

Economics 406 – Fall 2009
Professor Jeffrey Smith
Lecture: Review C
Version of September 14, 2009

Populations, Parameters and Random Sampling

We now turn to consideration of statistical inference.

This means obtaining estimates of population parameters by drawing samples, and testing hypotheses about population parameters and constructing statistics from them.

Taking a sample requires defining a population or universe of interest.

A sample can be drawn from a population in different ways. These include (1) random sampling; (2) stratified sampling; (3) network sampling and (4) importance sampling.

Explain each type.

Random sampling is the simplest. A random sample consists of independent observations drawn from a common distribution.

In addition to random sampling, throughout most of the course, we will also assume that the observations are “i.i.d.” for independently and identically distributed. These are separate assumptions; either one can be true without the other.

Example: mean height in the class. Explain the population and the parameter of interest. Explain about random sampling with $n = 5$ using the sample mean as the estimator.

Classical and Bayesian Statistics

There are two main schools of statistics, the classical and the Bayesian.

The main difference between the two is that Bayesian statistics formally incorporates prior beliefs or knowledge about the values of parameters. The data are then employed to “update” the prior, producing a (tighter) posterior distribution of beliefs about the parameter value.

The classical, or frequentist, approach treats each data set in isolation and leaves the integration of new and old evidence as an exercise, with the exception of formal meta-analysis.

Bayesian approaches tend to be more analytically and computationally demanding.

However, in practice, most economists in practice are what I call “casual Bayesians”. They do not employ the formal apparatus but they reason like Bayesians. They treat classical statistical evidence in the form of hypothesis tests and so on as providing useful information that they use to update their informal beliefs.

A very interesting book on this that I read in graduate school is called:

Howson, Colin and Peter Urbach. 1993. *Scientific Reasoning: The Bayesian Approach*, 2nd Edition. Open Court Publishing Company.

This book outlines the “casual Bayesian” view I have just described, without using the term.

In this course, we will focus on the classical approach, as it dominates the literature, but with a sprinkling of casual Bayesian intuition on top.

Finite Sample Properties of Estimators

Estimators and estimates

An estimator is a rule for constructing estimates.

Formally, an estimator W is defined as $W = h(Y_1, Y_2, \dots, Y_n)$.

Estimators have statistical properties, which we will turn to shortly.

There are many possible estimators of any given parameter. We will discuss ways of coming up with estimators later on.

An estimate is a number that is obtained when an estimator is applied to a particular data set.

Estimates are random variables. The distribution of estimates that results from applying an estimator to repeated, independent samples is called the sampling distribution of the estimator.

Example: mean height in the class.

Estimator 1: sample mean of five randomly chosen members of the class

Estimator 2: sample mean of 10 randomly chosen members of the class

Estimator 3: sample mean of five randomly chosen members of the back row

Estimator 4: 5’11”

Unbiasedness

An estimator is unbiased if $E(W) = \theta$, where θ is the population parameter of interest.

A particular estimate from an unbiased estimator may be very far from the population parameter of interest. Explain why: the unbiasedness is a property of the rule, and describes how the estimates compare to the population value on average.

The sample mean of a random sample is an unbiased estimator of the population mean.

Show this if time.

The sample variance, defined as

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2,$$

is an unbiased estimator of the population variance.

Unbiasedness is a desirable property of estimators, but it is not the only desirable property and in some cases we may be willing to trade of a bit of bias for other desirable characteristics.

The sampling variance of estimators

The sampling variance of the estimator is just that, its variance in repeated, independent samples.

In formal terms $\text{var}(W) = E[(W - E(W))^2]$. Note that the variance indicates variation about the expected value of the estimator.

For example, we showed before that $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$.

In general, unbiased estimators with a lower variance are preferred to unbiased estimators with a higher variance. It is good to get close to the correct answer (the population parameter) on average.

Draw the picture.

Efficiency

Consider two unbiased estimators of some population parameter θ , denoted W_1 and W_2 .

We say (or people who say these sorts of things say) that W_1 is more efficient than W_2 if

$$\text{var}(W_1) \leq \text{var}(W_2)$$

for all values of θ , with strict inequality for at least one value of θ .

All this says is that one unbiased estimator is more efficient than another if its variance always not higher and at least sometimes lower.

Be clear that statistical efficiency is not the same thing as economic efficiency.

Mean squared error

Draw the picture of one estimator with no bias and big variance and another with some bias but a small variance.

Ask: which is better?

The general point here is that sometimes you may be willing to trade off some bias for some variance.

This tradeoff is captured in the notion of mean squared error.

The mean squared error (MSE) of an estimator is given by

$$MSE(W) = E[(W - \theta)^2] = \text{var}(W) + [\text{bias}(W)]^2.$$

Explain the middle term in the formula.

Explain the third term in the formula. Note that when an estimator is unbiased, the MSE is just the variance.

In the econometrics literature, estimators that minimize the MSE are often preferred.

Asymptotic or large sample properties of estimators

Consistency

Most reasonable estimators use all of the available data to construct their estimates, based on the intuition that more information is better than less information.

We would expect estimators that use all of the available data to get closer, on average, to the population parameter as the size of the sample increases.

Consistency formalizes this notion of doing better as the sample size increases.

Formally, let W_n be an estimator of θ based on a sample of size n . Then W_n is a consistent estimator if, for every $\varepsilon > 0$,

$$\Pr(|W_n - \theta| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Another notation for the same thing is

$$\text{plim}(W_n) = \theta.$$

Explain the definition of consistency in detail.

If time, give the example of an estimator that is biased but consistent:

$$W_n = \begin{cases} \bar{X}_n & \text{with probability } (n-1)/n; \\ n & \text{with probability } 1/n. \end{cases}$$

Show that

$$E(W) = \left(\frac{n-1}{n}\right)\mu + 1$$

so that it is biased, but $\text{plim}(W) = \mu$ so that it is consistent.

If time, given the example of an estimator that is unbiased but not consistent, such as using the first five observations to estimate the mean regardless of sample size.

Emphasize that these are just toy examples. The key point is that unbiasedness is a property that is independent of sample size, whereas consistency is a property that describes what happens as the sample size gets large.

Probability limits

Probability limits have a number of useful properties:

$$\text{plim}(T_n + U_n) = \text{plim}(T_n) + \text{plim}(U_n),$$

$$\text{plim}(T_n U_n) = \text{plim}(T_n) \text{plim}(U_n),$$

$$\text{plim}\left(\frac{T_n}{U_n}\right) = \frac{\text{plim}(T_n)}{\text{plim}(U_n)},$$

where T_n and U_n are estimators.

Note that only the first of the three properties holds for the expected value operator, which suggests (correctly) that many consistent estimators will not be unbiased.

Asymptotic normality

We would also like to know, for purposes of hypothesis testing and confidence interval construction, what the shape of the sampling distribution of an estimator looks like as the sample size gets large.

The limiting distribution of an estimator as the sample size gets large is called its asymptotic distribution.

Definition of asymptotic normality:

Let $\{Z_n : n = 1, 2, \dots\}$ be a sequence of random variables, such that for all numbers z ,

$$\Pr(Z_n < z) \rightarrow \Phi(z) \text{ as } n \rightarrow \infty.$$

Then Z_n is said to have an asymptotic standard normal distribution.

The central limit theorem states that means are (almost) always asymptotically normal, regardless of the distribution of the individual observations, so long as the variance of the individual observations is finite. Formally,

Let $\{Y_1, Y_2, \dots, Y_n\}$ be a random sample from a distribution with mean μ and finite variance σ^2 . Then

$$Z_n = \frac{\bar{Y}_n - \mu}{\sigma / \sqrt{n}}$$

has an asymptotic standard normal distribution.

This is a really strong result, because it holds even for really odd distributions of the underlying random variables.

This is a really important result, because it allows inference using normal distributions, which we know a lot about, in samples of reasonable size.

We can also write this as:

$$\sqrt{n} \frac{\bar{Y}_n - \mu}{\sigma} \square AN(0,1)$$

where the distribution is asymptotic.

Explain the role of the \sqrt{n} and briefly discuss rates of convergence.

Note that the sample analog with s in place of σ is also asymptotically normal.

Note that the central limit theorem, important though it is, is not necessary when the distribution of the underlying individual variables is normal, because of the properties of the normal distribution that we already discussed.

General approaches to parameter estimation

There are three important general methods for coming up with estimators that have some or all of the fine properties we have just discussed. These are the method of moments, the method of maximum likelihood and the method of least squares. We now consider each of these briefly in turn.

Method of moments

A moment is just an expectation. The mean is the first moment about the origin. The variance is the second moment about the mean.

This usage always confused me, but there it is. I think it comes from mechanics.

In the method of moments, sample moments replace population moments. Thus, the method of moments estimator of the population mean is the sample mean, the method of moments estimator of the population variance is the sample variance and so on.

Less trivially, the method of moments estimator of the correlation coefficient is the ratio of the sample covariance (a moment) to the product of the sample standard deviations (square roots of moments).

For the reasons we just discussed, because sample moments tend to be consistent estimators of population moments, method of moments estimators are usually consistent estimators of the corresponding parameters.

To see this, suppose you want to estimate a population parameter $\theta = g(\mu)$. The method of moments estimator just replaces the population moment μ , with the corresponding sample moment \bar{X} .

Of course, if the function $g(\mu)$ is non-linear, then the estimator will be consistent but not unbiased.

Method of maximum likelihood

In the method of maximum likelihood, a particular distribution is assumed for the random variables in question. Note that this is a different, and often stronger, sort of assumption than is made in the case of the method of moments.

Given the distributional assumption, the parameters of the distribution are chosen so that the probability of observing the sample actually observed is maximized.

To see this in the abstract, let $\{Y_1, Y_2, \dots, Y_n\}$ be a random sample from the population distribution $f(y; \theta)$.

Explain what this notation means, using the mean of variables from a normal distribution as the example.

The likelihood function is then

$$L(\theta; Y_1, Y_2, \dots, Y_n) = f(Y_1; \theta), f(Y_2; \theta), \dots, f(Y_n; \theta).$$

Explain the notation and why the likelihood is just the product of all the densities for independent observations.

This is just a calculus problem (see – there are reasons we make you take it!)

The maximum likelihood estimator is the one that maximizes the likelihood function, which is usually written in log form for simplicity.

In many contexts, the MLE is the minimum variance unbiased estimator, when the distributional assumptions are right.

If time, do an example with a binomial random variable and show that the maximum likelihood estimate of the population probability is the sample proportion of ones.

Method of least squares

The method of least squares is another estimator based on solving calculus problems. In this case, the calculus problem is the sum of squared differences between the estimator and the data.

For example, the least squares estimator of the sample mean, call it m , solves the calculus problem:

$$\min_m \sum_{i=1}^n (Y_i - m)^2.$$

Applications to linear regression

We will spend most of the semester working with the so-called ordinary least squares (OLS) estimator of the parameters of a regression equation.

Foreshadow the problem:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Choose $(\hat{\beta}_0, \hat{\beta}_1)$ to minimize $\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$.

This is the least squares calculus problem. We will show the estimators that solve this problem and spend the bulk of the semester studying their properties.

It turns out that the OLS estimator is also a method of moments condition, which can be obtained by imposing sample analogues of the conditions that

$$E(x_i \varepsilon_i) = 0 \text{ and } E(\varepsilon_i) = 0.$$

If we assume that ε_i has a normal distribution, then it also turns out that the OLS estimator is the maximum likelihood estimator.

Life is good.

Interval estimation and confidence intervals

The nature of interval estimation

Up to this point we have been talking about point estimates – single numbers that provide an estimate of a particular parameter.

In some cases, we might like instead to construct intervals that contain the true parameter with some fixed probability in repeated samples.

We now consider how to construct such intervals, which we will call “confidence intervals.”

Confidence intervals for the mean of a normally distributed random variable

Suppose we want to construct a 95 percent confidence interval for the mean of n normal random variables drawn from a $N(\mu, \sigma^2)$ distribution.

We know that $\bar{Y}_n \square N(\mu, \sigma^2 / n)$.

Review why we know this.

We know that $\frac{\bar{Y} - \mu}{\hat{\sigma} / \sqrt{n}} \square t_{n-1}$.

Explain that $\hat{\sigma} = s$ in the book's notation. This is the sample standard deviation, which of course is an estimator of the population standard deviation.

Why we know that the standardized mean has a t distribution

Recall that the last lecture argued that the ratio of a standard normal random variable to a chi-squared random variable divided by its degrees of freedom has a t distribution.

We know that $\frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \square N(0,1)$.

The sample variance is given by $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$.

Also, $\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \square \chi_{n-1}^2$

which is clear if you think about dividing the sample variance by σ^2 , and multiplying it by $(n-1)$, which produces a squared standard normal.

Then take the ratio of the standard normal and the square root of the chi-squared random variable divided by its degrees of freedom to yield:

$$\frac{\left(\frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}\right)}{\sqrt{\frac{(n-1)\hat{\sigma}^2}{\sigma^2} / (n-1)}} = \left(\frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}\right) \sqrt{\frac{\sigma^2}{\hat{\sigma}^2}} = \left(\frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}\right) \frac{\sigma}{\hat{\sigma}} = \frac{\bar{Y} - \mu}{\hat{\sigma} / \sqrt{n}}.$$

Back to the confidence intervals

By definition, $\Pr(-t_{n-1, \alpha/2} < t_{n-1} < t_{n-1, \alpha/2}) = 1 - \alpha$, where $t_{n-1, \alpha/2} = c$ in the book's notation.

Explain the notation and the probability statement. Symmetry of the distribution is important here.

Draw a picture of the distribution to make the point.

Plugging the standardized mean into the definitional probability statement gives:

$$(-t_{n-1,\alpha/2} < \frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} < t_{n-1,\alpha/2}) = 1 - \alpha .$$

Some algebra (go through the steps) yields:

$$(\bar{Y} - t_{n-1,\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{Y} + t_{n-1,\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}) = 1 - \alpha ,$$

which is what we wanted.

Explain how to operationalize this by choosing a specific value of α .

What the confidence interval means

Correctly interpreting confidence intervals seems harder in practice than constructing them.

The key point is that the population parameter is fixed and the confidence interval is random.

Thus, it makes no sense to say that “the probability that the parameter lies in the interval is 0.95”. The parameter is or is not in the interval you have constructed.

What does make sense to say is that “the probability that any particular interval constructed in this way contains the population parameter is 0.95”

Reinforce the difference between these two statements.

Asymptotic confidence intervals for non-normal populations

By the central limit theorem, means of random variables generally have normal distributions, regardless of the distribution of the individual random variables. As such, once the sample size is “large enough” for the CLT to apply, an approximate 95 percent confidence interval is given by:

$$[\bar{y} \pm 1.96\hat{\sigma}_{\bar{y}}] .$$

Explain where the 1.96 comes from.

Explain the $\hat{\sigma}$ notation.

The formula for the confidence interval is sometimes written as

$$[\bar{y} \pm Z_{\alpha/2} \hat{\sigma}_{\bar{y}}]$$

for the more general case.

Explain what the $Z_{\alpha/2}$ indicates.

Note that we can do this for any parameter that has a normal distribution asymptotically, not just means.

Stata commands

The command `ci [varlist], level(#)` calculates a confidence interval around the mean of each of the variables specified in the variable list, for the confidence level specified in the command. For a 95 percent confidence interval, the level is set to 95.

The command `cii #obs #mean #sd, level(#)` is the command to do an “immediate” confidence interval in Stata. Stata’s immediate commands do not use the data in memory; they only use their arguments. This makes them very handy for quick calculations.

Hypothesis testing

Fundamentals of hypothesis testing

The basic intuition of hypothesis testing is to identify a test statistic that has a different distribution in the states of the world where the hypothesis of interest is true and is false. Then you can formally figure out the probability of observing the data in both these worlds.

Classical statistical tests privilege one particular hypothesis, called the null hypothesis. This is a key difference between classical and Bayesian statistics.

Call the null hypothesis H_0 .

As Wooldridge notes, it has the status of the defendant at an American criminal trial: you want to provide evidence that the null is false beyond a reasonable doubt.

The alternative hypothesis is denoted H_A . Typically, the null hypothesis is expressed as a point hypothesis and the alternative is expressed as an interval hypothesis, as in:

$$H_0 : \mu = 0$$

$$H_A : \mu \neq 0$$

The next step is to choose a test statistic. The test statistic is some function of the data that has a different distribution under the null than it has under the alternative.

In tests about population means, the test statistic is usually the standardized sample mean under the null.

After choosing a test statistic, you pick a significance level. The significance level gives the probability of a Type I error, which is the probability of (incorrectly) rejecting the null when it is in fact true.

The significance level is usually denoted by α .

Typical significance levels are 0.01, 0.05 and 0.10. Picking a significance level implicitly makes a statement about your tolerance for different types of mistakes.

The significance level, plus some additional reasoning, gives you a rejection region and, at the same time, a critical value or values. The rejection region is the region of possible values of the test statistic for which the null is rejected for the chosen significance level. The size of the rejection region depends on the significance level; the higher the level the bigger the rejection region.

In addition, the rejection region is chosen so as to include as large a fraction of the real line as possible, given the significance level. Explain how this leads to putting the rejection region in the tails.

Draw a picture showing the distribution of the test statistic under the null hypothesis along with the rejection region for different values of α .

The critical value is just the boundary of the rejection region. For a two-sided null hypothesis, there are two critical values, one the negative of the other. If the absolute value of the test statistic exceeds the critical value, then the null hypothesis is rejected.

Now consider Type II errors, which occur when you fail to reject the null hypothesis even though it is false. The probability of a Type II error depends on the distribution under the alternative hypothesis and the rejection region.

Draw a picture showing the distribution under the null, the rejection region and the distribution under the alternative. Show graphically the probability of making a Type II error. Emphasize again that there are two different worlds – the world of the null and the world of the alternative – and that we do not know which world we are in. Indeed, that is what the statistical test is trying to help us find out.

Make the point that the probability of a Type II error depends on the particular alternative hypothesis, of which there are infinitely many in the case of a composite alternative.

Update the picture to show how the probability of a Type II error declines as the alternative hypothesis becomes more different from the null.

The fact that the probability of a Type II error varies with the alternative hypothesis means that the power of the test varies with it as well. In particular, because

$$\text{Power} = 1 - \Pr(\text{Fail to Reject } H_0 \mid H_0 \text{ is False}),$$

the power of the test increases as the alternative hypothesis becomes more different from the null. This makes sense – it is easier to distinguish between the world of the null and the world of the alternative when the two worlds are quite different than when they are quite similar.

The last step is to use the data to calculate the test statistic and compare it to the critical value. If the value of the test statistic lies in the rejection region, then you reject the null hypothesis. If it does not, then you “fail to reject” the null hypothesis. Some would say that you “accept” the null hypothesis while others do not like that usage. I am among those who tend to avoid that usage. Notice again the asymmetrical treatment here of the null and the alternative.

At this point, you should also determine the p-value of the test, a topic we will return to shortly.

To reprise, there are two possible worlds. One is the world of the null hypothesis and the other is the world of the alternative hypothesis. The point of the test is to provide information about which of these worlds we are in. To this end, you choose a test statistic with a known distribution under the null and under the alternative. Using the distribution under the null and a chosen significance level, you determine a rejection region with associated critical values. You then compute the test statistic and compare its value to the critical value. If the test statistic exceeds the critical value in absolute value (for a two sided alternative) then you reject the null. Otherwise, you fail to reject the null.

Testing hypotheses about the mean in a normal population

Now we will go through the steps for an example null hypothesis regarding the population mean in a normally distributed population.

Suppose that we want to test the null that the population mean in some population equals 10. The implicit alternative is that the population mean does not equal 10.

We have a sample of size 121 of observations randomly drawn from a normal population. Explain that a normal population is one in which the random variable of interest has a normal distribution. The sample mean equals 12 and the sample standard deviation equals 10.

In terms of the usual notation, we have

$$H_0 : \mu = 10$$

$$H_A : \mu \neq 10$$

$$\bar{X} = 12$$

$$\hat{\sigma} = 10$$

$$n = 121$$

We choose a significance level $\alpha = 0.05$.

Under the null hypothesis $t = \frac{\bar{X} - \mu_0}{\hat{\sigma} / \sqrt{n}} \sim t_{120}$.

Under the alternative, the distribution has the same shape but is shifted along the real line, depending on the particular value selected for the population mean. If the population mean under the alternative exceeds that under the null, then the distribution is shifted to the right relative to the null; otherwise it is shifted to the left relative to the null.

Given the chosen significance level and the sample size, the critical value from Table G.2 in Wooldridge is 1.98 and the rejection region consists of all values of the test statistic greater than 1.98 in absolute value. In notation, the rejection region is

$$(-\infty, -1.98) \cup (1.98, \infty).$$

Now calculate the test statistic. We have

$$t = \frac{\bar{X} - \mu_0}{\hat{\sigma} / \sqrt{n}} = \frac{12 - 10}{10 / \sqrt{121}} = \frac{2}{10/11} = \frac{22}{10} = 2.2.$$

Comparing the value of the test statistic to the rejection region, we see that we barely reject the null for $\alpha = 0.05$.

Question: What would be the result of the test if $\bar{X} = 11$?

Question: What would be the result of the test if $\alpha = 0.01$?

Question: What would be the result of the test if $\alpha = 0.10$?

Performing t-tests of means in Stata

The `ttest` command in Stata performs tests of differences in means between groups and tests of a single mean against a constant.

The syntax in the first case, where the null is equal population means of two variables, is

```
ttest var1 var2
```

The syntax in the second case, where the null is that the population mean equals a particular value, is

```
ttest var1 #
```

You can also test the null that a given variable has the same population mean for two subgroups using the syntax

```
ttest var1, by(subgroup)
```

For more information, type “help ttest” in Stata.

Asymptotic tests for non-normal populations

Because of the Central Limit Theorem, standardized means of non-normal random variables have approximately normal distributions for large sample sizes.

Recall as well that the t distribution converges to the standard normal as the sample size increases. As a result, the testing apparatus just described for normally distributed observations also applies to non-normal observations, so long as the sample is sufficiently large.

In practice, a sample size of 30 is usually enough for approximate inferences and sample sizes above 60 will usually yield a very close approximation.

This illustrates the usefulness of the result embodied in the Central Limit Theorem.

Using p-values

Reporting the results of a test only in terms of whether or not the null hypothesis is rejected or not rejected throws a way a lot of information regarding the strength or weakness of the result. For example, if we are doing a test with a critical value of 1.96 and our test statistic equals 2.04, this is a much different situation in terms of the strength of the rejection than when our test statistic equals 4.26. Reporting only that the null is rejected equates these two results and omits the information that the second rejection is much stronger.

It is generally standard practice in economics to also report what is known as the “p-value”. The p-value can be described in a number of ways, all of which are substantively equivalent but which provide different intuition.

First, the p-value is the largest value of α for which the null would not be rejected. For example, in the standard two-tailed t-test with a large sample, if the test statistic equals 1.96 then the p-value equals 0.05, because that is just the margin of the rejection region. If we had chosen $\alpha = 0.10$ in this case, we would reject the null, because the observed test statistic of 1.96 is greater than the corresponding critical value of 1.65.

Similarly, again in the standard two-tailed t-test with a large sample, if the test equals 1.65, then the p-value equals 0.10, because we would reject the null for any larger value of α .

Draw the picture and show both cases.

Note how much additional information the p-value provides relative to just indicating rejection or non-rejection of the null. Once we know the p-value we know whether we would have rejected or not rejected the null for any value of α . This is much more information.

The second way to think about the p-value is as the value of α for which the observed test statistic is the critical value. This follows from the fact that the p-value indicates the largest value of α for which we would fail to reject the null hypothesis.

The third way to think about the p-value is as the probability of observing a value of the test statistic as large, or larger, than the one we actually observed (in absolute value for a two-sided test) under the null.

Use a picture to illustrate this definition.

Note that smaller p-values imply stronger evidence against the null. A p-value of 0.04 and a p-value of 0.0001 both imply rejection when $\alpha = 0.05$ but in the second case the observed test statistic is incredibly unlikely under the null, while in the first case we expect a value of the test statistic at least as large as the one we obtained in four out of 100 samples. On the other side, p-values of 0.06 and 0.82 both lead to a failure to reject the null with $\alpha = 0.05$, but the first case indicates very marginal rejection, suggesting that with additional data we might reject. In either case the p-value provides very valuable information that is lost by simply reporting on whether the test rejects the null or not.

In general, I prefer to more or less put aside the rejection business altogether, as the cutoff value is arbitrary in any case, and focus on the strength of the evidence against the null as indicated by the p-value. As this is not the way that most researchers proceed, we will continue to think about rejecting and not rejecting in this course.

Computing p-values

Stata prints out the p-values associated with t-tests automatically with the `ttest` command. It also prints out p-values for tests of the null hypothesis that each population

coefficient equals zero when it estimates linear regression models with the `regress` command.

You can also figure out p-values by hand using the t-table in the back of the book, although some interpolation may be required. To see this, consider the following example. Suppose that we obtain a test statistic of 2.45 with a sample of 200 and that we are doing a two-sided test. This puts us in the last line of Table G.2. The table indicates that the p-value for a test statistic of 2.326 is 0.02 and that for a test-statistic of 2.576 is 0.01. As 2.45 is roughly midway between 2.326 and 2.576, the p-value in this case is approximately equal to 0.015.

The relationship between confidence intervals and hypothesis testing

As the similarity in the probability statements underlying their construction suggests, there are strong links between confidence intervals and hypothesis testing.

In particular, a $(1 - \alpha)$ percent confidence interval is informative about a two-sided test of a null hypothesis that the population mean equals a particular value with a significance level of α . We can reject the null that the mean in the population equals a particular value if the confidence interval does not contain that value. In particular, if the confidence interval does not contain zero we can reject the null that the population mean equals zero at the indicated level of significance.

Practical versus statistical significance

In conducting hypothesis tests, it is easy to become overly focused on whether or not we reject a particular null hypothesis. Such rejections indicate that a coefficient is statistically significant, or that a difference between a population mean and a particular value (or between two population means) is statistically significant.

In the real world, magnitudes matter as well as statistical significance. Suppose that we are interested in determining whether individuals who take GRE preparation courses do better on the GRE than similar individuals who do not. We obtain some data on a random sample of students who take the GRE and compare the mean GRE scores (perhaps conditional on individual characteristics) of those who took preparation courses and those who did not. Suppose that we estimate a mean difference of five points. With a large enough sample, this difference will be statistically significant; that is, with a large enough sample we can reject the null that it equals zero.

However, a mean difference of five points is not large enough to make a substantive difference. That is, while the five point difference may be statistically significant (in a large enough sample) the magnitude of the difference is not very large, so we would not say that it is substantively significant.