Slide 1 – Title Slide

Hello and welcome to Week 8, Part 2 of EGM101: Hypothesis testing. In this lesson, we'll learn more about inferential statistics, and how we can use hypothesis testing to make conclusions about our data.

Slide 2 – Inferential Statistics

At the very beginning of this section of the module, we introduced the definitions for descriptive statistics – what we use to to organize and summarize our data. In a workflow, this might be part of what makes up our observations about something.

Now, we're finally ready to start to introduce inferential statistics – how we can use our observations to draw conclusions. Formally, the process of inferential statistics, of drawing conclusions from "good" data, is known as hypothesis testing.

Once we have our observations, or even before we have our observations, we come up with a hypothesis – for example, that a particular treatment or medicine has had an effect on an illness, or that there are differences between different categories of our data. In hypothesis testing, we call this the "alternative hypothesis," and we will denote it as H_a. We are comparing this to the "null hypothesis" – the contradiction, or the logical negation, of the alternative hypothesis. So, for example, if our alternative hypothesis is "the treatment had an effect", the null hypothesis would be "the treatment made no difference." If the alternative hypothesis is "there are differences between different categories", the null hypothesis would be "there are no differences between the categories." Once we have formulated the null and alternative hypotheses, we have to select a statistical test, and conduct a statistical experiment, to test whether or not there is enough evidence to allow us to reject the null hypothesis in favor of the alternative hypothesis.

This is a subtle, but important point – our goal is not to "prove" that the alternative hypothesis is true or correct, because that's not something that we can do. Instead, we our goal is to falsify, or refute, the alternative hypothesis by attempting to refute the null hypothesis, based on the evidence that we have, and the test we have chosen.

Slide 3 – Choose your hypothesis

We will look at a number of examples of tests and hypotheses, but we'll start with tests of the population mean, denoted "mu" – we're testing whether our observations support our hypothesis about what the population mean is. If we only have one group of observations, we have a "one sample" test – that is, we're testing the observations against an assumed population mean, using a particular probability distribution. Later this week, we will look at other examples, such as testing differences between sample means.

When we perform a test of the population mean, our null hypothesis can usually be formulated like this: that the population mean is equal to some assumed value, mu naught.

There are a few different forms that the alternative hypothesis can take, depending on the kind of test we use. If all we're trying to do is test whether the population mean is different from mu naught, it's a two-tailed test – we're testing the probability that our observations are in either tail of the probability distribution.

If, instead, we're testing whether the mean is different in a particular direction, we use a one-tailed test. For example, if we're testing whether the mean is less than mu naught, the test is called "left-tailed", because we're testing the probability that our observations are in the left tail of the probability distribution. If we're testing whether the mean is greater than mu naught, the test is called "right-tailed", because we're testing the probability that our observations are in the right tail of the probability distribution.

Slide 4 – What are we "testing" anyway?

Hopefully, an example of how this might work will help. Let's say that we are looking at salmon lengths, and that we assume that these fish are sampled from a population with a mean length of 150 cm, and a standard deviation of 15 cm. We have a sample of 30 fish, with a sample mean of 142 cm, and a sample standard deviation of 20 cm. What we want to know is, do our observations support the hypothesis that the population mean is actually less than 150 cm?

To formally state the hypotheses, then, the null hypothesis is that the population mean is equal to 150 cm. The alternative hypothesis is that the population mean is less than 150 cm. Note that here, all we are doing is determining whether we have enough evidence to say that the population mean is less than 150 cm – we're not testing a particular value.

The question that we are attempting to answer with our statistical test is, "what is the probability of seeing observations at least this extreme, assuming that the null hypothesis is correct?" Or, in terms of conditional probability, what is the probability of our evidence, given that the null hypothesis is correct? In this example, we are calculating the probability that a sample of 30 fish, taken from a population with the given mean and standard deviation, would produce a sample mean of 142 cm, and a sample standard deviation of 20 cm.

The test procedure that we will use here is very similar to what we saw at the end of Week 7 – we will effectively calculate a z-score for our sample mean, using the population mean and the standard error of the mean. This comes directly out of the central limit theorem that we learned last week – from the central limit theorem, we know that the distribution of sample means is going to have a mean equal to the population mean, and a standard deviation equal to the standard error of the mean.

Plugging in the values that we have, we see that the z-score of this sample mean and sample size is –2.921. Now, the question is: what is the probability of getting a z-score this low? Using a calculator or computer, or an ancient table of z-scores, we find that the probability is equal to 0.002 – in other words, it's very low. On the plot of a normal distribution with the population mean and standard deviation equal to the standard error of the mean, the red dot here represents our sample mean – it's way out in the tail. In fact, because it is so low, this is strong evidence that the population mean is most likely less

than 150 cm. Formally, we would say that we reject the null hypothesis, in favor of the alternative hypothesis.

Slide 5 – Type I & Type II Errors, revisited

In the previous lesson, we looked at type I and type II errors in the context of statistical tests – a type I error, or a false positive, is where the test incorrectly says that the condition is present, and a type II error, or a false negative, is where the test incorrectly says that the condition is not present.

In the context of hypothesis testing, this is a little bit different. In hypothesis testing, a type I error or a false positive is where the test result leads us to incorrectly reject the null hypothesis. That is, the reality is that the null hypothesis is correct, but our test result leads us to incorrectly reject it. We use the lowercase Greek letter alpha to symbolize the probability of a false positive, or a type I error.

A type II error, then, is where the reality is that the null hypothesis is not correct, but the test result leads us to incorrectly accept, or fail to reject, it. The probability that we make a type II error is symbolized using the lowercase Greek letter beta.

These probabilities, alpha and beta, are what we use to make our decision. In order to reject the null hypothesis, we want the outcome of the test to be less likely than making a false positive – that is, the probability should be less than alpha.

To determine how good the test is, we have something called the “power of the test”, given by 1 minus beta. This tells us the probability of getting a true positive, where the outcome leads us to correctly reject the null hypothesis. We will talk about this a bit more later in the week, but we want the power of the test to be very high, and we want both alpha and beta to be low – we want to minimize the chances of making both a type I and a type II error.

And finally, for completeness, the probability of a true negative is given by 1 minus alpha. We don’t often use this value, though – typically, we’re more concerned with alpha and 1 minus beta.

Slide 6 – Statistical Significance

In the context of statistics, significant means a very specific thing. Most importantly, it does not mean that a result is important, useful, noteworthy, or newsworthy, as it often does in popular usage.

It especially does not mean that the alternative hypothesis is true – remember, this is not something that we can do with hypothesis testing. Instead, it means very specifically that the probability of seeing the evidence, given that the null hypothesis is correct, is less than or equal to the probability of committing a type I error, alpha.

The choice of what value of alpha, or what level of significance, that we use, depends on what field of study we are in. Some fields of study use a relatively low significance level, while others tend to use relatively higher levels. As you will see, a typical value of alpha is 0.05, meaning that we have a 1 in 20 chance of committing a type I error.

When choosing a significance level, we also need to consider what sample size we have, and the cost of a false positive. In the context of a criminal trial, the cost of a false positive might be very high – in committing a type I error, we would be sending an innocent person to prison or even worse punishment. In that context, we would hopefully choose a much lower value of alpha than in a situation where the stakes were not so high.

Most importantly, though, we choose the significance level before we conduct the experiment – we don't make the decision of where we draw the line after we've seen the test results.

Slide 7 – Critical values

In the context of the probability density function, we can define two regions – the first is the rejection region, the values of the test statistic that lead us to reject the null hypothesis. In this left-tailed example, these are values of t that fall within the blue shaded region on the plot. The blue-shaded region has an area, or a probability, equal to alpha – if our test statistic falls in this region, the probability of seeing that test statistic is less than or equal to alpha, and we would reject the null hypothesis.

The nonrejection region, then, are all of the values of the test statistic that lead to non-rejection of the null hypothesis – on the graph here, these are values in the non-shaded region under the curve. If our test statistic falls in this region, the probability of seeing that test statistic is greater than alpha, and we would not reject the null hypothesis.

The critical value is the value of the test statistic that separates the rejection and non-rejection regions. On this graph, which is again a left-tailed example, it's the location where the the blue line starts – alternatively, it's the value of the test statistic where the cumulative distribution function is equal to alpha. For a right-tailed test, it's the value of the test statistic where the cumulative distribution function equals 1 minus alpha, and for the two-tailed test, it's the values where the cumulative distribution function equals alpha over 2 and 1 minus alpha over 2.

We then have to compare the critical value for the chosen alpha to the test statistic. For the left-tailed example shown here, if the test statistic is less than or equal to the critical value, we reject the null hypothesis. For the right-tailed example, the test statistic is greater than or equal to the critical value of 1 minus alpha, and for the two-tailed example, the absolute value of the test statistic is greater than the critical value of 1 minus alpha divided by 2.

For the left-tailed example, if the test statistic is greater than the critical value, we do not reject the null hypothesis. For the right-tailed, we don't reject if the statistic is less than the critical value for 1 minus alpha, and for the two-tailed test, we compare the absolute value of the test statistic to the critical value for 1 minus alpha over 2.

Finally, the area under the probability density function in the rejection region is referred to as the p-value, which is something that we'll learn more about ...

Slide 8 – *p*-values

… right now. The p-value tells us the probability of seeing observations at least as extreme as ours, assuming that the null hypothesis is correct – written as conditional probability like this.

Instead of calculating a critical value and comparing it to our test statistic, we can also compare the p-value directly to alpha – in the examples to come, we'll see both ways of doing this.

Just like with many other things that we have seen, it's important to remember what the p-value tells us, and what it does not tell us. The p-value tells us how likely it is to see our observations, given that the null hypothesis is correct. When we compare it to our chosen significance level, it also tells us whether or not we can reject the null hypothesis. And that's it.

The p-value absolutely does not tell us any of the following things – it does not tell us the probability that the null hypothesis is true or correct; it does not tell us the probability that the alternative hypothesis is correct, given the available evidence, and it does not tell us whether or not the alternative hypothesis is false.

All it tells us is the probability of seeing our observations, given that the null hypothesis is correct.

Slide 9 – *p*-hacking

One important question to ask is, "what happens if we test lots of hypotheses?" When we do this, we will probably find some "significant" relationships or results, such as a correlation between the number of math doctorates awarded and the amount of uranium stored at US nuclear power plants, or the correlation between the divorce rate in Maine and the per capita consumption of margarine in the US. As we discussed when we first learned about correlation, these are examples of spurious correlations – they aren't a result of a causal relationship between the two variables.

The reason for this is easy to see – a p-value of 0.05 means that there's a 1 in 20 chance of seeing the same observations by random chance alone under the assumption that the null hypothesis is correct. As we increase the number of tests that we do, we're bound to see some seemingly significant results, purely by chance.

You may have heard this term before, but p-hacking is when we do lots of different analyses, or lots of different statistical tests, and select only the "best" ones – it's a way of finding "significant" results by chance. A better way to proceed, instead of trying lots of analyses to find a significant result, would be to split the data – use one part of the data for developing a hypothesis, and use the other part to actually test the significance. Alternatively, you could develop the hypothesis, then go out and collect new data to test the significance, though this may not always be possible. By separating the steps of hypothesis development and significance testing, though, we decrease the chances of finding significant results by chance alone.

Above all, though – like the correlation coefficient, remember that the p-value should not be the only evidence you have for a causal relationship – it should be one part of the case that you are making.

Slide 10 – Summary

In this lesson, we've discussed how hypothesis testing is the formal process of drawing conclusions from our data. We introduced the term "null hypothesis", which is the hypothesis that there is no effect or difference, and the alternative hypothesis, which is that there is some effect or difference.

We also discussed how it's important to remember that we are not proving that the null hypothesis is false, nor are we proving that the alternative hypothesis is true. Instead, we are determining whether or not the available evidence is good enough to reject the null hypothesis at a particular level of significance – in other words, we conclude that it is very unlikely that we would see our observations based purely on random chance.

Slide 11 – Additional resources

You can read more about the topics we've discussed here in the textbooks – Illowsky and Dean, Chapter 9; Caswell, Chapter 15; and Weiss, Chapter 9.

I've also included links to two articles about statistical inference and understanding significance levels in statistics that provide a bit more background and depth. There's also a video about hypothesis testing from Khan Academy. Finally, on the subject of p-hacking, there's a video from TED-Ed, and an interactive page from FiveThirtyEight which helps illustrate how p-hacking works.

That's all for this lesson – I hope you found it interesting, and if you have any questions, please don't hesitate to e-mail me or post in the discussion forum on blackboard. Bye!