Economics 250 — Hypothesis testing notes

Hypothesis testing

Sometimes what we care about isn't necessarily the precise value of an estimator, but rather whether it is *significant*: in other words, does a sample mean support or contradict some pre-existing idea about the mean.

For example, suppose I want a mean grade on the first assignment of 74. If I sample some of the tests, how can I use the sample to test whether the population mean is actually 74?

We call this idea we want to test ("the mean is 74") a "null hypothesis."

We denote it as:

$$H_0 : \mu = 74$$

This is something we can never absolutely prove from a sample. We **might** get a value very close to 74, but that doesn't prove the population is 74, it just gives us some evidence that it is close to 74. The population mean could be 76 for example, and we just happened to draw a sample with one exceptionally low score that gave us a sample mean of 74.

We also can't prove absolutely that that it is **not** 74: even if our sample had a mean of 56, there is still some small possibility that the sample happened to have some atypical, extreme values.

If we are willing to allow for a little possibility of making a mistake, what we can do is use the data to attempt to contradict the null hypothesis with some level of confidence. We call the contradictory statement the "alternative" hypothesis: it's simply the opposite (or contradiction) of the null hypothesis. So our alternative in the example is "the mean is not 74."

We denote it as:

$$H_a : \mu \neq 74$$

We typically write these two statements together:

$$H_0 : \mu = 74$$
$$H_a : \mu \neq 74$$

What we do next is calculate a "test statistic" that either rejects or fails to reject the null hypothesis, H_0.

The general idea is that if our sample provides evidence that the population parameter is sufficiently far away from 74, we can reject the idea that $\mu = 74$, that is, we can reject H_0.

The hypothesis above is called a "two-sided test" because we will reject if we see a value sufficiently far away from 74 on either side: both large values (e.g. 85) and small values (e.g. 60) provide evidence for rejecting $H_0 : \mu = 74$.

We can also do a "one-sided test" where we only reject on one side or the other. An example of such a test is testing whether the mean is at least 74. Then our hypotheses looks like this:

$$H_0 : \mu \geq 74$$
$$H_a : \mu < 74$$

We reject only if we find evidence from our sample that the mean is sufficiently below 74. Since larger values such as 85 don't contradict the statement "the mean is at least 74," we don't reject in this situation if we see large values, even if they are much larger than 74.

You can, of course, also perform the test in the other direction ($H_0 : \mu \leq 74, H_a : \mu > 74$), in which case you reject only for large sample means.

Hypothesis testing and confidence intervals

Let's think about the two-sided test ($H_0 : \mu = 74, H_a : \mu \neq 74$). Suppose we take a random sample of 25 midterms, calculate $\overline{x}$, then calculate a confidence interval (as we did in previous classes). We'll assume for now that σ is known to us.

Let's say we get: $[75, 79]$ for our 95% confidence interval. This would allow us to **reject** our null hypothesis with 95% confidence: if there's a 95% chance that this sample came from a population with mean between 75 and 79, that means the chance that this sample came from a population with mean 74 is *less than* 5%.

While confidence intervals are related to hypothesis testing, we usually don't use confidence intervals for this; instead we use test statistics.

Test statistics

We've already seen one test statistic before: A z-score is a test statistic.

What we want to test is how likely we are to see the value of $\overline{x}$ we got from our sample if the null hypothesis is true. For example, if we got $\overline{x} = 77$ from our sample, what we want to know is how likely we are to see a *sample* mean of $\overline{x} = 77$ when the *population* mean really is $\mu = 74$.

We can answer this using a z-score, where we use our $\overline{x}$ value, and the null hypothesis mean as μ:

$$z = \frac{\overline{x} - \mu}{s.d.(\overline{x})}$$

where $s.d.(\overline{x})$ is the standard deviation of our test statistic, $\overline{x}$. As we learned in previous classes this is: $s.d.(\overline{x}) = \frac{\sigma}{\sqrt{n}}$, so the z-score is:

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

Note, however, that some of the statistics we will see later involve a different denominator: they will have different standard deviation formulas.

We will also sometimes use *s.e.* instead of *s.d.* to indicate "standard error" instead of "standard deviation": the main difference is that standard error uses an estimate, such as s_x, instead of the population parameter, σ_x.

Suppose our sample mean is $\overline{x} = 77$, from a sample of $n = 25$, and we know $\sigma = 8$. If our hypothesis is:

$$H_0 : \mu = 74$$
$$H_a : \mu \neq 74$$

then we plug in the values to calculate z:

$$\begin{aligned} z &= \frac{77 - 74}{8/\sqrt{25}} \\ &= 1.88 \end{aligned}$$

p-values

A test statistic is just a number: what we want to do is convert that number into a probability that tells us how likely it is that a population with mean μ (i.e. a population where the null hypothesis is true) would produce a sample mean that gives us the value of our test-statistics.

We call this probability a "p-value". It is just the probability we're looking for: how likely would we be to get a sample value as large (or larger) than the one we found **if** the null hypothesis was true?

We already know how to find a number for this when we know the distribution of our test statistic: it's just the area under the distribution further out in the tails than the test statistic we found. For a z-statistic, we use the standard normal (from the textbook's "Table A").

(Later on we will see a t-statistic (for which we use "Table D"). The t-distribution is shaped similarly to a standard normal distribution (i.e. a z-distribution), but with tails that are a little bit larger.)

The one catch here is that we have to realize that, since we are doing a 2-tailed test, we are going to reject for very large **and** very small values. That is, for the $z = 1.88$ value above, our probability of getting a value **at least** as far away from 0 as 1.88 is the area to the right of 1.88 in the z-distribution, *plus* the area to the *left* of -1.88. The reason for this is that we

only care about how far away we are from 0: both 1.88 and -1.88 are a distance of 1.88 from 0, so we need to add up both tails.

So what we need to calculate is:

$$\begin{aligned} & P(z \leq -1.88) + P(z \geq 1.88) \\ &= P(z \leq -1.88) + [1 - P(z \leq 1.88)] \\ &= 0.0301 + (1 - 0.9699) \\ &= 0.0602 \end{aligned}$$

where the third line is just substituting the values from "Table A".

There is an easier way to calculate this, however: just look up the probability for the *negative* z-value (-1.88), then double it. (Since the z-distribution is symmetric, this works. t-distributions, which we'll see later, are also symmetric, so we can do this there as well).

Note: for one-tailed tests, we **do not** do this doubling of probabilities to find the p-values: there we only care about the probability of a test statistic in one tail, and so only calculate the probability for one tail.

Reject or not?

In the example above, our p-value was 0.0602. How do we use this to decide whether or not to reject the null hypothesis?

Before we can decide that, we need to decide on a rejection criterion. This is the answer to the question "how often are you willing to be wrong?" In other words, how often are we willing to make a mistake by rejecting H_0 when it is actually true?

We denote this criterion as α. It is most common to choose $\alpha = 0.05$ (that is we allow up to 5% chance of rejecting when we shouldn't), but it's important to realize that this choice is arbitrary. It was probably originally chosen because it's a round number that is fairly small but not too small, and then it stuck. We could also require more certainty by picking a smaller α (such as $\alpha = 0.01$), or allow more inaccuracy (but more rejection probability) by picking a larger α (such as $\alpha = 0.1$).

α is very closely related to our "confidence level": a confidence level is simply $(1-\alpha) \times 100\%$. So $\alpha = 0.05$ is the same thing as having 95% confidence, $\alpha = 0.01$ is the same thing as 99% confidence, etc.

Whether we reject or not is simple a matter of determining whether our p-value is above or below this α threshold. The general rule is that smaller p-values reject: a small p-value tells us that the probability of getting our test statistic when H_0 is true is very small, thus giving us evidence that H_0 is not true.

In the example above, 0.0602 is larger than $\alpha = 0.05$ so we would fail to reject at the $\alpha = 0.05$ level (alternatively: fail to reject with 95% confidence). We would similarly fail to reject at

smaller α levels, such as $\alpha = 0.025$. If we chose $\alpha = 0.1$, on the other hand, we would reject at the $\alpha = 0.1$ significance level.

A bigger α lets us reject more often, but gives us less confidence in that rejection. Another way to think about this is that smaller p-values give us stronger evidence that H_0 is false.