through all 10 pairs, and you pick eight out of 10 correct. This eight out of 10 is our test statistic. Right, so that's the thing we're going to measure in our experiment. And so now we're really at step three of our hypothesis testing, right. How unlikely is that test statistic if our null had been true? But we really haven't formalized our experiment yet in terms of this null and alternative hypothesis, right? So you're trying to prove that you're skilled at differentiating Coca-Cola versus generic soda.

So our null hypothesis here, our status quo, would have been that you're not skilled. And so the chance of you picking you is simply one half or 50/50. So you're basically just guessing. Is it Coke or is it generic soda? Our alternative hypothesis is that the chance of picking Coke correctly is greater than a half, i.e. you are skillful at differentiating Coca-Cola from our store brand. In our next session, we'll actually go through and formalize and assess you skill level by calculating your p-value or the probability of getting a eight out of 10 correct and comparing that to our preset alpha and determining whether or not we have statistically significant results.

Back to Table of Contents

Apply the Hypothesis Testing Process

Instructions:

You are required to participate meaningfully in all discussions in this course.

Discussion topic:

Consider a decision you face, or a time when you faced a decision, or asked a question based on data from an experiment. This should be a situation for which you have not employed a hypothesis testing framework but could do so. Create a post in which you:

Describe the situation and the decision faced or the question to be answered.
Articulate the null and alternative hypotheses you would use.
Discuss how the use of hypothesis testing might impact the way you thought about or communicate about the situation.

To participate in this discussion:

Use the **Reply** button to post a comment or reply to another comment. Please consider that this is a professional forum; courtesy and professional language and tone are expected. Before posting, please review eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Case Study: How Revinate Used Hypothesis Testing

Hypothesis testing can be applied to test several outcomes simultaneously.

In order to get valid results, it is often necessary to index numerical data from your experiment using control data.

In an effort to evaluate the impact of encouraging customers to provide reviews at TripAdvisor, we conducted a series of before and after tests. We did this in conjunction with online reputation management firm Revinate (see revinate.com for company background) using the Revinate Surveys tool. Revinate Surveys is a survey product that allows hotels to collect private and public guest feedback simultaneously. The survey product allows hotels to send a post-stay short format survey to guests. This survey includes an optional TripAdvisor review form.

Our null hypothesis was that sending a post-stay survey to customers and asking them to review hotels on TripAdvisor would have no impact on metrics used to assess performance of the hotel. Our alternative hypothesis was that one or more performance metrics would improve as a result of encouraging hotel guests to provide feedback in this way.

We collected data from 80 hotels operated by five different management firms, including review data from TripAdvisor as well as hotel performance data compiled by Smith Travel Research (see str.com for company background). In addition to the individual hotels' data, we had similar measurements from the property-determined competitive set. Our review and performance data comprised a period from 12 months before until 12 months after each hotel launched a Revinate Survey. This allowed us to compare the average performance before and after the survey launch. Performance was measured in terms of several factors, including TripAdvisor metrics as well as more traditional hotel performance metrics.

To control for seasonality, all analyses employ indices, each of which is calculated as the hotel parameter divided by the average value from the competitive set. The table summarizes average indexed values across a series of six metrics for the 80 hotels in these before and after tests. The first column in the table summarizes review scores (TripAdvisor Score), showing an average index increase from 99 to 100.8. This increase of nearly 2% indicates that encouraging consumers to post reviews is positively related to an increase in the scores of those reviews in comparison to the hotels' competitive sets.

Similarly, the number of reviews dramatically increased. The test hotels were lagging their competitors considerably before launch (index of 86.1), whereas in the post-launch window these properties had more than double the number of reviews compared with their competitive sets (224.4). Prior to launch, the test hotels had slightly fewer positive reviews (index of 99.9) increasing by 3% (to 102.9) post launch. All three of these TripAdvisor metrics are statistically significant at the 0.05 level, so statistically speaking the indices are different post launch. The hotel performance metrics are not statistically different, though they do show slight gains across all three indicators measured (ADR, occupancy, and RevPAR).

	Trip Advisor Metrics			Hotel Performance Metrics		
	TripAdvisor Score	No. of Reviews	% Positive Review	Average Daily Rate	Hotel Occupancy	R A
Before	99.00	86.1	99.9	94.3	103.4	
After	100.80	224.4	102.9	94.3	104.5	
significance	0.00350	0.000000	0.03190	0.48860	0.18920	0

As you see, six distinct hypothesis tests were performed on the data. For the three relating to the TripAdvisor metrics, it was possible to reject the null hypothesis and provide support for the alternative hypothesis. For the three relating to hotel performance, it was not possible to reject the null even though the data show there were positive impacts on performance that resulted from asking customers to rate the hotel.

To read the full published report of the Revinate study, download the Hotel Performance Impact of Socially Engaging with Consumers PDF.

Back to Table of Contents

Watch: Probability and Extreme Outcomes

A probability measures the chance of a specific event occurring. In hypothesis testing, we are looking at a p-value as an expression of probability. This p-value represents the chance of outcomes happening assuming a null value. The null value is based on the null hypothesis defined before the test.

Your hope is to find outcomes that are unlikely, which are represented by p-values close to zero. These unlikely outcomes can be thought of as extreme outcomes not because they are necessarily unusual, but because they could be considered unusual if you pre-supposed that the null hypothesis is true. If you do observe these outcomes, and if the p-value indicates the chance of observing them is very close to zero, then you have statistical significance. Finding statistical significance for the test statistic allows you to reject the null hypothesis.

In this video, Professor Anderson continues with an example from earlier. In this example, the p-value is not sufficiently close to zero to reject the null hypothesis, and so the hypothesis test does not give support to the alternative hypothesis.

In this video you continue to examine the Coke™ versus generic cola example. Keep in mind that in this example the test statistic of 8 correct guesses represents the ability of one person. We can't say whether the demonstrated abilities of this person are typical or even relevant. The question of what is typical relates to what we're trying to test, and as a starting point we're assuming a typical person would have to make an uneducated (blind) guess, in which case that person should get a correct result 50% of the time. If there were such a thing as skill in our example, and if our test subject possessed that skill, the p-value should have provided evidence of that skill.

Transcript

So in our Coke versus store brand or generic soda example, our null hypothesis is that you're unskilled at differentiating Coke from generic soda; our alternative hypothesis is that you're skillful. Right, so said more formally, your null hypotheses is the chance at picking Coke correctly is simply .5 and in our alternative hypothesis is the chance at picking Coke correctly is greater than .5. Right, so we have our sort of null, sort of conceptually and then formalized very specifically, and our alternative, looser, simply greater than 0.5. So we have our test statistic, which was we picked eight out of 10 correct.

Now we need to sort of focus on the chance of that happening, right. So recall that our p-value is the probability of getting a statistic as extreme, or more extreme, as the one you observed by chance alone, if the null hypothesis is true. So if you're lucky and you're simply guessing, what's the probability that you would get eight or more correct out of 10 pairs? So it happens, right, that the chance of getting eight, nine, or 10 correct, right. So, at least as extreme as you observe. So, we observed eight. So, keep in mind that at least as extreme would be eight, nine, or 10 correct. The chance of that happening is 0.055. All right, so there's a 5.5% chance, that if you were guessing that you could get eight or more correct. As that 5.5% chance is greater than our typical alpha of .05, then we won't reject the null. Right, so we don't have any support for our alternative hypothesis that you're skillful.

Back to Table of Contents

Watch: Calculating Probabilities (p-values and the BINOMDIST function)

Performing sophisticated statistical analyses doesn't necessarily require specialized software. Microsoft Excel's built-in statistical functions are sufficient for all the calculations in this course and for many others besides.

In this video, Professor Anderson explains generically how p-values are calculated. He introduces the BINOMDIST function in Excel and describes how it can be used to calculate the probability, expressed as the p-value.

Note: In this video Professor Anderson exposes some of the mathematical calculations as he works through an example. You need not be too concerned with the mathematical details, especially if you are not working in a technical capacity. It is enough to follow in a general way the concept being discussed.

Transcript

In our Coca-Cola versus generic store brand cola exercise we indicated there was a 5.5% chance that you could've correctly picked Coke eight, nine, or 10 times out of 10. And so the question is, well how did we calculate that 5.5? Right, so given our null hypothesis is that you're lucky, we can think of, as picking Coke correctly as analogous to flipping a coin. Right? It's a 50/50 chance that you get heads versus tails. And it's a 50/50 chance that you could've correctly picked Coke when you are simply guessing. So if we think about flipping a coin, right, and what's the probability you get heads? It's simply 0.5. Now, if I do two back-to-back coin flips, and I ask, what's the probability I get a head on both of those? It's simply .5 for the first flip multiplied by .5 for the second flip.

So .25. Now it gets a little more complex though, as we sort of say, well what's the probability I could get two heads out of three flips? Because now I have many different combinations, specifically three combinations. Or I could have flipped two heads and then a tail. Or I could have flipped a tail first followed by two heads or head, tail, head. Right, so we can think of each of these coin flips as an independent event, right. So they're not linked to each other, you have no particular skill in flipping coins, right. And we can generalize this process where we're thinking about calculating in a probability of getting x heads out of say, n flips, given you have a 0.5% chance on any given flip of getting a head, right. And so that generalization we refer to as our binomial distribution, right, where this binomial distribution literally calculates the probability of heads and tails and then enumerates these different combinations of getting this case, two heads and one tail.

So, if we look at this in Excel, we could calculate those probabilities for our Coke versus generic coke example. So we're going to use the BINOMDIST function in Excel. This BINOMIDST function has four inputs. First of all, it is the x of concern, right. So what's the probability of getting eight correct? Our second entry is our number of trials n, in this case 10. Our third entry is the probability of that success, so 0.5. And our last entry is either a zero or a one. It's a zero if I want the probability of getting exactly x correct, and it's a one if I want the probability of getting x or less correct, right. So we could calculate the probability of getting eight correct, right. That is going to be 0.044. The probability of getting nine correct, 0.009. And the the probability of getting 10 out of 10 correct. That's a very, very small number, right. If we were to sum those three probabilities up, they sum up to .055. Right, so the probability of getting eight correct, plus the probability of getting nine correct, plus the probability of getting 10 correct is the chance of you getting eight or more correct, given you were just lucky, right. That probability, 5.5% or .055, was greater than our alpha and hence we did not reject our null.

Back to Table of Contents

Hypotheses and Errors

It is important to be attuned to the potential for errors when testing hypotheses. In this quiz, you will review a scenario and decide which of two outcomes represents a Type I error and which represents a Type II error.

You must achieve a score of 100% on this quiz to complete the course. You may take it as many times as needed to achieve that score.

Back to Table of Contents

Watch: Statistical Significance and Size Effect

In this video, Professor Anderson examines the relationship between sample size, p-value, and statistical significance. Specifically, he discusses how to interpret large effects seen in small samples and small effects seen in relatively large samples.

Transcript

So, in our Coke versus store brand cola example, it seems kind of strange that while I picked eight out of 10 pairs correctly, that we did not reject the null, right? That seems skillful but in the end we don't reject the null that you're simply guessing. This is, you know, largely the effect of our sample size. In fact, if we had maintained that same level of skill but had a sample twice as big, right, so if you picked 16 out of 20 pairs correctly, then that would have had a p-value of 0.006, again, which is less than our typical alpha of 0.05. So, in that context we would have rejected the null, right. So the same sort of effect but a larger sample. In fact, even further, if we had, you know, less skill, right, so say I picked 60 out of 100 correct, that would have a p-value, or a probability, of getting 60 out of 100 correct, by simply guessing, of 0.028, which is also less than alpha at .05. So again, I would have rejected that null.

Right, so we see here the interplay between p-values, the effect, and the sample size. Right, so, if the effect is small, the sample size has to be large in order to have a p-value sufficiently small to reject the null, right. The flip side though is if you have a small sample and you end up rejecting the null, then that means you have a pretty impactful or practically significant effect, right. And so we see this sort of interplay between those. Also keep in mind that if I have a really small effect, right, or almost no effect at all, and I have a large sample, then I could still reject that null hypothesis, right. So we have this direct linkage between the size of the effect, the size of the sample, the resulting p value and then whether or not we accept or reject that null hypothesis.

Back to Table of Contents

Watch: Calculating Goodness of Fit (the CHIDIST Function)

So far the examples in this course have involved sampling quantitative variables. In this video, Professor Anderson shows an example of hypothesis testing with a qualitative, or categorical, variable. Specifically, this example considers outcomes of a roulette wheel that is being tested for fairness. Though they are expressed as numbers, each possible outcome on the wheel is a separate category of a qualitative variable.

When looking at more than two categories of a qualitative variable, p-values can be calculated using the CHIDIST function in Excel. The CHIDIST function is used to perform a test known as the Chi-Squared test, which is just another way to measure how closely observations (data) align with the null hypothesis.

Note: In this video Professor Anderson exposes some of the mathematical calculations as he works through an example. You need not be too concerned with the mathematical details, especially if you are not working in a technical capacity. It is enough to follow in a general way the concept being discussed.

Transcript

So we have this four-step process for hypothesis testing. Those four steps are relatively consistent across different tests, except for steps two and three. Steps two and three will depend upon the nature of the test, what data we collect, what statistic we measure, and then how we calculate the p-value, or probability, of getting a value at least as extreme as that test statistic, given the null. So let's focus on an example to illustrate. So we have a casino that's concerned whether or not its roulette wheel, is functioning properly, right. So during 18 hours of play the wheel is spun 912 times. Across these 912 times a ball could land into one of 38 different cups. The cups are numbered one through 36 and then there is a zero and a double zero. So, if our wheel was working effectively or properly then we would expect the ball to land into one of these 38 cups with a one-in-38 chance, right. So we would expect each cup having the same chance of the ball landing in it. If the wheel was not working properly, our alternative might be that those probabilities are not equal, right.

So our null: it's working properly, probabilities are equal, alternative: not working properly, probabilities not equal to one over 38. And so now, how do we sort of calculate a test statistic, given that null? Well, if we spun the wheel 912 times, it had 38 possible outcomes, on average, we'd expect the ball to land in each cup 24 times, 912 divided by 38. If we compare what actually happened versus what expected happened, or what we expect to happen under the null, right. So under the null we expect it to land 24 times. If we take the difference between the observed and the expected. If we square that difference to make sure pluses and minuses don't cancel out. And then if we want to compare this from experiment to experiment, we have to standardize this to the size of our experiment. Right, so let's divide that observed minus expected squared, divided by our expected, Given that if we had done this 9,000 times, we would expect a very different set of numbers, right.

So we have this measure of distance between the observed and the expected. Observed minus expected squared, divided by the expected. That's the difference between the observed and our null. If we sum that up over all 38 possible outcomes, we have this total measure of fit. How well does our data fit our null hypothesis? We actually refer to this as a goodness of fit test. And it's formally called a Chi-Squared Test, where we are now going to compare this distance between what we observed and what we had expected. So how do we translate this distance to a p-value? Right, so in our Coke versus generic coke example we used the BINOMDIST function to calculate the probability of getting eight, nine, or 10 correct when I was guessing.

So for our goodness of fit test we use what's called the chi distribution or the CHIDIST to calculate the probability of getting this difference or this distance, given our null hypothesis. Now obviously if I have more outcomes, then that difference will be different. And so you can think of the number of outcomes as what we call the degree of freedom. Right, so even though we have 38 outcomes, we have 38 minus 1 degrees of freedom, because if I give you the total distance, and you have 37 of the 38 different distances, then you could calculate the 38, right?

So you really only have 37 unknowns once you know the total distance. Right, so in Excel we have this CHIDIST function. We give the CHIDIST function the dist, the distance, right, the sum of these observed minus expected squared over expected, comma our degrees of freedom, it returns this p-value, or probability, of 0.75, right. Because that 0.75 is bigger than our alpha, our traditional alpha of 0.05, we cannot reject the null that the wheel was working properly, right. So it is not unlikely to get those observations if the null was true if the wheel was working properly.

Back to Table of Contents

Watch: Getting the Null and Alternative Hypotheses Right

For a given data set there are a multitude of ways you can structure the null and alternative hypotheses. The choice you make for these will determine what you use as a test statistic. More importantly, the choices you make will impact what conclusions you can draw after you get test results.

In this video, Professor Anderson works through another example involving a categorical variable. In this case, the question centers on bias in selecting participants in a youth soccer league. The results of the hypothesis test can be used to say something meaningful about the situation, but if the null and alternative hypotheses had been chosen differently it's possible that an even stronger statement could have been made.

Transcript

So when we're designing hypothesis tests, we have to be cognizant of the implications of what our null and our alternative are, and how we structure the test statistic around those. Because depending upon how we measure that test statistic, it's going to limit what we can say vis-à-vis those null and the alternative. Let's look at an example to make this a little more concrete. So let's focus on participation in youth elite soccer, or football. And so we're focused on boys participating in a 16-and-under league. We have data on 288 boys participating on this team. And our data is what month they were born in, specifically, we have the number of boys that were born in January, the number of boys that were born in February, etc., through to December.

So, we might sort of formulate some hypothesis around when you were born, and whether that, whether or not that's impacting your participation in this elite sports team. So, if we said that our null was that birth month was not impacting participation, that you were equally likely to be on this team if you were born in January versus December, then that means if we had 288 observations, or 288 boys across 12 months, on average we'd expect 24 boys to be born in each of the 12 months. If we actually compare the number born in each month versus the expected, right, so the expected is given our null, right, what's the difference between what we observed and that null, if we take that observed minus expected and we square it and we divide by expected and we sum that up, we have this total difference between our sample and our null. If we compare that now or we sort of measure the probability of that distance using our CHIDIST function. Right, so what's the probability, we could have this distance, right, given we have 11 degrees of freedom. We have 11 degrees of freedom because we have 12 different outcomes. But if I give you the sum, then that means that the other, if I give you the sum in 11 of the 12 counts, you could calculate the 12th, right? So you really have only 11 unknowns once you know the sum. If we sort of give our CHIDIST function that distance and that degrees of freedom, it returns the probability, that probability is really, really small. Right, so it basically says there's almost no chance that you could get these number of counts for each month, given your null, right. So if the null was true it would be very unlikely to observe this distribution of birth months. Now, so we're going to reject our null, right. And provide support for alternative. But our alternative is simply that birth month is impacting participation. We can't really go any deeper than that. We can't sort of say, which months are impacting. When we clearly look at the data, we see a much higher propensity for boys to be born earlier in the year versus late in the year. But that's not the test statistic we calculated. That's not the null that we looked at. We just looked at are they equal, versus not. Right, so depending upon how you structure the null and the resulting alternative, will dictate how you measure that test statistic, and what sort of conclusions you can make. If we wanted to sort of say, being born in the first half of the year was different than born in the last half of the year, we would have a very different null and a different alternative and a different test statistic, right. So, we have to be cognizant of what we're trying to show, what the resulting null and alternative are, and how we measure that vis-à-vis our test statistic.

Back to Table of Contents

Determine Appropriate Null and Alternative Hypotheses

In this quiz you will review several scenarios and determine which null or alternative hypothesis best address the situation and the question being asked.

In the first two questions, you will identify the most appropriate null hypothesis for a given scenario. In the second two questions, you will be given a scenario and a null hypothesis, and you will identify the most appropriate alternative hypothesis.

You must achieve a score of 100% on this quiz to complete the course. You may take it as many times as needed to achieve that score.

Back to Table of Contents

Watch: Testing Association (the CHITEST Function)

Hypothesis testing can be adapted to consider more complex questions. In this video, Professor Anderson describes a situation in which two categorical variables are being considered simultaneously. The goal is to determine if there is an association, or link, between the two variables. For example, if you wanted to demonstrate a gender-based preference for a specific genre of movie (say, horror films) over each of several other genres, you could use an association test to support your hypothesis.

In Microsoft Excel, the CHITEST function takes inputs that can include a range of values spanning more than one row and more than one column simultaneously. The need for such a data range results from a situation in which at least one of your categorical variables has more than two categories. The two inputs required by the CHITEST function are the expected values and the actual values. From these values, the function calculates a p-value that describes the goodness of fit between the actual values and the values we would expect to see if the null hypothesis is true. As with all previous tests, this p-value indicates whether there is statistical significance that allows us to reject the null hypothesis.

Transcript

So we can use hypothesis testing and our hypothesis testing framework to determine whether or not factors are impacting each other. Right, so, if we had two categorical variables, and we're concerned or inquisitive of whether or not the outcomes of one categorical variable are impacting the other, we could test that association with a hypothesis test. So let's do an example. So we have a firm, and this firm is looking at some cost-cutting measures. And it's interested in employee attitudes between two alternatives, right, basically layoffs or wage freeze. And so in order to get an assessment of employee attitudes towards these two opportunities, they survey a subset of their employees.

And so basically, the firm surveys 120 employees, these 120 employees are across three different job categories, frontline, salespeople, and administrative. And each of these different levels of job category indicates their preference for a wage freeze or layoffs. And our firm has noticed some differences across preference for each of these different outcomes. Right, so, it looks like administrative staff sort of prefer a wage freeze whereas sales individuals are more in favor of layoffs. And so the question really is, are these differences we see here substantive? Or in our language, are they statistically significant? So we might formulate some hypotheses and test those hypotheses. Right, so, one hypothesis might be or one null hypothesis might be that there's no relationship between job category and cost-cutting measure preference. The alternative would be that there is a relationship. And so now we have to figure out how do we test those hypotheses, right.

So we could look at the differences between what we observed and what we would expect, if the null was true, right. So, for example, we have roughly, or one-third of our staff, 40 out of 120, are frontline staff. Right, so, a third of our staff are frontline. If we look at staff who are in favor of a layoff, there's 77 individuals who are in favor of a layoff. So if there's no relationship between job category and cost-cutting measure preference, then we would expect those 77 individuals in favor of layoffs to be proportionately split across those three categories. Specifically, we would expect one-third of those 77 to be frontline staff. We would expect one-half, or 60 over 120 of the 77 to be sales individuals and one-sixth, or 20 over 120, to be administrative. And so those would be our expected distribution of those favoring layoffs across those job categories given the makeup of those who responded to our survey, right. So those would be our expected counts. And so we have a set of expected counts, we have actually what we observed, and so we could compare those observed to those expected, the expecteds given the null.

Again, using our CHIDIST or chi-distribution test, basically our goodness of fit test, right. So if we pop over to Excel we can sort of do that test very simply, right, where we have our set of responses across our three categories and two different cost-cutting measures. We can calculate our expected counts, where our expected counts are each of the preferences for the cost-cutting measures, either the 43 or the 77, times the proportion of our respondents, right, so times the 40 over 120, the 60 over 120 and the 20 over 120. Those would be our expected counts. And now we simply want to compare the expected to the observed. We can do that with our CHITEST function, right.

So in Excel we had this CHITEST function and we simply say, CHITEST observed, expected. Right, so we don't actually have to calculate the distance and square it and divide by the expected. We can use this CHITEST function, and it returns the resulting p-value. That p-value is very, very small. That p-value is less than 0.05. So therefore, we reject the null. We reject the null that all job levels view these two different cost-cutting measures the same. Right, so clearly there is preference, right, across job category for different cost-cutting measures. Right, so we can use our hypothesis testing framework In many different contexts as long as we can frame that null and those alternatives and calculate the corresponding test statistic.

Back to Table of Contents

Watch: A Summary of Hypothesis Testing

In this video, Professor Anderson briefly summarizes the hypothesis testing process.

Transcript

So we use statistical tests on sample data to assess a claim about the population. Those samples might be 10 taste tests for Coke versus generic coke. They might be a series of spins of a roulette wheel, or responses from surveys, survey responses from employees on their attitudes toward cost-cutting measures. Our statistical tests are usually formalized around two competing hypotheses: Our null hypothesis is our no effect, no difference, or sort of an equality. And our alternative hypothesis is really our research question. What are we seeking evidence for?

If our data are statistically significant, it indicates that there was very little chance to observe the characteristics of our sample, if our null was true, providing support for our alternative hypothesis. Our four-step, sort of, recipe toward hypothesis testing is going to be consistent across different approaches and different tests. We're going to have different data we collect and different test statistics that we enumerate as a function of that data. But once we have that test statistic, we're always going to calculate a probability, compare that to our alpha, and make some accept-reject decisions around those null hypotheses.

Course Project, Part One—Set Up the Hypothesis Test

Scientific decision making begins with a hypothesis. You have practiced casting a question as a set of null and alternative hypotheses, and you have considered several factors relevant to hypothesis testing. In this part of the course project, you will write hypotheses to test a question that is relevant to you personally or to your business. You'll establish your degree of sensitivity to potential testing errors, and you'll clearly articulate what and how much data will be needed to reach a conclusion that you consider valid and useful.

Completion of all parts of this project is a course requirement.

Instructions:

1. Download the course project document.
2. Complete Part One.
3. Save your work.
4. You will submit your completed project at the end of the course for grading and credit.

Do not hesitate to contact your instructor if you have any questions about the project. You will add to this document as the course proceeds and will submit it to the course instructor at the end of the course.

Before you begin:

Before starting your work, please review the **rubric** (a list of evaluative criteria) for this assignment. Also review eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Module Wrapup: Define a Hypothesis

In the world of statistical inference, the null and alternative hypotheses are essential to drawing meaningful conclusions in the same way that a foundation is essential for the support of a tall building.

In this module, you cast your question as null and alternative hypotheses. You examined your risk tolerance in terms of Type I and Type II errors, and you considered what you might do to mitigate those errors. You identified the test statistic you'll use to test the null hypothesis. Next, you'll be taking a closer look at how hypothesis testing works and how you build confidence in the conclusions you want to draw. Once you have a solid understanding of the mechanics of hypothesis testing, you'll be ready to use some tools that are provided later in the course to support calculations of probabilities.

Back to Table of Contents

Module Introduction: Test the Hypothesis

Anyone can make a claim, and without too much effort they can find (or fake) statistics to back up their claim. In this module, you will examine sampling more closely and perform calculations that help you understand what goes into a statistical claim.

You have developed testable hypotheses based on a question you are trying to answer, and now you are ready to put your null hypothesis to the test. Remember that your goal is to try to refute your null hypothesis and, as a result, give support for your alternative hypothesis. As you work toward that goal, you'll develop a concrete, statistical understanding of what is involved in making strong, defensible statements about the null.

Back to Table of Contents

Watch: Sampling to Support a Valid Conclusion

At some point you will need to generate a sample of your data so you can make an inference about the population. You want this sample to be representative, but even a representative sample can lead to a sample statistic that is not accurate for the population. So what *can* you say about the population based on this sample?

In this video, Professor Anderson examines more closely the question of what you can say about a population based on a sample taken from that population. In particular, he considers the impact of sample size and discusses how the structure of the hypotheses needs to take into account the sampling process.

Transcript

All right, so now our focus is going to be on the role of sampling within hypothesis testing. So let's take an abstraction for a minute, and just suppose we have this random event and this random event is true 40% of the time. And so if we were to do 10 trials of this random event, you wouldn't necessarily be surprised if two times this random event came up true. Now you'd probably be a little bit more surprised, though, if we did this random event 1,000 times and it came up true 200 of those 1,000 times. And so keep this sort of in mind as we're starting to talk about sampling. Because sampling is, we have this population, and we take a subset of this population, which we'll call our sample. Ideally, that sample is representative of the population, just smaller in nature.

And then we describe that sample with some descriptive statistics, and then based upon those descriptive statistics, make some inference or draw some insight into the population.

Now, just like we had this sample of 10, or that sample of 1,000 of this random outcome, our sample is a chance outcome, or chance representation of the population. And, we might diligently sample through things like randomization, where we make sure every member of the population has an equally likely chance of being in the sample, right. That will help us reduce bias and make the sample more representative. We might use things like experimentation, which allows us to make more causal statements about the sample and ultimately about the population, and allows us to avoid sort of secondary issues or confounds, right. So we might have a well-designed sample. But, at the end of the day, this sample, the sample is still a sort of chance representation of the population. And this process of sampling, though, is for us to make some definitive or conclusive statements about the population. And because of this chance outcome nature of sampling, hypothesis testing sort of really informs us that we can't make definitive conclusions about the population. Instead we're making these sort of probabilistic statements about the population. And that's why we design this null and this alternative hypothesis. Because if we can make a probabilistic statement about the null, ultimately, hopefully rejecting the null, then we can infer, or provide support for the alternative hypothesis.

So let's go back to this sort of random event. Instead of it being just some general random event, which comes up true 40% of the time, think of it as a referendum. And we have this issue, which is potentially contentious with the population and we're trying to get some insight into what fraction of the population is in favor of this issue, so when the referendum goes to full vote, will it pass or not. Right, so, keep in mind that, you know, the population is 40% sort of against this issue and 60% in favor of it. But we don't know that. We don't know the general characteristics of the population. So instead we sample a subset of those. And based upon that sample of 10 or that sample of 1,000, we're trying to infer the characteristics of the population. And so I could never definitively say that 40% of the population is against this issue.

I might be able to make some probabilistic statements around whether or not this issue or this referendum would pass, right? Would the majority of the population be in favor of this issue? And so while I sample 10 or 1,000 representatives of the population, then I might structure a null around that hypothesis, right, is 50% of the population in favor of this issue, with my alternative being that 50% of the population or greater than 50% of the population is in favor of this issue. And so hopefully I reject the null and provide support for the alternative, which indicates that if this issue went to vote, it would pass the referendum. Right, so my goal is not to find the 40%, but to get some insight into whether or not this issue would pass a popular vote. And so we get an idea of where hypothesis testing is involved in this sampling process.

Back to Table of Contents

Watch: Getting a Result You Can Rely On (Confidence Intervals and the Normal Curve)

Chances are, even if you can convince yourself and others that your sample is unbiased and that your hypotheses are structured properly, this is no guarantee that your conclusions based on an analysis of this sample will be valid. It's impossible to ignore the fact that any sample is an inexact estimate of the population.

You will derive statistical measurements using samples, and those measurements will be precise numbers. But it is unclear how valid those numbers are for the population. The mathematical construct referred to as the "confidence interval" serves the purpose of addressing the uncertainty that goes along with sampling. By specifying a confidence interval, you are in effect saying there is range which includes nearly all possible values of the statistic measured.

In this video, Professor Anderson uses an example to introduce confidence intervals and further explore the question of how to get reliable results. He introduces the normal curve, which is also known as the bell curve or the Gaussian curve. And he discusses how sample size affects the shape of the normal curve and, in turn, the reliability of your results.

Transcript

So we have a sample and we take some descriptive statistics of that sample and then we use those descriptive statistics to make inference about the population. But, you know, we might ask ourselves how comfortable are we in making that inference? How good of an estimate is that sample descriptive of the true population value?

And so as an example, let's, let's be a turkey farmer. So we're raising turkeys. And we're trying to get a sense of the weight of our flock of turkeys for the upcoming holiday season. Obviously, we don't want to go out and weigh all our turkeys so instead we just go out and we just randomly catch nine turkeys and weigh these nine turkeys. Turns out they have an average weight of 15.6 pounds. And a standard deviation of 4.8, where that standard deviation is a measure of spread about the 15.6 average. Now, just for curiosity, we again take another sample of nine, and we weigh those nine turkeys, right. We take another sample of nine and we repeatedly re-sample from our flock of turkeys, calculating the mean for each of those instances. And if we were to plot that as a sort of histogram, or a dot plot, we would get a sense of the distribution of the means of those samples of nine.

Now if you wanted to do this a little bit differently, instead of taking samples of nine, we might take samples of 81. So now I go out and catch 81 turkeys, and I calculate the average of those 81, which, you know is 18.5 for this one sample, and it has a standard deviation of 4.5. If I was to, you know, again resample another 81 turkeys, calculate their average weight, resample another 81 turkeys etc., and you know plot this similar distribution of these mean weights of these samples. Interesting enough, both of these samples tend to have this nice bell shape. We call this bell shaping normal distribution. It has these nice characteristics that the distribution is symmetric about the mean. And we can describe it with just two parameters, the mean of the distribution and the standard deviation of the distribution.

What's really cool about this distribution is the range of outcomes and how we can quantify those possible ranges just with the mean and the standard deviation. So we can show that about 68% of outcomes are within one standard deviation of the mean. 95% are within two standard deviations. And a little more than 99 are within three standard deviations of the mean. Right, so, these distributions of sample means follow this bell shape. And this bell shape can be described by these two parameters. And the range of outcomes can be quantified by these two parameters, the mean and the standard deviation. Now when we look at these two samples of means, of samples of nine and the samples of 81, we see they have a consistent middle point. So the means of these samples of means tend to be consistent and actually they're, you know, basically very close to the population mean. The distribution of samples of nine tends to be a little fatter than the narrower distribution drawn from samples of 81. And so we're going to use these characteristics of these distributions to help us make some inference about the population parameter.

Back to Table of Contents

Watch: Finding the Population Mean Indirectly

If you wanted to study a population using just one sample, the ideal sample would be one that is identical to the population in every way except size. While you can never rely on selecting this ideal sample, it is still possible to estimate the mean for a population indirectly using samples.

In this video, Professor Anderson explains the conditions necessary to assume a normal distribution for sample statistics. He goes on to describe the relationship between the standard deviation of the population and the standard deviation of a sample. Establishing the relationships between population statistics and sample statistics is important because this makes it possible to look at sample statistics and draw conclusions about the population with a known degree of confidence.

The example in this video deals with sampling turkeys and calculating average mass. This example may not seem immediately relevant on its surface, but in fact you can think of turkey sampling as an experiment in describing product yield. As with many of the examples you work through, if you can think of the example in generic terms it will increase the chance that you can find its relevance for your own situation.

Transcript

So recall that our normal distribution has these nice properties.

That's 68% of all possible outcomes are within one standard deviation of the mean, 95 within two, and over 99 within three standard deviations of the mean. And so, we like these nice properties and so if we're focused on the mean of a sample and if we were re-sampling the means of different samples, we illustrated that those distributions of those sample means follow this normal distribution. That happens if either one of two cases occur. If either the population, in this case of turkeys, if the population of turkey weights follows a normal distribution, then the distribution of sample means will also follow a normal distribution.

If the underlying is normal, then the mean of a series of samples will follow a normal distribution. Or if turkeys themselves, or the population of turkeys, did not follow a normal distribution, so it was some other kind of shape, the distribution of sample means will still follow normal distribution as long as I take larger samples. And by larger samples, we mean samples of size 30 or bigger. Right, so if the population is normal or if I have larger samples, 30 or larger, then the distribution of means of samples will follow a normal distribution. And in addition to having this normal distribution, the center of those distributions, or the means of those distributions, is actually the mean of the population, right?

So, for our turkey distribution from samples of nine and our distribution from samples of 81, they basically had the same peak, and that peak turns out to be the population mean. Now, the two different distributions though, one from nine and one from samples of 81, had different spreads. And it turns out that the standard deviation of those sampling distributions is a function of two things:

the underlying population standard deviation, so the population, the standard deviation of the underlying turkeys, and the sample size. Right, so if our sample size is larger than the distribution of those means is narrower, and if we have a smaller sample that distribution of means is wider.

And it turns out that the standard deviation of the sampling distribution equals the standard deviation of the population, divided by the square root of the sample size. Right, so for our, based upon our samples of nine turkeys, that has a standard deviation equal to the standard deviation of the population, divided by root nine, or three, versus the standard deviation of the population divided by nine, or the square root of 81, for our larger samples. Right, so we have these really nice properties that the sample mean follows this normal distribution and has a mean of the population and a standard deviation which is a function of the underlying population standard deviation and the sample size.

Back to Table of Contents

Watch: Estimating the Population Proportion

Categorical variables are typically summarized using proportions. So when you are looking at categorical variables, you will have defined the categories in a useful way and you will form a hypothesis around how prevalent one category is relative to the others. For example, a shop clerk may hypothesize that six out of every 10 visitors to their store leaves without buying something. In much the same way as with quantitative variables and means, you can use the characteristics of a sample to infer the properties of the population. For the shop clerk, sampling might mean tallying paying customers versus browsers for one hour a day over a the course of a month.

In this video, Professor Anderson discusses how statistical values of the sample can be used to make statements about the population with some degree of confidence. As with mean values, the statements that can be made will either make it possible to reject the null hypothesis or they won't. As you watch the video, think about how you would generically characterize the test statistic, which is the proportion of sick turkeys, to better understand how this example could be relevant for some problem in your own situation.

Note: This video includes a number of formulas and symbols that might not be familiar. Before watching the video, you may find it helpful to download and have on hand the Summary Stats and Testing Reference tool and the Statistical Symbols tool highlighted later in the module.

Transcript

So, just like we used the mean to help describe a quantitative variable, we use proportions to help us describe categorical or qualitative variables. Right, so I might focus on the mean weights of turkeys. I also might be interested in what fraction or what proportion of turkeys are exhibiting some disease. Now, for our sample, our distribution of sample means, we indicated that would follow a normal distribution. It turns out that our distribution of sample proportions will also be approximately normal, as long as a couple of conditions are satisfied. Right, first and foremost, we have to have this population which is exhibiting these categorical or fixed proportion characteristics. We have to have randomly sampled from that population, and our sample has to be of sufficient size that we observe at least 10 observations in each case.

Right, so if we're focused on estimating the fraction of turkeys that are diseased, when we sample those we have to get, in our sample we need at least 10 disease turkeys and at least 10 healthy turkeys in order to estimate the fraction that are diseased. And so if we were to do this, if we were to sample these different proportions from a series of samples, just like the distribution of sample means follows a normal distribution, this distribution of sample proportions follows this normal distribution, the mean of which is the same as the population proportion. Right, so our sampling distribution of proportions normally distributed with a mean equal to the population proportion and just like our means, the distribution of proportions is a function of the underlying population proportion. It has a standard deviation equal to the population proportion times one minus the population proportion divided by the sample size. Right, so we have, just like for means, we have these nice characteristics of the sample proportion which now we can use to make inference about the population.

Back to Table of Contents

Watch: Using Standard Error to Generate a Confidence Interval

Your goal throughout this hypothesis testing process is to be able to say something useful about a population. You have seen that statistical calculations based on sample data can be used to draw conclusions about the population, both for quantitative and categorical variables. Now you will see how it is possible to make probabilistic statements about the population. These statements will allow you to make judgments about population parameters.

In this video, Professor Anderson shows specifically how to calculate the upper and lower values of a confidence interval for a population parameter. He introduces the notion of Standard Error, which you can think of as akin to standard deviation. The difference is, Standard Error describes how far, on average, sample means (or sample proportions) are from the population mean (or proportion). Standard Error is a measure of a collection of samples. Standard deviation, on the other hand, deals with individual values in a sample and how far they might be from the mean of the sample. So it is more of a test of how tightly clustered the values are around this sample mean. For the purposes of generating a confidence interval, Standard Error is the measure you need to calculate.

For the example shown in this video, a confidence interval was used to estimate the proportion of British people who would vote against a referendum proposing that Britain exit the EU. You might communicate the results of the analysis in this way:

The sample of 3,533 people showed that 48% of respondents were in favor of remaining part of the EU.
As a result, I am 95% sure the population proportion is within 1.7 percentage points of 48%. That is, I am 95% sure the population proportion opposed to Brexit lies between 46.3% and 49.7%.

Transcript

So we've established, under certain conditions, that the sampling distribution for sample means will follow a normal distribution, the mean of which is the same as the population mean, and the standard deviation of which equals the standard deviation of population, divided by the square root of the sample size. So, what that means is that 95% of the time, the mean of any sort of random sample is going to be within two standard deviations of the true population mean. And so we can just flip that around. Right, so instead of the sample being within two means of the population mean, we can say that if I have a sample mean, plus or minus two standard deviations, that interval will contain the true population mean 95% of the time. Right, so that's a pretty cool insight.

We can talk about the population based upon the mean of the sample. If we wanted to sort of make a more confident estimate, we would go plus or minus three standard deviations and that would contain the true population mean 99% of the times. And this also holds for the proportions in the same way that it holds for means. Now unfortunately, when we talk about the distribution of the sample means, we don't have the standard deviation of the true population. So we approximate the standard deviation of the population with the standard deviation of the sample. And so because we're using this approximation, we'll refer to this as standard error.

Right, so the standard error for our estimate of the population mean equals the standard deviation of the sample divided by the square root of sample size. We're just substituting the standard deviation of the sample in for the population. In concert with this substitution, we're going to sort of replace the normal distribution with this other distribution called the T-distribution, but for all intents and purposes they have the same characteristics. So, let's look at an example. Say we're trying to sort of determine or estimate what fraction of the UK population is going to vote to leave the European Union. Right, so we have a sample of voters, right. 3,533 voters and of that sample, 48% of them has said they want to remain in the EU.

And so now, we're trying to make a definitive statement about the potential population mean of British voters. We know that that true population mean will be within plus or minus two standard errors of our estimates, right? So that's 48%, plus or minus 1.7 percent, where the 1.7 comes from 2 times .48, times 1 minus .48 divided by 3,533 and taking the square root of that. So we can take this sample, the characteristics of our normal, and now make some relatively refined estimates of the population. And we say estimates because they have these intervals over which we're fairly confident that the true population parameter will lie within.

Back to Table of Contents

Calculate Confidence Intervals

So long as you have the necessary summary statistics, you should be able make a statement about outcomes that involves a range of values and a confidence in that range. You have seen how this works for both quantitative variables (mean values) and categorical variables (proportions).

In this exercise, you will perform calculations and make inferences that are based on sample means and standard deviations. You should be able to find the formulae you need for this quiz in the Summary Stats and Testing Reference Tool.

You must achieve a score of 100% on this quiz to complete the course. You may take it as many times as needed to achieve that score.

Back to Table of Contents

Watch: Bringing Confidence Intervals Back into Hypothesis Testing

In this video, Professor Anderson works through the four-step hypothesis-testing process, from translation of a question to a null hypothesis, through the generation of a confidence interval, and finally to using that confidence interval to reject the null hypothesis.

Video Transcript

So we have a sample statistic and we can take this sample statistic and generate an interval for that population parameter. What we do with hypothesis testing now is make a more refined statement about the underlying population based upon that interval. So let's walk through an example and sort of see that in context. So a group of Air Force cadets sampled bags of Chips Ahoy chocolate chip cookies, right. And their goal was to determine or evaluate the claim on the bag that there's 1,000 chips in every bag, right. So each student took one of these bags and they broke the cookies apart and calculated the number of chocolate chips in each bag.

It turns out, across these 42 students, they had an average of 1,262 cookies, chips in each bag. And there was standard deviation across the 42 bags of 118 chocolate chips. And so based upon that sample average of 1,262 and that sample standard deviation of 118, using our normally distributed sampling distribution of the mean, we can generate this interval for the population mean, that interval is going to be 1,226 to 1,298. So we're 95% sure that the population mean lies within this interval of 1,226 to 1,298. Right, so now can we take this information and allow Nabisco to make a more refined statement about the amount of chips in each bag? Right, so instead of focusing on the interval, maybe they want to actually label. Right, what does it mean to label 1,000 chips in every bag? Right, are they at some level guaranteeing that there's a 1,000 chips in every bag?

Right, so we might structure a null and an alternative hypothesis around this, where our null hypothesis is that there is 1,000 chips in each bag and that our alternative hypothesis is that there's more than 1,000 chips in each bag. And so here, our null is equal to 1,000, our sample has a mean of 1,262. Right, so our sample is 262 chips away from the mean. Now our standard error for the estimate of the mean is equal to the sample standard deviation divided by the square root of the sample size. Right, so it turns out that our sample is 14 standard errors away from the null hypothesis. And remember that owing to our normal distribution that 99% of the time, the true population mean is within three standard deviations of the sample mean and so now all of a sudden we're 14 of these away from the null hypothesis. It means there's basically almost no chance that the mean of these, the average number of chips in a bag is 1,000. Right, so we can reject that null and we have strong support for this alternative hypothesis. And so we can sort of see this link between this confidence interval, the normal distribution, and how that plays a role in our hypothesis testing.

Back to Table of Contents

Watch: Understanding Tailed Tests

In this video, Professor Anderson revisits the Chips Ahoy example and explains how a test known as a right-tailed test is used. Later in the course, it will become important to know and recognize the difference between left-tailed tests, right-tailed tests, and two-tailed tests.

The "tail" refers to the area underneath the ends of a curve. The area underneath the curve translates to a specific probability depending on where it is "cut off" from the curve. The place where the tail is "cut off" is defined by the test statistic. This test statistic is a function of the null hypothesis, and the null hypothesis is always stated as an equality. For example, "the population mean equals five" or "the population proportions are equal."

In terms of hypothesis testing, your choice of tails is dependent on the alternative hypothesis, which is stated as an inequality. You use a left-tailed test when the alternative hypothesis is that the population value is less than the value stated in the null hypothesis. You use a right-tailed test when the population value is greater than the value in the null hypothesis. When no directionality is indicated by the alternative hypothesis, you use a two-tailed test.

Transcript

So in our Chips Ahoy example, we had a sample of 42 bags of chocolate chip cookies. And across those 42 bags, we determined the average number of chips per bag was 1,262. And so, our question was, is that number really bigger than 1,000? Right, so our null hypothesis is that the average number of chips in a bag is 1,000 and our alternative is that average number's bigger than 1,000. Now it turns out the standard deviation of the number of chips in a bag is 118. So given we have 42 bags, our standard error for the estimate of that mean is 18. And so our sample has 262 more chips than the null. That 262 turns out to be 14 standard errors away from the null. And so the last part of hypothesis testing is calculating the probability of getting a test statistic which is 14 standard errors away from the null. And so we do that with our sampling distribution.

Right, so in this case our sampling distribution is for the mean number of chips. And we're trying to calculate the probability of getting a test statistic that's at least as big as 14. Given that our alternative is that the mean number of chips is greater than 1,000, we're really looking at what we refer to as a one-tail test because we're really focused on these test statistics that are at least as big as 14 and that are averages bigger than the 1,000. So we're focused on this part of the curve that's to the right of our test statistic. This is all the sort of samples that would have a mean which was at least as big as 14 standard errors away from the null. And so the sum of this area under the curve is the probability of getting a sample at least as extreme as the one we observed.

Alternatively, if we had a different null hypothesis, I mean a different alternative hypothesis, where we were focused on, say the mean number of chips being less than 1,000, then we would focus on the left tail of our distribution or the other side of our sampling distribution. But we're focused on this area to the left of our test statistic. That would be a left-tailed test. In some instances, we're not going to differentiate whether we're less than or greater than. Our alternative is that we're simply not equal to the null. And in that case, we sum up both areas, the area to the right, and the area to the left, as these possibilities of getting test statistics at least as extreme as the one you observe when we're trying to calculate those p-values. So basically our p-values are areas under the sampling distribution, and which areas we pick is a function of our alternative hypothesis that we have selected.

Back to Table of Contents

Watch: Testing the Difference Between Two Groups

The mechanism of hypothesis testing can be adapted to situations in which you are trying to compare two distinct groups. Group comparison is actually a very common goal. It is essential for everything from A/B testing to randomized controlled trials.

In this video, Professor Anderson shows how to structure the null and alternative hypotheses for comparative testing of two populations (i.e. groups). Most commonly, the approach to comparative testing involves hypothesizing that there is no difference between the two populations. We test this hypothesis using differences observed in *samples*, and then we calculate the probability that the samples would show the observed differences if in fact there were no difference between the *populations*.

Transcript

All right, so a classic age-old question is whether or not you're going to lose more weight if you diet versus if you exercise, right? So obviously if you did both you'd probably lose the most amount of weight, but let's focus on under which sort of treatment, right, dieting versus exercising, do we lose the most weight, right. So here we're looking at a classical hypothesis test where we're going to compare the mean weight loss under two different treatments. Right, so we would have a sample of participants who are in our diet-only group, and ideally another random sample of participants who are in our exercise-only group. We would measure their weight loss, let's say our diet-only group lost just over seven kilograms, whereas our exercise-only group lost four kilograms.

Not everyone loses the same amount of weight, so there's a standard deviation of weight loss in each of these groups. These groups were around the same size, but they weren't exactly the same size. But as a function of the standard deviation in the sample size, we can calculate the standard errors of those mean weight losses for each of those groups. And those standard errors and those sample means now provide an estimate of the population means for people who are on a diet-only group, who are on a diet-only program versus those who are on an exercise-only program. And we can calculate the 95% confidence intervals for those population mean weight losses.

Right, so now we would have a confidence interval for how much people would lose if they dieted versus a confidence interval, how much people would lose if they exercised. And the degree to which those confidence intervals don't overlap, right, that basically means that these groups have different means. So obviously if the confidence intervals overlapped, then there is a chance that the population mean weight loss for diet is the same as the exercise group. But if the confidence intervals don't overlap, then that means that those population means are, in fact, distinct. And that's in essence what we do with hypothesis testing, when we're trying to compare two different groups. Right, at some level we calculate these intervals for the means and we sort of check to see whether or not they overlap or what's the chance that they would overlap.

Now it gets a little bit tricky when we do this, but, perhaps, the easiest way to think about this is let's not focus on the means of each group, let's actually focus on the mean differences. Right, so if we were to take the mean of the exercise-only group, and subtract the mean of the diet-only group from that, then if that mean is positive or negative, then there's a difference between those two means. So we might structure a hypothesis around that difference. An obvious null would be that the mean difference is equal to zero. And a two-tailed alternative hypothesis would be that that mean is not equal to zero. And so now we could go and do a simple hypothesis test. The null is that it's equal to zero, the alternative is that it's not equal to zero. We could look at the mean difference, right?

So, in this case the mean difference is just over three kilograms. And so the question is, is that three kilograms statistically significant? Well, we measure that based upon the variation in those two samples, and so we create this sort of aggregate variation, which is a measure of the standard deviations of both of those groups. Right? So we calculate the standard error for our test statistic and that standard error for that test statistic is a function of the standard deviations of both of those groups provided together. Once we do that, we can calculate how many standard errors away from the mean we are. Once we have that, we can then translate that into a p-value. So again, when we have two groups, it's really kind of the same as one group, we just have to a little massaging in order to get to that final test statistic.

Make a Data-driven Statement About the Hypothesis

You want to be able to say that an alternative hypothesis has or doesn't have statistical support from your data. In order to do this, you need to work through the hypothesis testing process from start to finish.

In this exercise, you will create a null and an alternative hypothesis based on a question outlined in a fictitious scenario. You will calculate the standard error based on summary statistics provided, and you will determine a confidence interval. Finally, you will either reject or accept the null hypothesis and determine an answer to the question. You should be able to find the formulae you need for this quiz in the Summary Stats and Testing Reference Tool.

Scenario:

A cleaning service firm recently began to outsource some of its workload. In an effort to see what impact outsourcing had on customer satisfaction, it sent a series of surveys to a randomly selected sample of recent clients. The sample of 100 clients gave an average customer satisfaction rating of 4.5 with a standard deviation of 1. The firm wants to know if this result is meaningfully different from their benchmark score of 4.0.

You must achieve a score of 100% on this quiz to complete the course. You may take it as many times as needed to

achieve that score.

Back to Table of Contents

Tool: Summary Stats and Testing Reference

Summary Stats and Testing Reference

Statistical Symbols

It can be challenging to keep track of which formula to use, how to test, and how to interpret the outcomes. Likewise, it can be difficult to keep track of the meaning of the symbols used in statistics formulas.

Use the tools provided here to remind yourself of formulae and testing procedures for hypothesis testing, and to keep clear the meanings of the symbols used in these formulae.

Back to Table of Contents

Module Wrapup: Test the Hypothesis

In this module, you continued to practice writing null and alternative hypotheses, drafting them as mathematically testable statements. You calculated standard errors, and you used those standard errors to calculate test statistics and confidence intervals. These numerical results enabled you to make a definitive statement about a null hypothesis, and you used that result to make a judgment about the population in question.

A working familiarity with all steps of hypothesis testing is useful whether you are primarily a consumer of data analysis or the person responsible for analyzing data on a regular basis. Now that you've had this experience, your understanding of statistical inference should allow you to question critically statistical results that are presented to you. As you continue your study of hypothesis testing, you'll be focusing on which statistical test is most appropriate to apply.

Back to Table of Contents

Module Introduction: Testing and Conclusions

Recall that the first crucial step in hypothesis testing is framing your question as a null and an alternative hypothesis. The other crucial step is selecting the appropriate test to determine the p-value that, hopefully, allows you to reject the null and support the alternative. The results of the test is always a number, and the usefulness of that number depends entirely on whether the correct test is chosen. Hypothesis testing tools are available in many statistics packages. In this course, we focus on functions available in Microsoft Excel, since Excel is widely available and has all the functionality you need to perform most statistical analyses.

As you work through this module, you will become familiar with each of several tests used in hypothesis testing. You are given tools to help you understand when to use a given test, and you will see what information you need to use the tests correctly. You will consider how changing your alternative hypothesis might give you a more refined picture.

Back to Table of Contents

Watch: Testing Sample Means

Quantitative data, such as the average amount of money a person spends in an online store, is evaluated using a test of means. If, for example, you want to demonstrate a significant change in a mean value as a result of some action, you need to structure your hypothesis test around an expected null value with an alternative value representing the result you are hoping to see. In the case of online store sales, you would hope to be able to provide support for a claim of increased sales.

After reviewing the four-step hypothesis testing process, Professor Anderson works through an example that illustrates a means test in action. He introduces the Excel function T.DIST that takes just two parameters to calculate a probability.

Transcript

All right, so we have our four-step hypothesis testing process. What I want to focus on now is the mechanics of that process. Right, so we have our underlying research question, we translate that to a null and an alternative hypothesis, we generate a sample, we check that that sample doesn't violate any of our assumptions, we calculate our sample statistics, then we calculate our test statistics, which is a function of the sample in the null. And then ultimately, we translate that test statistic to a p-value, and then decide whether or not to accept or reject our null, and then translate that back to our original research question. So, if we're going to look at a single mean, right, let's talk about a very specific test, so we're testing a single sample, and we're going to have some research questions around the mean of that sample.

Right, so we have our null and our alternative. Our test statistic is simply the difference from the null that our sample is, and then that distance is measured in standard errors. Now because we don't have the population standard deviation, we have to approximate that with the sample standard deviation, and so because of that approximation and owing potentially to a small sample, we calculate the p-value using what's called the T-distribution and not using the normal distribution, right? So they're very similar but we make these small adjustments for these sort of technical complexities. So, when we go to calculate these p-values, remember we have three different potential tails that we could be focused on: right, the left tail, the right tail, and then both of those for a two-tailed test. And so depending upon what our alternative is, we'll pick one of those tails and then in order to calculate those p-values, we have to use one of three different functions in our spreadsheet. We can use the T.DIST function to calculate a single, left-tail probability, we can calculate the T.DIST.RT function to calculate the right-tail p-value, and then we can use the T.DIST.2T function if we want to calculate the two-tail sort of p-value.

All right, so depending on upon what our alternative is dictates what type of tail we're focused on, and then that dictates what test we use. So let's go to an example. Right, suppose we've sort of looked at a new weight loss program. We have a sample of participants who have participated in our experiment. And we're trying to determine whether or not this was a successful weight loss program. So, across our sample of participants, we calculate their average weight loss and basically now we're trying to determine if that average weight loss is significant. So, here what do we have a classic case where are null hypothesis is that there's no weight loss, right, our equality, and that our alternative hypothesis, you know, a logical one would be that, you mean their weight loss is greater that zero.

So we would do a single tail, a right-tail test, because we're focused on what's the probability that they actually lost some weight, right. And so we would calculate our sample mean. We would calculate our sample standard deviation. We would translate that sample standard deviation to a standard error, dividing by the square root of the sample size, and then we would take and calculate the distance from the null of our sample, right? So our sample mean, minus zero, in this case, divided by that standard error tells us how many standard errors away from the mean our sample was. And now, we simply have to translate that test statistic to a p-value. So because we're focused on the right tail, we're going to use our T.DIST.RT function. We're going to give it that test statistic. And the last thing we have to tell our function is what's referred to as our degrees of freedom, which is simply our sample size minus one.

Right, because we have one less degrees of freedom than our sample size because if you have the test statistic, you already have one observation, right? So we would give it the test statistic, comma, our sample size minus one, and it returns our p-value.

Turns out that p-value is considerably less than 0.05, so we would reject that null hypothesis at the 0.05 level. We would use different functions if we were focused on different alternative hypotheses. Right, if it was our left tail again we would use just the T.DIST function and if we were doing two tails we would do the T.DIST.2T function.

Back to Table of Contents

Watch: Testing Sample Proportions

When testing a hypothesis related to categorical data, you are trying to conclude that a significant fraction, or proportion, of the population is within a certain category.

In this video, Professor Anderson works through an example that illustrates a proportions test in action. He introduces the Excel function NORMSDIST, which uses only the test statistic to calculate a p-value. Recall that a p-value is a probability used in hypothesis testing.

Transcript

So when we're focused on hypothesis tests for categorical or qualitative variables, we typically structure those hypothesis tests around proportions or fractions of outcomes falling within a category. So let's say our categorical variable of interest is the gender of newborns. That categorical variable has two categories, male or female. And then we would structure our hypothesis test around the fraction of those newborns that are male. All right, so for our hypothesis test, we're going to have this fraction or proportion from our sample. We're going to have some hypothesized population proportion and we're going to compare that sample to that hypothesized population proportion.

Our null hypothesis is going to be that the population proportion equals some null proportion and then we're going to have one of three alternatives. That alternative hypothesis might be a non-directional two tail test where the population proportion is not equal to the null proportion or it might be a one tail test, a right tail test when the population proportion is greater than the null proportion or left tail test when the population proportion is less than that null proportion. So, if we focus on how we calculate those p values, we're going to have this sample proportion. We're going to have the difference between that sample proportion and our null proportion. Then we're going to scale that difference by our standard error to calculate our test statistic. Now, when we calculate the standard error of that estimate, we're going to use the null proportion, right? The standard error is going to be the square root of the null proportion times 1 minus the null proportion divided by the sample size.

Now because we're using the null proportion, say versus the sample proportion, then we can still use the normal distribution when calculating p values or when calculating the probability of getting observations at least just extreme as the one that we've observed given the null is true. So in Excel we can use our NORMSDIST function, so NORMSDIST of our test statistic gives us the area to the left of that test statistic, or the probability of getting a value less than or equal to that test statistic. For a right tail test we want the area to the right of our test statistic, right? So 1 minus NORMSDIST of that test statistic will give us the p value associated with a right tail test. Then for a two tail test we're simply going to take one of those areas and multiply it by two to get the corresponding two tail p value.

So let's go through an illustration. So I have a sample. And then I have a series of observations in my category of interest, and then the remainder of my sample in the other category. As long as the smallest of those two numbers is at least as big as ten, we can proceed with calculating our p values. We're going to have our estimate for that proportion by the number of observations in our category of interest divided by our sample size. We have the difference between that estimate or that proportion and our null proportion. We take that difference and we divide by our standard error to get our test statistic. So for a left tail test that test statistic is going to be negative and the p value associated with that test statistic will simply be NORMSDIST of that test statistic. For a right tailed test that test statistic is going to be positive.

Right, because our alternative is that the proportion is greater than the null proportion. So one minus NORMSDIST of that test statistic will give us the area to the right of that test statistic. And then for our two tailed test, we don't necessarily know ahead of time When they're not that test statistic is positive or negative. So let's make it positive by taking the absolute value of the test statistic. So now I can calculate the right tail mass, so (1-NORMSDIST(ABS) of that test statistic gives us the p value associated with a right tail test. If I multiply that by 2, that'll give the p value associated with that two tail test.

Back to Table of Contents

Watch: Testing Two Independent Sample Means

The statistical functions you use for hypothesis testing are versatile. You may already have noticed that the same tools and approaches are being used over and over in different testing scenarios, with slight variations.

In this video, Professor Anderson works through an example in which a two-means test helps determine whether the rating of one dining option is significantly better than the rating of another dining option. You will encounter the Excel function T.DIST once again. This time, pay close attention to a nuance of how the "degree of freedom" value is determined.

Transcript

So, sometimes we're focused on hypothesis testing across samples versus just within a sample. For example, we could have two samples, and we're interested in comparing the means of those two samples. Hypothesis testing for comparing means across two samples is very similar to hypothesis testing for a single sample. We're still going to calculate a test statistic, that test statistic is the difference between our null and our sample divided by that standard error. When we're comparing two samples, all that's different is we're going to have the difference between the means of the two samples, and we're going to compare that to our null. For example, the null might be that there's no difference or that the difference is zero. So, the mean of one sample minus the mean of the other sample minus our null, and our null is zero, divided by our standard error is going to be our test statistic.

All we have to do now is when we calculate that standard error is realize that that's comprised of, or originates from, two samples. So, instead of calculating the standard error based upon the standard deviation of the single sample, we're going to calculate that standard error based upon the standard deviations of both of those samples. Right, sort of like a pooled standard error. And we make that calculation by taking the square root of the standard deviation from the first sample squared, divided by its size, plus the standard deviation from the second sample squared divided by its size. Right, and so that's our standard error for this pooled sample. We're still going to use our T.DIST function to calculate the actual p-value. The only difference now when we use that T.DIST function, we have to specify the degrees of freedom and that degrees of freedom will be for the smallest of the two samples, right? So our samples may not of the same size, so the degrees of freedom will be the smallest sample minus one.

Right, so let's look at an example. Your interested in dinning at a food truck. You have the choice between two different food trucks: The Hot Truck and Louie's Lunch. You're unsure of which one to dine at so you go over to Yelp and look at a sample of reviews. The Hot Truck has 40 reviews. Louie's Lunch has 31 reviews. Louie's Lunch gets a slightly higher average score than The Hot Truck. And so we may structure a hypothesis test around the sample of reviews for both food trucks to see in fact if the population looks differently upon Louie's Lunch versus The Hot Truck.

Right, so our null hypothesis might be that there's no differences between the mean review scores of these two different food trucks. An alternative might be that Louie's Lunch has a higher mean than The Hot Truck or that the difference between Louie's Lunch mean and The Hot Trucks mean is greater than zero. So, if that was our alternative we would take the mean of Louie's Lunch, subtract the mean of The Hot Truck minus zero, that would be the distance from the null, and we would calculate our standard error. The standard error would be the square root of the square of The Hot Truck standard deviation divided by its sample size, plus the square of Louie's Lunch standard deviation divided by its sample size. Right, so the square root of all that, that's our standard error. The difference from, the difference between the means minus the null divided by that standard error gives us our test statistic. And then, we would simply calculate the p-value of that test statistic using our T.DIST functions. T.DIST for the left tail, T.DIST.RT for the right tail, and T.DIST.2T for the two-tail null, the two-tail alternative hypothesis.

Back to Table of Contents

Watch: Testing Two Linked Sample Means

Sometimes the data you collect will generate two samples that have an important link, or dependence. For example, you might be keeping track of scores of different sports teams and trying to compare team performance based on those scores. In some cases, ignoring the link between samples can have a distorting effect on the conclusions you draw from a hypothesis test. For tests of means, you may want to make pairwise comparisons between a datum in one sample and a corresponding datum in the other sample. Consider a decision of which basketball teams will enter the playoffs. The playoffs are structured around a series of competitive events in which there is one winner. So it wouldn't make sense to add up the number of points each team scored over the whole season and let the teams with the most points enter the playoffs. Instead you need to look pairwise at the difference in team points scored in individual games and perform tests on the differences.

In this video, Professor Anderson works through an example in which a test of two means requires linking of the two population samples. The scenario involves daily revenues over the course of a month. As you can imagine, external events can shape the revenues on a daily basis. So a "fair" comparison requires a linking of the samples.

Transcript

So quite often, when we're looking at comparing samples, there's something that links the samples together. So the samples become, at some level, dependent versus independent. You could think of this as a before-and-after test. So now we're not just looking at these samples in isolation, but there's something that binds them together. When that happens, we need to sort of modify the tests we run in order to see whether or not we have statistically significant results. So let's do a simple example. So you're an online advertiser and you're looking at improving your efficiency in how you display ads. So we have a new strategy, we'll call that strategy B. And we're looking at the revenues generated by strategy B versus the revenues generated under our current strategy, we'll call that strategy A.

So, we're looking at daily revenues over a month. So, we're looking at a sample of 30 revenues for strategy A. And similarly a sample of 30 revenues for strategy B. So we could do a two-sample comparison of means where we are sort of basically doing a hypothesis test to see whether or not the mean of strategy B is different than the mean of strategy A. So our null hypothesis is that the difference between those means is zero. So we could take the differences of those means and then divide that by our standard error to get our test statistic.

In this instance, our standard error is a function of the standard deviation of both of those samples. When we run that specific hypothesis test, it turns out our p-value is greater than 0.05, so we don't reject the null that these differences are zero. Right, so we have no support for our alternative. But here we're losing the fact that we run these sort of comparisons on a daily basis. And so perhaps we should somehow link the fact that I ran both of these tests on this day and both of these tests on the next day, etc. So we're in essence linked across these two samples by these daily revenues. So, instead of just taking the average of each strategy, let's take the difference between the strategies. And then take the average of the difference and see whether or not the average of the difference is different than zero.

So our null is that the average of the difference equals zero, our alternative is that average of the difference is greater than zero, that strategy B is outperforming strategy A. And so now we have a single sample that we're running this test of means on and we run that hypothesis test, it turns out the p-value is less than 0.05, so we do reject the null that they're equal and we have support for the alternative that this new strategy is outperforming the old strategy.

The reason we reject the null when we do the single-sample test on the differences versus the two-sample test on the the individual two samples, is because there is a lot of variance, there's a lot of difference in revenues from one day to the next. And that difference in revenue from one day to the next is probably bigger than the differences in revenue from one strategy to the next. And so that dwarfs the impact of the different strategies. But we're not really focused on the differences in revenue from day to day. We're focused on the differences in revenue from strategy to strategy. And so that's why it makes sense for us to focus on the difference because we're sort of linked by the day. We have these dependent samples versus independent samples and so we need to sort of run our test correspondingly.

Back to Table of Contents

Watch: An Easier Way to Calculate p-values (using the T.TEST function)

So far, calculating the probability, or p-value, has involved a number of intermediate steps. While its useful to have solved some problems in this way, in some cases it isn't necessary to go through these intermediate steps.

In this video, Professor Anderson introduces the T.TEST function in Excel. If you are performing a test of means with two variables, this function allows you to specify data ranges, input a few test parameters, and get your p-value automatically without having to do any intermediate calculations.

Transcript

So when we're comparing the means across two different samples, there's a couple of different ways to calculate the p-value when we're running our hypothesis test. One of those is to take the difference, get the standard errors, calculate the test statistics, and use our T.DIST function. What I want to focus on now is a slightly easier method where we use some built-in functions in our spreadsheet to simply this process.

So let's revisit our food truck example where we're trying to decide if we should stay at The Hot Truck or dine at The Hot Truck or dine at Louie's Lunch. We have 40 reviews from The Hot Truck 31 reviews from Louie's Lunch. Louie's Lunch has a slightly higher average review score. So, we're trying to decide, based upon these two samples, whether or not we think the population of diners will look differently upon Louie's Lunch over The Hot Tuck. Right, so we're going to compare the means of these two samples in order to make some inference about the population.

So we can use our TTEST function in order to run those tests. Our TTEST function, the first two sets of entries, are simply our data. So TTEST, we grab the data for sample one, comma, we grab the data for the second sample. Right? So we just simply grab our data. Now we have to specify do we want a one-tail or a two-tail test. We put in a one for a one tail, a two for a two tail. We don't have to specify which tail under the one tail, just whether or not we're doing a one- or two-tail test. And then the fourth entry into our TTEST function is really to determine how the standard error is calculated.

Right, so we have to tell our spreadsheet whether or not we're doing an independent versus dependent sample test. So if we're doing a paired sample, or a dependent sample test, we're going to input a one. If we're doing an independent sample test, we're either going to input a two or a three. We'll typically input a three because that indicates that our samples come from populations with different variances. Right, so you see here, this function is pretty straightforward. All we really need is our two data vectors, and then you have to specify the details of the test you want to run. We don't have to make interim calculations for means and standards deviations and calculate standard errors, it's just giving these functions our data. So at some level, less steps, so there's less chance for you to make a mistake along the process.

Back to Table of Contents

Watch: Testing Two Sample Proportions

Often we want to compare a behavior, a preference, or another quality across two populations. For example, if you run a cafe and are interested in comparing the buying habits of morning coffee drinkers and afternoon coffee drinkers, you might be interested to know if there is a preference for decaf versus caffeinated coffee at these different times of the day. In this case, you have two groups: morning and afternoon coffee drinkers. For each group, you are looking at the choice of coffee, with decaf and caffeinated as the two categories.

In this video, Professor Anderson returns to proportions testing and shows how the NORMSDIST function can be used to test how a specific behavior may or may not be related to a specific demographic.

Transcript

We can use hypothesis testing to compare differences across categorical variables. For example, we're interested in whether or not males have a different brand preference than females, right? Do males prefer brand A versus brand B in a different proportion than females? So let's look at a British game show in an effort to illustrate how we might go about comparing these proportions across these categorical variables. In this specific game show, participants are trying to maximize their winnings. They do so by deciding whether or not they should split or steal the jackpot.

Right, so we have two participants, each with a choice of split or steal, and so we have these four potential outcomes.

If they both decide to split or cooperate, then they share the jackpot 50/50. If they both, though, decide to steal, then neither of them win anything. Now it gets tricky, though, when one splits and one steals. So if one individual decides to cooperate and split, and the other individual decides to steal the pot, the individual who stole the pot gets the entire jackpot and the individual who decided to cooperate or split gets nothing. And so now the issue is, well, do males or females tend to cooperate more or less often than the other gender. So let's look at a sample of outcomes from this game show. So we have about just over 500 playings of this game show. And we can look at whether or not males or females had a preference for splitting or stealing. Specifically, what fraction of the time did males decide to cooperate, or split, versus what fraction of the time did females decide to cooperate or split? Both of those proportions are a little bit more than 50%, and there's a small difference between that proportion for males versus females.

Now the question is, is that difference statistically significant, right? We're going to test that structurally with our hypothesis test. So basically a null might be that those proportions are equal. And an alternative would be that they're not equal. Right, so we're not doing a directional alternative we're just saying those proportions are not equal. And so we can simply take the difference between those proportions. But now, we have to scale that by the standard error in order to calculate that test statistic. Right, so our standard error calculation is pretty straightforward. It's a function of the proportions for both of those categories. But we have to make sure we use the right proportion consistent with our null. So our null is that the proportions are the same. And so we want to have the standard error calculated based upon that null hypothesis.

So if we go back to our data, we can calculate this thing which we call the pooled proportion, were basically if we look at the total number of individuals who choose to cooperate or split over the total number of times the game was played, so we're not differentiating males from females we're just looking at this aggregate proportion. So that's really, that aggregate proportion would be the one that represents the null then be equal. So we put that aggregate proportion into our standard error calculation and now we take our differences between the proportion. We scale it by that standard error to get our test statistic. Now we just have to calculate the probability of getting a test statistic at least as extreme as the one we've observed.

So because our alternative is that they are not equal, our alternative is a two-tail test. So in our NORMSDIST function we take the absolute value of that test statistic, One minus the NORMSDIST of that absolute value gets us the probability of getting a value, at least as extreme as that. For a right-tail test, we multiply that by two to get a two-tailed test and we see that that probability is quite large. Much bigger than .05, so we're not going to reject our null that those proportions are equal. We have no support for the alternative that males and females have a different tendency to cooperate.

Tool: Test Function Flowcharts

Means Test Flowchart

Proportions Test Flowchart

By now, you have been exposed to a handful of examples involving tests of means and proportions, and it may seem like they are running together into a confusing mess. At some point though, you will need to determine which test you need to perform on your data in order to get a valid result.

You can simplify your decision of which test to use by answering a few very simple questions. The first questions is: Are the data quantitative or qualitative (categorical)? Quantitative data indicate a test of means. Categorical data indicate a test of proportions is needed. From there, the questions you need to ask will differ.

Use the two tools linked on this page to help you determine which Excel function you will use to calculate a p-value that might allow you to reject your null hypothesis. You may wish to print them for future reference. Later in the module, Professor Anderson will give you a guided tour through these flowcharts.

Tool: Test Templates Workbook

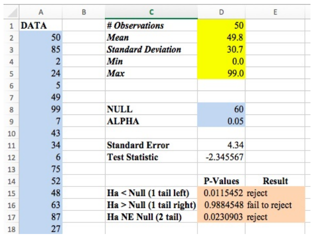
Test Templates Workbook

The attached Excel workbook includes seven sheets, or tabs, each of which has a layout useful for a different kind of statistical testing scenario. The sheets are named:

TEST for Mean
TEST for PROPORTION
TEST for 2 PROPORTIONS
TEST for 2 Means (Ind)
TEST for 2 Means (Dep)
TEST for Association (1 Var)
TEST for Association (2 Var)

Each of these sheets corresponds to one of the paths defined on the printable Means Test Flowchart and Proportions Test Flowchart. On the flowcharts you can find the place on each path where one of these seven sheets is mentioned. By selecting the appropriate sheet, pasting your data in the sheet, and modifying the parameters as needed, you can calculate the p-value for your hypothesis test.

On each sheet in the tool, you will see blue, yellow, and orange shaded cells. The blue areas are cells you will need to fill in, either with data or with other information you need to provide. The yellow cells are calculated values that will change depending on the values that are in the blue cells. The orange cells show results, specifically whether you can or cannot reject your null hypothesis for a given test. Detailed instructions for using each sheet are included on the sheets themselves.



	A	B	C	D	E
1	DATA		*# Observations*	50	
2	50		*Mean*	49.8	
3	85		*Standard Deviation*	30.7	
4	2		*Min*	0.0	
5	24		*Max*	99.0	
6	5				
7	49				
8	99		NULL	60	
9	7		ALPHA	0.05	
10	43				
11	34		Standard Error	4.34	
12	6		Test Statistic	-2.345567	
13	75				
14	52			P-Values	Result
15	48		Ha < Null (1 tail left)	0.0115452	reject
16	63		Ha > Null (1 tail right)	0.9884548	fail to reject
17	87		Ha NE Null (2 tail)	0.0230903	reject
18	27				

You will use this tool to complete the course project. You may also want to download it and keep it handy for use with future hypothesis tests.

Back to Table of Contents

Watch: Choosing a Test Function

Choosing the correct statistical test is essential to getting a valid result. Since all the tests give you is a numerical probability, you need to be confident that you have arrived at that number through the correct choices.

In this video, Professor Anderson leads you stepwise along the paths that are on the means and proportions flowcharts included in this course. You may find it helpful to have a printed version of the flowcharts in front of you as you watch this video.

Transcript

We can use hypothesis testing to compare differences across categorical variables. For example, we're interested in whether or not males have a different brand preference than females, right? Do males prefer brand A versus brand B in a different proportion than females? So let's look at a British game show in an effort to illustrate how we might go about comparing these proportions across these categorical variables. In this specific game show, participants are trying to maximize their winnings. They do so by deciding whether or not they should split or steal the jackpot. Right, so we have two participants, each with a choice of split or steal, and so we have these four potential outcomes. If they both decide to split or cooperate, then they share the jackpot 50/50. If they both, though, decide to steal, then neither of them win anything. Now it gets tricky, though, when one splits and one steals. So if one individual decides to cooperate and split, and the other individual decides to steal the pot, the individual who stole the pot gets the entire jackpot and the individual who decided to cooperate or split gets nothing. And so now the issue is, well, do males or females tend to cooperate more or less often than the other gender.

So let's look at a sample of outcomes from this game show. So we have about just over 500 playings of this game show. And we can look at whether or not males or females had a preference for splitting or stealing. Specifically, what fraction of the time did males decide to cooperate, or split, versus what fraction of the time did females decide to cooperate or split? Both of those proportions are a little bit more than 50%, and there's a small difference between that proportion for males versus females. Now the question is, is that difference statistically significant, right? We're going to test that structurally with our hypothesis test. So basically a null might be that those proportions are equal. And an alternative would be that they're not equal. Right, so we're not doing a directional alternative we're just saying those proportions are not equal. And so we can simply take the difference between those proportions.

But now, we have to scale that by the standard error in order to calculate that test statistic. Right, so our standard error calculation is pretty straightforward. It's a function of the proportions for both of those categories. But we have to make sure we use the right proportion consistent with our null. So our null is that the proportions are the same. And so we want to have the standard error calculated based upon that null hypothesis. So if we go back to our data, we can calculate this thing which we call the pooled proportion, were basically if we look at the total number of individuals who choose to cooperate or split over the total number of times the game was played, so we're not differentiating males from females we're just looking at this aggregate proportion. So that's really, that aggregate proportion would be the one that represents the null then be equal. So we put that aggregate proportion into our standard error calculation and now we take our differences between the proportion. We scale it by that standard error to get our test statistic.

Now we just have to calculate the probability of getting a test statistic at least as extreme as the one we've observed. So because our alternative is that they are not equal, our alternative is a two-tail test. So in our NORMSDIST function we take the absolute value of that test statistic, One minus the NORMSDIST of that absolute value gets us the probability of getting a value, at least as extreme as that. For a right-tail test, we multiply that by two to get a two-tailed test and we see that that probability is quite large. Much bigger than .05, so we're not going to reject our null that those proportions are equal. We have no support for the alternative that males and females have a different tendency to cooperate.

Back to Table of Contents

Course Project, Part Two—Test, Analyze, and Conclude

In the first part of the project, you stated a null and an alternative hypothesis that were related to a business question you wanted to answer. In this part of the course project, you will use data to determine whether you can reject your null hypothesis. The Test Templates Workbook is an essential part of your project, and you will need to turn this in as part of your project documentation.

Note: If you do not have data to test your hypothesis, contact your instructor. They will supply you with some substitute data you can use to complete your project. You may need to provide them with some information, including but not limited to the null and alternative hypotheses you defined in the first part of the project.

Completion of all parts of this project is a course requirement.

Instructions:

1. Open your saved course project document. (If needed, download it again now.)
2. Download the Test Templates Workbook
3. Complete Part Two.
4. Save your work.
5. Once you've finished, review the entire document, making any final additions or revisions, and then **submit it for instructor review using the Submit Assignment button on this page.**

Do not hesitate to contact your instructor if you have any questions about the project. You will add to this document as the course proceeds and will submit it to the course instructor at the end of the course.

Before you begin:

Before starting your work, please review the **rubric** (a list of evaluative criteria) for this assignment. Also review eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Consider Other Alternative Hypotheses

Instructions:

You are required to participate meaningfully in all discussions in this course.

Discussion topic:

By now, you should have completed the project for this course. It's time for you to reflect on what you've accomplished, and in particular to consider what other information you might be able to extract from your data. Create a post in which you:

Restate your research question along with your original null and alternative hypotheses.
Indicate if you were able to reject your null. If you did reject the null, say what you were able to infer about your research question.
List any other alternative hypotheses you might consider testing, and what you would hope to learn from testing with them.

Reply to at least one other student post indicating what your reaction would be as a consumer of the testing outcome. What other alternative hypotheses, if any, would you suggest to the original poster?

To participate in this discussion:

Use the **Reply** button to post a comment or reply to another comment. Please consider that this is a professional forum; courtesy and professional language and tone are expected. Before posting, please review eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Module Wrapup: Testing and Conclusions

In this module, you learned to reason through a plan for testing hypotheses using statistics. This is a great accomplishment. You may or may not feel completely confident in your ability to select the most appropriate statistical test, but the fact that you've put in the time and effort to consider the options gives you a deeper appreciation for what goes into statistical analyses.

We set out to focus on the connection between data and inference. You used your familiarity with a question of interest to create a null and an alternative hypothesis, and you wrote these hypotheses in testable, mathematical terms. You derived a test statistic from data, and you calculated a probability based on that test statistic. Based on that probability, you either did or did not reject your null hypothesis and support your alternative hypothesis.

Back to Table of Contents

Read: Thank You and Farewell

Chris Anderson
Associate Professor
School of Hotel Administration
Cornell University

Congratulations on completing *Scientific Decision Making*. I hope your work in this course has left you better prepared to think scientifically about the decisions you make and has given you the tools you need to make evidence-based arguments to support these decisions.

From all of us at Cornell University and eCornell, thank you for participating in this course.

Sincerely,

Chris Anderson