Econ 452 - A Review of Hypothesis Testing

Connor Cole

September 20, 2015

Introduction

This document is meant to provide a review of the basic framework, tools, and procedures of hypothesis testing. The material in here should be familiar to those with an intermediate probability theory class. Wooldridge's textbook has an excellent appendix (C.6 in my edition) that will be of use to you.

The Structure of Hypothesis Testing

Hypothesis testing is an organized method of statistically assessing the believability of various hypotheses about the world. To run a hypothesis test, first we need a hypothesis. Formally, we need two hypotheses. We need a null hypothesis, or some statement about the nature of the world that we would like to assess (e.g. 53% of Americans believe in ghosts, the mean level of hourly wages for full-time the U.S. is $13.00, etc.), and we also need an alternative hypothesis, or some statement about the world that could be true if null hypothesis were false (e.g. the share of Americans who believe in ghosts is not 53%, the mean level of hourly wages for full-time workers in the US. is not $13.00, the share of Americans who believe in ghosts is 42% etc.). Thus, the null and alternative hypotheses should be mutually exclusive. We generally denote the null hypothesis as H_0 and the alternative hypothesis as H_1.

Next, we need some form of a test that can help us assess the plausibility of the null hypothesis. Although there are many tests possible, econometricians tend to use tests based off of the central limit theorem, such as t-tests. To assess the utility of one test or another, statisticians look at type 1 and type 2 error. Type 1 error is a false rejection of the null hypothesis in the world where the null hypothesis is true. Type 2 error is a failure to reject the null hypothesis when the alternative hypothesis is true and, by extension, the null hypothesis is false.

The size of a test, usually denoted α, is the probability of a type 1 error, or $\alpha = P(\text{reject}|H_0 \text{ is true})$. The power of a test is the probability of rejecting the null hypothesis when the null hypothesis is false, or $\beta = P(\text{reject}|H_0 \text{ is false})$. Note that $\beta = 1 - P(\text{don't reject}|H_0 \text{ is false}) = 1 - P(\text{type 2 error})$.

While both type 1 and 2 error are 'bad', there is often a tradeoff in trying to minimize one type of error or the other. For example, if our 'test' were to simply reject the null hypothesis under all circumstances, our 'test' would result in a type 2 error of 0, but a potentially significant type 1 error.

T-Tests

As mentioned previously, there are many different kinds of tests we can to look at the plausibility of a null hypothesis. In this class, we will generally focus on t-tests.

We will discuss how to do t-tests as a practical matter in a bit, but it useful to first look at why they make sense. As you remember from your probability theory class, the central limit theorem implies that the asymptotic distribution of a mean of random variables from the same distribution, if scaled(multiplied) by $\sqrt{n}$, is normal with variance $V(x)$. Technically, we say that the scaled mean of random variables converges in distribution to a normal distribution. That is:

$$\sqrt{n}(\bar{X} - \mathbb{E}[X]) \xrightarrow{d} N(0, V(X))$$

Using simple algebra, we can restate this relationship as:

$$\sqrt{n}\frac{(\bar{X} - \mathbb{E}[X])}{\sqrt{V(X)}} \xrightarrow{d} N(0, 1)$$

For our purposes, it is sufficient to just remember the simplified and practical version of the central limit theorem that most applied researchers use - the sample mean scaled by the square root of n minus the expectation of X divided by the standard deviation of X has a normal distribution "as n grows large."

This powerful statistical fact enables us to test the believability of hypotheses about $\mathbb{E}[X]$, as we can test the probability of observing a certain specific value of $\sqrt{n}\frac{(\bar{X}-\mathbb{E}[X])}{\sqrt{V(X)}}$ for observed data. By convention, we refer to the value on the right hand side of the equation above, $\sqrt{n}\frac{(\bar{X}-\mathbb{E}[X])}{\sqrt{V(X)}}$, as a test-statistic. Under a null hypothesis that $\mathbb{E}[X] = \mu_0$, we know that the distribution of potential test statistics will become approximately normal "as n grows large." Hence, we can see where our particular test statistic falls under that null hypothesis, and calculate the probability of observing a test statistic as extreme as we observe under our null hypothesis, a quantity known as the p value.

Intuitively, smaller p values make it less believable that our null hypothesis is true, as smaller p values imply that it is less likely that we observe the test statistic that we see under the null. Note that it is still possible that the null is true, but only less probable. As a matter of convention, most statisticians regard p values smaller than 0.05 as sufficient evidence to reject the null. Technically

speaking, we would say at that level that we reject the null at the 5% confidence level. Note that this rejection rule implies that if the null were in fact true that we would still reject the null 5% of the time. Thus, the size of this particular test is 5%. A smaller p-value cut off point (referred to as a rejection rule in the literature) results in a smaller test size.

So far, we have been looking only at tests for the value of a mean for a single distribution. We might also want to test whether the values of the mean are the same for two different groups or variables. Some simple algebra shows that:

$$\frac{(\bar{X_1} - \mathbb{E}[X_1]) - (\bar{X_2} - \mathbb{E}[X_2])}{\sqrt{V(\bar{X_1} - \bar{X_2})}} \xrightarrow{d} N(0,1)$$

Thus, under the null hypothesis that the two means are equal, $\mathbb{E}[X_1] = \mathbb{E}[X_2]$.

$$\frac{\bar{X_1} - \bar{X_2}}{\sqrt{V(\bar{X_1} - \bar{X_2})}} \xrightarrow{d} N(0,1)$$

As before, if we calculate this test statistic we can see where our particular test statistic falls under that null hypothesis of equal means, and calculate the probability of observing a test statistic as extreme as we see.

Thus as a practical matter, there is a straight-forward progression in how to run a t-test:

1. Specify alternative and null hypotheses
2. If using a t-test, calculate the test statistic:
   - If testing the null hypothesis that the mean of a variable is some value μ, then use test statistic $T_n = \frac{\bar{X} - \mu}{\sqrt{\frac{V(X)}{n}}}$
   - If testing the null hypothesis that the mean of two variables is the same, or is the mean of a variable is the same for two subgroups, use the test statistic $T_n = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{\frac{V(X_1)}{n_1} + \frac{V(X_2)}{n_2}}}$
3. Look up a normal distribution table, or use statistical computing software to calculate the probability of observing a test statistic as extreme as the statistic you observe. Thus, calculate $P(|\epsilon| > T_n)$ where ϵ is a standard normal distributed variable.
4. State the p-value and reject the null hypothesis with 5% confidence if the p-value is less than 0.05.

Suppose we are measuring output per hour at a firm with 30 different observations, and the variance of output per hour is known to be $\sigma^2_{Output} = 10$. We have a null hypothesis that mean

output per hour is 20, but we observe a mean of 25 per hour. Applying a t-test since we are judging the value of a mean, we state the null hypothesis as the belief that output is 20 units per hour, and leave as an alternative hypothesis the claim that output per hour is not equal to 20. Then, calculating our test statistic, we see:

$$T_n = \frac{25 - 20}{\sqrt{\frac{\sigma^2_{Output}}{30}}} = \frac{5}{\sqrt{\frac{10}{30}}} = 8.7$$

Next looking at the normal distribution, we see that observing a value as extreme as 8.7 or greater occurs with probability approximately equal to 0. Thus, we reject our former hypothesis with a high degree of confidence.

Confidence Intervals

Note that our 5% confidence level implies that we expect T_n to fall within approximately -1.98 and 1.98 95% of the time. If we are looking at a t-test where we test the hypothesis that the mean of a distribution is some value μ, then we can rearrange this statement as follows:

$$P(-1.98 < \frac{\bar{X} - \mu}{\sqrt{\frac{V(X)}{n}}} < 1.98) = .95$$

$$P(-\bar{X} - 1.98\sqrt{\frac{V(X)}{n}} < -\mu < 1.98\sqrt{\frac{V(X)}{n}} - \bar{X}) = .95$$

$$P(\bar{X} - 1.98\sqrt{\frac{V(X)}{n}} < \mu < \bar{X} + 1.98\sqrt{\frac{V(X)}{n}}) = .95$$

Technically speaking, we can interpret this statement as an interval estimate for μ where the interval estimate $[\bar{X} - 1.98\sqrt{\frac{V(X)}{n}}, \bar{X} + 1.98\sqrt{\frac{V(X)}{n}}]$ will contain μ with 95% probability before the sample is drawn. Note that this statement does not imply that our interval contains μ with 95% probability after the sample is drawn. Either μ is or isn't in this interval after the sample is drawn. Thus, 95% confidence intervals are a statement about the probability of μ being within in an interval before the sample is drawn.

Chi-Squared Tests

While t-tests work well when either testing some hypothesis about the value of the mean or comparing mean values of some continuous variable across subgroups, they don't work when comparing two categorical values with each other. If we would like to assess the way two categorical variables

Table 1: Observed Values

	Below Median Income	Below Median Income	ROW TOTAL
Black	1170	642	1812
Hispanic	758	713	1471
Mixed Race	32	37	69
Non-Black and Non-Hispanic	1494	1989	3483
COLUMN TOTAL	3454	3381	6835

Table 2: Predicted Values

	Below Median Income	Below Median Income	ROW TOTAL
Black	915.6	896.3	1812
Hispanic	743.35	727.6	1471
Mixed Race	34.9	34.1	69
Non-Black and Non-Hispanic	1760.1	1722.9	3483
COLUMN TOTAL	3454	3381	6835

relate to each other, we use chi-squared tests. Let's say we had data on race and income status. Chi-squared tests allow us to test the null hypothesis that the income status category is unrelated to race against the alternative hypothesis that there exists some form of association between the two. Technically, this test works by comparing the observed number of observations in some category with the 'expected' number of observations, where the expected number of observations are computed by using the total number of observations in a row multiplied by the total number of observations in a column divided by the total number of observations in the data to predict for each category combination the number of observations found in that cell under the null hypothesis that the percentages of observations in each category would be equal. To see how this test works in practice, consider the categorical values in Table 1.

Using the previously described method, we would predict for the Black and Below Median Income categories $\frac{1812 \cdot 3454}{6835} = 915$ observations. Similarly, we would predict for other observations the predicted values in Table 2. Note that the row totals and column totals in Table 2 are unaffected, and that we are only 'predicting' cell values.

Next, we compute the following sum across all cells:

$$T_n = \sum_{i=1}^{n} \frac{(Observed - Expected)^2}{Expected}$$

For this data, we acquire a test statistic of 225.181 with 4 degrees of freedom. Under the null of no association between group values, we would expect to see an outcome as extreme as this or greater approximately 0% of the time. Hence, we reject the null hypothesis of no association between the categorical variables with a high degree of confidence.