Wooldridge, Introductory Econometrics, 4th ed.

Appendix C: Fundamentals of mathematical statistics

A short review of the principles of mathematical statistics. Econometrics is concerned with statistical inference: learning about the characteristics of a population from a sample of the population. The **population** is a well-defined group of subjects–and it is important to define the population of interest. Are we trying to study the unemployment rate of all labor force participants, or only teenaged workers, or only AHANA workers? Given a population, we may define an economic model that contains parameters of interest–coefficients, or elasticities, which express the effects of changes in one variable upon another.

Let Y be a random variable (r.v.) representing a population with probability density function (pdf) $f(y;\theta)$, with θ a scalar parameter. We assume that we know f,but do not know the value of θ. Let a random sample from the population be $(Y_1, ..., Y_N)$, with Y_i being an independent random variable drawn from $f(y;\theta)$. We speak of Y_i being iid – independently and identically distributed. We often assume that random samples are drawn from the Bernoulli distribution (for instance, that if I pick a student randomly from my class list, what is the probability that she is female? That probability is γ, where γ% of the students are female, so $P(Y_i = 1) = \gamma$ and $P(Y_i = 0) = (1 - \gamma)$. For many other applications, we will assume that samples are drawn from the Normal distribution. In that case, the pdf is characterized by two parameters, μ and σ^2, expressing the mean and spread of the distribution, respectively.

Finite sample properties of estimators

The finite sample properties (as opposed to asymptotic properties) apply to all sample sizes, large or small. These are of great relevance when we are dealing with samples of limited size, and unable to conduct a survey to generate a larger sample. How well will estimators perform in this context? First we must distinguish between estimators and estimates. An **estimator** is a rule, or algorithm, that specifies how the sample information should be manipulated in order to generate a numerical **estimate**. Estimators have properties–they may be reliable in some sense to be defined; they may be easy or difficult to calculate; that difficulty may itself be a function of sample size. For instance, a test which involves measuring the distances between every observation of a variable involves an order of calculations which grows more than linearly with sample size. An estimator with which we are all familiar is the

sample average, or arithmetic mean, of N numbers: add them up and divide by N. That estimator has certain properties, and its application to a sample produces an estimate. We will often call this a **point estimate**–since it yields a single number–as opposed to an **interval estimate**, which produces a range of values associated with a particular level of confidence. For instance, an election poll may state that 55% are expected to vote for candidate A, with a margin of error of $\pm 4\%$. If we trust those results, it is likely that candidate A will win, with between 51% and 59% of the vote. We are concerned with the **sampling distributions** of estimators–that is, how the estimates they generate will vary when the estimator is applied to repeated samples.

What are the finite-sample properties which we might be able to establish for a given estimator and its sampling distribution? First of all, we

are concerned with unbiasedness. An estimator W of θ is said to be **unbiased** if $E(W) = \theta$ for all possible values of θ. If an estimator is unbiased, then its probability distribution has an expected value equal to the population parameter it is estimating. Unbiasedness does not mean that a given estimate is equal to θ, or even very close to θ; it means that if we drew an infinite number of samples from the population and averaged the W estimates, we would obtain θ. An estimator that is **biased** exhibits $Bias(W) = E(W) - \theta$. The magnitude of the bias will depend on the distribution of the Y and the function that transforms Y into W, that is, the estimator. In some cases we can demonstrate unbiasedness (or show that bias=0) irregardless of the distribution of Y; for instance, consider the sample average $\bar{Y}$, which is an unbiased estimate of the population mean μ :

$$E(\bar{Y}) \;=\; E(\frac{1}{n}\sum_{i=1}^{n} Y_i)$$

$$
\begin{aligned}
&= \frac{1}{n} E(\sum_{i=1}^{n} Y_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} E(Y_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mu \\
&= \frac{1}{n} n\mu = \mu
\end{aligned}
$$

Any hypothesis tests on the mean will require an estimate of the variance, σ^2, from a population with mean μ. Since we do not know μ (but must estimate it with $\bar{Y}$), the estimate of sample variance is defined as

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

with one degree of freedom lost by the replacement of the population statistic μ with its sample estimate $\bar{Y}$. This is an unbiased estimate of the population variance, whereas the counterpart with a divisor of n will be biased unless we

know μ. Of course, the degree of this bias will depend on the difference between $\left(\frac{n}{n-1}\right)$ and unity, which disappears as $n \to \infty$.

Two difficulties with unbiasedness as a criterion for an estimator: some quite reasonable estimators are unavoidably biased, but useful; and more seriously, many unbiased estimators are quite poor. For instance, picking the first value in a sample as an estimate of the population mean, and discarding the remaining $(n-1)$ values, yields an unbiased estimator of μ, since $E(Y_1) = \mu$; but this is a very imprecise estimator.

What additional information do we need to evaluate estimators? We are concerned with the precision of the estimator as well as its bias. An unbiased estimator with a smaller sampling variance will dominate its counterpart with a larger sampling variance: e.g. we

can demonstrate that the estimator that uses only the first observation to estimate μ has a much larger sampling variance than the sample average, for nontrivial n. What is the sampling variance of the sample average?

$$\begin{aligned} Var(\bar{Y}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n} Var(Y_i)\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n} \sigma^2\right) \\ &= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

so that the precision of the sample average depends on the sample size, as well as the (unknown) variance of the underlying distribution of Y. Using the same logic, we can derive the

sampling variance of the "estimator" that uses only the first observation of a sample as σ^2. Even for a sample of size 2, the sample mean will be twice as precise.

This leads us to the concept of **efficiency**: given two unbiased estimators of θ, an estimator W_1 is efficient relative to W_2 when $Var(W_1) \leq Var(W_2)\ \forall\theta$, with strict inequality for at least one θ. A relatively efficient unbiased estimator dominates its less efficient counterpart. We can compare two estimators, even if one or both is biased, by comparing mean squared error (MSE), $MSE(W) = E\left[(W-\theta)^2\right]$. This expression can be shown to equal the variance of the estimator plus the square of the bias; thus, it equals the variance for an unbiased estimator.

Large sample (asymptotic) properties of estimators

We can compare estimators, and evaluate their relative usefulness, by appealing to their large sample properties–or **asymptotic** properties. That is, how do they behave as sample size goes to infinity? We see that the sample average has a sampling variance with limiting value of zero as $n \to \infty$. The first asymptotic property is that of consistency. If W is an estimate of θ based on a sample $[Y_1, ..., Y_n]$ of size n, W is said to be a **consistent** estimator of θ if, for every $\epsilon > 0$,

$$P\left(|W_n - \theta| > \epsilon\right) \to 0 \ as \ n \to \infty.$$

Intuitively, a consistent estimator becomes more accurate as the sample size increases without bound. If an estimator does not possess this property, it is said to be **inconsistent**. In that case, it does not matter how much data we have; the "recipe" that tells us how to use the data to estimate θ is flawed. If an estimator is

biased but its variance shrinks as $n \to \infty$, then the estimator is consistent.

A consistent estimator has probability limit, or plim, equal to the population parameter: $\text{plim}(\bar{Y}) = \mu$. Some mechanics of plims: let θ be a parameter and $g(\cdot)$ a continuous function, so that $\gamma = g(\theta)$. Suppose $\text{plim}(W_n) = \theta$, and we devise an estimator of γ, $G_n = g(W_n)$. Then $\text{plim}(G_n) = \gamma$, or

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

This allows us to establish the consistency of estimators which can be shown to be transformations of other consistent estimators. For instance, we can demonstrate that the estimator given above of the population variance is not only unbiased but consistent. The standard deviation is the square root of the variance: a nonlinear function, continuous for positive arguments. Thus the standard deviation S is

a consistent estimator of the population standard deviation. Some additional properties of plims, if $\text{plim}(T_n) = \alpha$ and $\text{plim}(U_n) = \beta$:

$$\begin{aligned} \text{plim}\,(T_n + U_n) &= \alpha + \beta \\ \text{plim}\,(T_n U_n) &= \alpha\beta \\ \text{plim}\,(T_n / U_n) &= \alpha/\beta, \beta \neq 0. \end{aligned}$$

Consistency is a property of point estimators: the distribution of the estimator collapses around the population parameter in the limit, but that says nothing about the shape of the distribution for a given sample size. To work with interval estimators and hypothesis tests, we need a way to approximate the distribution of the estimators. Most estimators used in econometrics have distributions that are reasonably approximated by the Normal distribution for large samples, leading to the concept of *asymptotic normality*:

$$P\,(Z_n \leq z) \to \Phi\,(z) \;\; as \;\; n \to \infty$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (cdf). We will often say "$Z_n \tilde{} N(0,1)$" or "Z_n is asy N." This relates to one form of the central limit theorem (CLT). If $[Y_1, ...Y_n]$ is a random sample with mean μ and variance σ^2,

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

has an asymptotic standard normal distribution. Regardless of the population distribution of Y, this standardized version of Y will be asy N, and the entire distribution of Z will become arbitrarily close to the standard normal as $n \to \infty$. Since many of the estimators we will derive in econometrics can be viewed as sample averages, the law of large numbers and the central limit theorem can be combined to show that these estimators will be asy N. Indeed, the above estimator will be asy N even if we replace σ with a consistent estimator of that parameter, S.

General approaches to parameter estimation

What general strategies will provide us with estimators with desirable properties such as unbiasedness, consistency and efficiency? One of the most fundamental strategies for estimation is the **method of moments**, in which we replace population moments with their sample counterparts. We have seen this above, where a consistent estimator of sample variance is defined by replacing the unknown population μ with a consistent estimate thereof, $\bar{Y}$. A second widely employed strategy is the principle of **maximum likelihood**, where we choose an estimator of the population parameter θ by finding the value that maximizes the likelihood of observing the sample data. We will not focus on maximum likelihood estimators in this course, but note their importance in econometrics. Most of our work here is based on the

least squares principle: that to find an estimate of the population parameter, we should solve a minimization problem. We can readily show that the sample average is a method of moments estimator (and is in fact a maximum likelihood estimator as well). We demonstrate now that the sample average is a least squares estimator:

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

will yield an estimator, m, which is identical to that defined as $\bar{Y}$. We may show that the value m minimizes the sum of squared deviations about the sample mean, and that any other value m' would have a larger sum (or would not be "least squares"). Standard regression techniques, to which we will devote much of the course, are often called "OLS": ordinary least squares.

Interval estimation and confidence intervals

Since an estimator will yield a value (or point estimate) as well as a sampling variance, we may generally form a confidence interval around the point estimate in order to make probability statements about a population parameter. For instance, the fraction of Firestone tires involved in fatal accidents is surely not 0.0005 of those sold. Any number of samples would yield estimates of that mean differing from that number (and for a continuous random variable, the probability of a point is zero). But we can test the hypothesis that 0.0005 of the tires are involved with fatal accidents if we can generate both a point and interval estimate for that parameter, and if the interval estimate cannot reject 0.0005 as a plausible value. This is the concept of a **confidence interval**, which is defined with regard to a given

level of "confidence" or level of probability. For a standard normal ($N(0,1)$) variable,

$$P\left(-1.96 < \frac{\bar{Y}-\mu}{1/\sqrt{n}} < 1.96\right) = 0.95.$$

which defines the interval estimate $\left(\bar{Y} - \frac{1.96}{\sqrt{n}}, \bar{Y} + \frac{1.96}{\sqrt{n}}\right)$. We do not conclude from this that the probability that μ lies in the interval is 0.95; the population parameter either lies in the interval or it does not. The proper way to consider the confidence interval is that if we construct a large number of random samples from the population, 95% of them will contain μ. Thus, if a hypothesized value for μ lies outside the confidence interval for a single sample, that would occur by chance only 5% of the time.

But what if we do not have a standard normal variate, for which we know the variance equals unity? If we have a variable X, which we conclude is distributed as $N\left(\mu, \sigma^2\right)$, we arrive at

the difficulty that we do not know σ^2 : and thus cannot specify the confidence interval. Via the method of moments, we replace the unknown σ^2 with a consistent estimate, S^2, to form the transformed statistic

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_n$$

denoting that its distribution is no longer standard normal, but "student's t" with n degrees of freedom. The t distribution has fatter tails than does the normal; above 20 or 25 degrees of freedom, it is approximated quite well by the normal. Thus, confidence intervals constructed with the t distribution will be wider for small n, since the value will be larger than 1.96. A 95% confidence interval, given the symmetry of the t distribution, will leave 2.5% of probability in each tail (a two-tailed t test).

If c_α is the 100(1-α) percentile in the t distribution, a 100(1-α)% confidence interval for the mean will be defined as:

$$\bar{y} - c_{\alpha/2}\frac{s}{\sqrt{n}},\ \bar{y} + c_{\alpha/2}\frac{s}{\sqrt{n}}$$

where s is the estimated standard deviation of Y. We often refer to $\frac{s}{\sqrt{n}}$ as the standard error of the parameter–in this case, the standard error of our estimate of μ. Note well the difference between the concepts of the standard deviation of the underlying distribution (an estimate of σ) and the standard error, or precision, of our estimate of the mean μ. We will return to this distinction when we consider regression parameters. A simple rule of thumb, for large samples, is that a 95% confidence interval is roughly two standard errors on either side of the point estimate–the counterpart of a "t of 2" denoting significance of a parameter. If an estimated parameter is more than

two standard errors from zero, a test of the hypothesis that it equals zero in the population will likely be rejected.

Hypothesis testing

We want to test a specific hypothesis about the value of a population parameter θ. We may believe that the parameter equals 0.42; so that we state the null and alternative hypotheses:

$$\begin{aligned} H_0 &: \theta = 0.42 \\ H_A &: \theta \neq 0.42 \end{aligned}$$

In this case, we have a two-sided alternative: we will reject the null if our point estimate is "significantly" below 0.42, or if it is "significantly" above 0.42. In other cases, we may specify the alternative as one-sided. For instance, in a quality control study, our null might be that the proportion of rejects from

the assembly line is no more than 0.03, versus the alternative that it is greater than 0.03. A rejection of the null would lead to a shutdown of the production process, whereas a smaller proportion of rejects would not be cause for concern. Using the principles of the scientific method, we set up the hypothesis and consider whether there is sufficient evidence against the null to reject it. Like the principle that a finding of guilt must be associated with evidence beyond a reasonable doubt, the null will stand unless sufficient evidence is found to reject it as unlikely. Just as in the courts, there are two potential errors of judgment: we may find an innocent person guilty, and reject a null even when it is true; this is **Type I error**. We may also fail to convict a guilty person, or reject a false null; this is **Type II error**. Just as the judicial system tries to balance those two types of error (especially considering the consequences of punishing the innocent, or even

putting them to death), we must be concerned with the magnitude of these two sources of error in statistical inference. We construct hypothesis tests so as to make the probability of a Type I error fairly small; this is the **level of the test**, and is usually denoted as α. For instance, if we operate at a 95% level of confidence, then the level of the test is $\alpha = 0.05$. When we set α, we are expressing our tolerance for committing a Type I error (and rejecting a true null). Given α, we would like to minimize the probability of a Type II error, or equivalently maximize the **power of the test**, which is just one minus the probability of committing a Type II error, and failing to reject a false null. We must balance the level of the test (and the risk of falsely rejecting the truth) with the power of the test (and failing to reject a false null).

When we use a computer program to calculate point and interval estimates, we are given the

information that will allow us to reject or fail to reject a particular null. This is usually phrased in terms of $p - values$, which are the tail probabilities associated with a test statistic. If the p-value is less than the level of the test, then it leads to a rejection: a p-value of 0.035 allows us to reject the null at the level of 0.05. One must be careful to avoid the misinterpretation of a p-value of, say, 0.94, which is indicative of the massive failure to reject that null.

One should also note the duality between confidence intervals and hypothesis tests. They utilize the same information: the point estimate, the precision as expressed in the standard error, and a value taken from the underlying distribution of the test statistic (such as 1.96). If the boundary of the 95% confidence interval contains a value δ, then a hypothesis test that the population parameter equals δ will be on the borderline of acceptance and rejection at the 5% level. We can consider these

quantities as either defining an interval estimate for the parameter, or alternatively supporting an hypothesis test for the parameter.