

Notes 13: Hypothesis Testing

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

1 Hypothesis Testing

You should have a decent basis on hypothesis testing from your pre-requisite courses. However, in my experience, this is the part of the material where students seem to have less basic knowledge. Somehow hypothesis testing doesn't really stick in their mind. This is not surprising. The topic is a bit hard conceptually, and unless you see the use, there is no reason why you would remember it. Freedman's book chapter 26 has a brilliant discussion of the topic. What is nice about it is that it applies just as well to a Stat introductory class as it does to this class. We will discuss these concepts in a slightly different manner, a little faster too. Remember that you have this extra resource to understand this topic more in depth.

And now we begin. Everything we did so far was about discovering the values of the coefficients in the linear model. Now we will step out of that world. Now we begin to do science. Remember in the beginning of the course how we discussed scientific questions? This is when we begin to answer them.

Hypothesis testing refers exactly to what the name implies: we make hypotheses, and then we test them. If we reject them, then we abandon them. Remember that we can never scientifically prove something right, but we can say that something has survived all tests we threw at it given the current technology.

So, first we need a hypothesis. Let's begin with a scientific question to guide us. Suppose that we wanted to know what is the effect of the mother's smoking during pregnancy on the baby's birth weight. Suppose that the world satisfies the model

$$\text{birthweight} = \beta_0 + \beta_1 \text{cigarettes} + \beta_2 \text{income} + \beta_3 \text{education} + \beta_4 \text{married} + u,$$

and $\mathbb{E}[u | \text{cigarettes}, \text{income}, \text{education}, \text{married}] = 0$. Moreover, suppose that assumptions 2 and 3 are also satisfied. We ran an OLS regression, and we got $\hat{\beta}_1 = -10$. We also calculated $SE(\hat{\beta}_1) = 8$. If we wanted, we could have calculated the confidence interval (check yourself that the 95% confidence interval is roughly $[-26, 6]$).

Now we begin to ask questions. The first and most natural question is: does smoking affect birth weight? If we look at $\hat{\beta}_1 = -10$, we are tempted to say that it does. However, remember that $\hat{\beta}_1$ doesn't get it exactly right. We should do this with care. Now that we have a question, we can write down the hypotheses, which are the mathematical expression of our question. There is always the **null hypothesis**, and the **alternative hypothesis**. The null hypothesis is always that which we want to test. The alternative is all the other possibilities. So, in our case:

We always denote the null hypothesis H_0 . The alternative hypothesis is usually denoted H_1 , but sometimes you can see it denoted H_A .

Let's just get the general setup here, just so that the formulas below will make sense. For a number a ,

Now we need to find a rule which will tell us whether we reject the null hypothesis or not. See the subtlety there? We can't "accept the null." We can only reject it, or not reject it. If we don't reject it, the hypothesis survives. It doesn't mean it's right, it just means we cannot rule it out with this technology.

Before we figure out the rule, we need to understand the risks. There are two kinds of mistakes we could do:

• **Type I error:** _____

• **Type II error:** _____

Ideally we would like to avoid committing both errors, but this cannot be managed. In fact, there is a trade-off between the errors. The more we protect ourselves from one, the more we expose ourselves to the other. It is actually quite easy to see this. If we fear

rejecting a hypothesis that is right, our rejection rule will be conservative, it will tend to reject less. But if we reject less overall, we may be not-rejecting when we should, that is, when H_0 is false!

We could adopt many strategies to deal with those errors. We could be looking for a rule which minimized the sum of both, for example, or a rule that minimized the maximum of both. All of these were studied, and the fights weren't pretty, but one strategy won the war. It is the most used of all, perhaps because it is so easy to apply. The fact is that it has been so long since its reign was established that now we understand it so well we just keep using it. The idea is to fix the probability of making a type I error, and then minimize the probability of making a type II error.

So, this is what we do. First, fix the probability of type I error. This has a name: significance.

- **Significance:** _____

The most used significance level is 0.05, but 0.01 and 0.1 are also used. Why? For the same reasons we are comfortable with confidence of 95%. I mean, if we consider 95% to be confidence enough, then we should be satisfied in making a mistake with about 0.05 probability.

All right, now for the rule. The rule is always written like this: _____

- $T(a)$ is a formula, which in our context always involves the OLS estimators, and the number a we are testing. This formula is called

- $c(\alpha)$ is a number, which always has to do with a probability distribution, such as normal, chi-square, etc., and it depends on the significance level α . This number is called

So, in the case of this particular test of whether $\beta_1 = a$, what is the test statistic, and what is the critical value?

- The test statistic is

this statistic is usually called a t -statistic. This name is given to statistics that are a ratio with the estimator in the numerator and the standard errors in the denominator.

- The critical value is

and it is exactly the same as in the confidence interval case.

So, a test has 3 components:

1. _____
2. _____
3. _____

If I ask you to setup a test, you must write down all the components. So, for example, if we want to test whether smoking has any effect on birth weight at 0.05 significance, the test is:

In our example,

$$\left| \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \right| = \frac{10}{8} = 1.25.$$

So, do we reject the null hypothesis or not? _____

In other words, we cannot rule out that smoking might not affect birth weight at all. Do you see the importance of the standard errors? If $SE(\hat{\beta}_1) = 4$, we would have ruled it out.

The null hypothesis $H_0 : \beta_1 = 0$ is very special. It is the most commonly used one. When we reject this null hypothesis, we say that $\hat{\beta}_1$ is _____
However, observe that we only say this when $H_0 : \beta_1 = 0$. It doesn't work with other null hypotheses, like for example, $H_0 : \beta_1 = 3$. In this case, we say simply that the null was rejected.

2 *p*-values

Another commonly used concept that you need to learn is the *p*-value. The *p*-values are particularly well liked in the medical literature, but they are common enough that you should know them anyway. We begin with the concept:

- **The *p*-value is:** _____

Wow, what? Don't expect to understand it just yet. Observe:

- At the 0.01 significance level, reject H_0 if

$$\left| \frac{\hat{\beta}_1 - a}{SE(\hat{\beta}_1)} \right| > 2.58$$

- At the 0.05 significance level, reject H_0 if

$$\left| \frac{\hat{\beta}_1 - a}{SE(\hat{\beta}_1)} \right| > 1.96$$

- At the 0.1 significance level, reject H_0 if

$$\left| \frac{\hat{\beta}_1 - a}{SE(\hat{\beta}_1)} \right| > 1.64$$

The larger the significance level, the easier it is to reject. There's got to be a threshold below which you don't reject, and above which you reject. Suppose that when you used

the formula in your sample, you got

$$\left| \frac{\hat{\beta}_1 - a}{SE(\hat{\beta}_1)} \right| = 1.8$$

We know that we would not reject the null at the significance levels 0.01, and 0.05, but we would reject it at the significance level 0.1. There is a threshold somewhere between 0.05 and 0.1 where we go from not rejecting to rejecting. That is the p -value. How do you find it? The p value is the α that solves the equation

So, the question is, 1.8 is which percentile of the standard normal distribution? If you go and search for this, either in a table, or in Stata, you will see that it is more or less the 96.5th percentile. So, $1 - \frac{\alpha}{2} = 0.965$ implies that $\alpha = 0.072$. That is exactly the p -value. It means that for a significance level below 0.072, we would not reject the null, for a significance level above, we would reject it. The general formula is:

$$p = 2 \left[1 - \Phi \left(\left| \frac{\hat{\beta}_1 - a}{SE(\hat{\beta}_1)} \right| \right) \right],$$

where Φ is the c.d.f. of the standard normal distribution. Should you memorize it? I don't recommend it. I don't know it by heart myself, and I don't think it will help you. You can always figure it out if you know what it means.

Now, why do we care about the p -value? The p -value has a lot more information than a simple test response. A test either rejects or it doesn't at the given significance level. When you report the p -value, you are giving more information. You are reporting the results of tests for any significance level. For example, if $p = 0.0001$, we would reject the null at any significance level larger than 0.0001. That means that we would reject it at the significance level 0.01, 0.05, and 0.1. We would most definitely reject it. If $p = 0.3$, the we would reject the test only if the significance was over 0.3. It means we wouldn't reject it at 0.1, 0.05, and 0.01. We just wouldn't reject it at all. You can keep playing this game. What does a p -value of 0.02 tell you?