Hypothesis Testing

Today, we are going to begin talking about the idea of *hypothesis testing*—how we can use statistics to show that our causal models are valid or invalid.

We normally talk about two types of hypothesis: the *null hypothesis* and the *research* or *alternative hypothesis*.

The null hypothesis

The hypothesis we actually test is called the null hypothesis. The null hypothesis means "there is no difference." For example, if we want to test whether or not the means of two variables (x and y) are equal, our null hypothesis is that they are equal. We'd write this as:

$$H_0 : \bar{x} = \bar{y}$$

H_0 is just a fancy way to write "the null hypothesis."

The research hypothesis

The other hypothesis is called the research or alternative hypothesis. This hypothesis means that there **is** a difference. For example, if we are testing the equality of two means, we would write the alternative hypothesis as:

$$H_A : \bar{x} \neq \bar{y}$$

Again, H_A is just another way to write "the research hypothesis."

An example

Last time, we thought about the IQs of political science majors at Ole Miss. Let's assume we found that the mean IQ of all political science majors, based on a sample, was $\mu_x = 107 \pm 8$. Now, imagine we found the mean IQ of the students in this class, $\bar{x}$. Let's test whether the mean IQ of the class was the same as the mean of all political science majors. What would our null and research hypotheses be?

$$H_0 : \quad \bar{x} = \mu_x$$

$$H_A : \quad \bar{x} \neq \mu_x$$

Why do we use the null?

Why do we test the null hypothesis, if it's usually what we want to disprove? The reason is that it's easier to disprove a specific hypothesis (x is not, in fact, equal to y) than it is to prove a non-specific hypothesis.

When we disprove the null hypothesis, we call that "rejecting the null" (or accepting the research hypothesis). Conversely, when we can't disprove the null (we don't find a statistical difference), we "fail to reject the null."

So, how do we decide whether or not to reject the null hypothesis?

A short lesson in rejection

It turns out we use the same concepts we use in confidence intervals: specifically, whether we reject the null depends on the *alpha level*. The alpha level is a statement of how confident we are that the null is correctly rejected; for example, when $\alpha = .05$, we are 95% sure we correctly rejected the null, but 5% of the time we may have actually rejected it when the null is actually true.

The other possibility is that we may accept (or fail to reject) the null hypothesis. If we fail to reject the null, that means we cannot prove our research hypothesis at a given alpha level—we aren't confident that the research hypothesis is valid.

Types of error

We have "Type I Error" when we reject the null when the null is in fact true. We have "Type II Error" when we fail to reject the null when the null is in fact false.

	H_0 is true	H_0 is false
Reject H_0	Type I Error	Correct rejection
Fail to reject H_0	Correct fail	Type II Error

The Z Test

Now, we can start talking about some actual statistical tests. For example, let's think about taking a (possibly non-random) sample from a population, and then trying to decide if our sample is typical of the population. This means we want to test $H_A : \mu \neq \bar{x}$.

This sort of test is called a "Z test." The formula for the Z test is pretty simple:

$$Z_{\text{ob}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Z Test Example

Let's revisit our example of Ole Miss students. Say the mean IQ of the class $\bar{x} = 115$, the mean IQ of all political science undergrads $\mu = 107$ with a standard deviation $\sigma = 14$, and we have 33 students in the class (n).

$$Z_{\text{ob}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{115 - 107}{14/\sqrt{33}} = \frac{8}{14/5.74} = \frac{8}{2.44} = 3.28$$

Comparing to the critical value

Now we have $Z_{\text{ob}} = 3.28$. We call this "Z obtained." To decide whether the Z test is statistically significant, we have to compare Z_{ob} to the critical value of Z for a given alpha level (Z_{crit}).

It turns out that we find Z_{crit} the same way we found critical values for confidence intervals. Specifically, we use $Z_{\alpha/2}$:

$$Z_{.05/2} = Z_{.025} = 1.96 \quad \text{and} \quad Z_{0.01/2} = Z_{.005} = 2.58$$

The final part of the test

So, now all we have to do is compare Z_{crit} to Z_{ob}:

$$H_0 : \text{is} \begin{cases} \text{rejected if } Z_{\text{crit}} \leq Z_{\text{ob}} \\ \text{accepted if } Z_{\text{crit}} > Z_{\text{ob}} \end{cases}$$

Of course, the converse applies to H_A; if we reject H_0, we accept H_A, and vice versa. In this example, we would reject the null at both $\alpha = .05$ and $\alpha = .01$, because 3.28 is greater than either critical value of Z.

The dependent sample t test

Often (like in the case of confidence intervals) we may know the population mean but be unaware of the population standard deviation. To solve this problem (again, testing whether $\mu = \bar{x}$), we use something called the *dependent sample t test*:

$$t_{\text{ob}} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

The formula looks a lot like the Z test, but since we only have the sample standard deviation we have to use the t distribution. Again, we use the same information we used to construct confidence intervals with the t distribution: an alpha level (usually .05 or .01) and a number of degrees of freedom (still $n - 1$).

Dependent sample t test example

For example, let's think about the clothing expenditures of Greek students relative to other students. We know (somehow) that the average male student spends \$50.00 a month on clothing, but we don't know the standard deviation. Now, imagine that we take a sample of 25 fraternity members, and find that the mean expenditure is \$60.00 with a sample standard deviation of \$20.00. Using this information, we can do a dependent sample t test, with an alpha level of .02:

$$t_{\text{ob}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\$60.00 - \$50.00}{\$20.00/\sqrt{25}} = \frac{\$10}{\$20/5} = \frac{\$10}{\$4} = 2.5$$

Compare t obtained with t critical

The next step, like the Z test, is to compare the obtained value of t with the critical value of t for the given alpha level and number of degrees of freedom. For $\alpha = .02$ and $df = 25 - 1 = 24$, the critical value is 2.492. And, again, all we do is compare:

$$H_0 : \text{is} \begin{cases} \text{rejected if } t_{\text{crit}} \leq t_{\text{ob}} \\ \text{accepted if } t_{\text{crit}} > t_{\text{ob}} \end{cases}$$

Since $2.492 \leq 2.5$, we reject the null hypothesis and conclude that frat guys do spend a different amount than non-frat guys on clothing.

The independent samples t test

This is all well and good, but what if we want to compare two samples from the same or different populations? For example, what if we want to compare the views of blacks on affirmative action to the views of whites, rather than the views of the population as a whole?

The solution to this problem is called the *independent samples* t *test*. The formula for the problem is pretty ugly, but bear in mind that there's nothing here you can't already do: it's just addition, subtraction, multiplication, division, and square roots. (Also bear in mind that it will be on a formula sheet on an exam!)

The formula in all its glory

The formula is generally divided into two parts:

$$t_{\text{ob}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_{\text{pooled}}(1/n_1 + 1/n_2)}}$$

$$s_{\text{pooled}} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The s_{pooled} part "weighs" the two sample standard deviations to produce a weighted standard error. The formula is much simpler when n_1 and n_2 (and s_1 and s_2) are equal, but this is rare.

How do we calculate this?

It's probably easiest to calculate the s_{pooled} part first, then plug the result into the larger formula. Let's say we obtain survey data on the number of social events attended by fraternity members and non-fraternity males in a given month, and want to see if frat guys go to more events than non-frat guys.

Frat Data

We find the following data (where group 1 is frat guys and group 2 is non-frat guys):

$$\bar{x}_1 = 15 \qquad \bar{x}_2 = 11$$

$$s_1 = 3 \qquad s_2 = 4$$

$$n_1 = 33 \qquad n_2 = 29$$

Let's decide whether the difference is statistically significant, with $\alpha = .01$.

Calculations

$$s_{\text{pooled}} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(33-1)3^2 + (29-1)4^2}{33+29-2} =$$

$$= \frac{(32)(9)+(28)(16)}{60} = \frac{288+448}{60} = 73660 = 12.266\overline{6}.$$

$$t_{\text{ob}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_{\text{pooled}}(1/n_1 + 1/n_2)}}$$

$$= \frac{15-11}{\sqrt{(12.266\overline{6})(1/33+1/29)}} = \frac{4}{\sqrt{(12.266\overline{6})(.0\overline{303}+.0344)}}$$

$$= \frac{4}{\sqrt{(12.266\overline{6})(.0648)}} = \frac{4}{\sqrt{0.794}} = \frac{4}{0.891} = 4.487$$

Determining t critical

Again, we need to compare the t-obtained (t_{ob}) with the critical value of t. We already know the alpha level (.01), as it was given in the problem; what we don't know is the number of degrees of freedom to use ($n - 1$ won't work, because we now have two n's!).

It turns out that we use $df = n_1 + n_2 - 2$ in this case, because we are dealing with two independent samples. (In terms of the free parameters problem, we'd only need $n_1 + n_2 - 2$ data points in addition to our statistics to figure out the data for both groups.) So, in this case, we use $29 + 33 - 2 = 60$ degrees of freedom. It turns out that, after we look it up in Appendix 2, $t_{\text{crit}} = 2.66$.

Determining the validity of our hypothesis

Now, we compare t_{ob} and t_{crit} the way we always do:

$$H_0 : \text{is} \begin{cases} \text{rejected if } t_{\text{crit}} \leq t_{\text{ob}} \\ \text{accepted if } t_{\text{crit}} > t_{\text{ob}} \end{cases}$$

Since $2.66 \leq 4.487$, we reject the null hypothesis and conclude that frat guys do have more fun.

The difference (matched or paired) t test

Now, let's imagine that instead of wanting to compare two groups, we want to compare the response of the same group of people in two situations. For example, we might want to compare voters' evaluations of Al Gore before and after he grew his beard, or compare the response of Bill the Cat to heavy metal and easy listening music.

This sort of comparison requires the use of the *difference* t *test* (or the matched or paired t test). This is the sort of test you'd use in an experiment, quasi-experiment, or panel study, where you want to examine the same subjects on different occasions.

The difference t test is a bit different in that we need to have the original data available to do the comparison; we can't use the mean and standard deviation of all the subjects from before and after the event. (Why not?)

The formula

The formula comes in three parts; luckily, they're simpler than the parts of the independent-samples t test:

$$t_{\text{ob}} = \frac{\bar{d}}{s_d/\sqrt{n}}, \quad \text{where} \quad \bar{d} = \frac{\sum d}{n}, \quad \text{and} \quad s_d = \sqrt{\frac{\sum d^2 - (\sum d)^2/n}{n-1}}$$

Note that the last two parts are really the sample mean and standard deviation formulas for d—we just put d in there instead of x. But what is d? It's simply the difference between the subject's "before" and "after" values. For example, if the subject gave Al Gore a rating of 60 when he was beardless and 40 when he had a beard, $d = 60 - 40 = 20$.

Al: Thanks for shaving

Here's some example data:

Individual	Rating of Gore w/o beard	Rating of Gore w/beard
1	80	60
2	85	87
3	17	21
4	18	27
5	50	40
6	10	10
7	98	54

Now, let's figure out whether there is a significant difference between people's attitudes towards Al Gore before and after he grew a beard, with a 95% confidence level ($\alpha = .05$).

Calculate d and d-squared

The first step is to find the mean and standard deviation of d. That will require us to know d, d^2, $\sum d$, and $\sum d^2$:

Individual	Rating of Gore w/o beard	Rating of Gore w/beard	d	d^2
1	80	60	+20	400
2	85	87	–2	4
3	17	21	–4	16
4	18	27	–9	81
5	50	40	+10	100
6	10	10	0	0
7	98	54	+44	1936
			+59	2537

Calculations

The mean difference is $\bar{d} = 59/7 = 8.42$ and the standard deviation of the differences is:

$$s_d = \sqrt{\frac{\sum d^2 - (\sum d)^2/n}{n-1}} = \sqrt{\frac{2537 - (59)^2/7}{6}} = \sqrt{\frac{2537 - 3481/7}{6}}$$
$$= \sqrt{\frac{2537 - 497.2857}{6}} = \sqrt{\frac{2039.7143}{6}} = \sqrt{339.95} = 18.438$$

Now, we can calculate the t_{ob} value:

$$t_{\text{ob}} = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{8.42}{18.438/\sqrt{7}} = \frac{8.42}{6.97} = 1.21$$

Determining the validity of our hypothesis

We use the same number of degrees of freedom we'd use for a dependent samples t test $(n-1)$; so, with $\alpha = .05$ and $df = 7-1 = 6$, $t_{\text{crit}} = 2.447$. Since $2.447 > 1.21$, we fail to reject the null hypothesis and thus conclude that Al Gore's beard did not significantly affect public opinion about him.

Homework

There was a lot of material in this chapter. I **strongly** recommend doing all of the practice exercises; however, I will not collect the assignment. If you have any questions on how to solve these problems, or want your work checked, see me.

You may want to read pages 155–56 about one-tailed test statistics. These are used when you want to test nulls involving comparisons, like $\mu \geq \bar{x}$. However, political scientists generally rely on two-tailed tests in our more sophisticated models.

The Road Ahead

Next time, we will begin discussing a technique called Analysis of Variance (ANOVA—Chapter 11). We will return to bivariate regression and correlation (Chapter 10) after we discuss ANOVAs.

Then, we will discuss how to interpret statistics presented in journal articles and briefly discuss the concepts behind other statistical techniques that are common in political science—multivariate regression and models with limited dependent variables (logit, probit)—with the goal of helping you understand the sorts of articles you might encounter in an upper-division course.