This material comes from Hansen's *Probability and Statistics for Economists* (PSE) and Len Goff's lecture notes along with some of my own comments.

Law of Large Numbers

So far, we have primarily focused on the mean and variance of estimators, particularly $\bar{X}$. One thing that is worth pointing out here is that, in order to derive these properties, we did not need to impose any strong conditions (either on the distribution of X_i or on the number of observations that we have access to). But we will often want to learn more about the sampling distribution of $\hat{\theta}$ or $\bar{X}$ than just its mean and variance.

One way to proceed would be to make additional distributional assumptions about X_i; the most common one would be to assume that X_i are normally distributed. In this case (which I'll talk about briefly in the next set of lecture notes), we could derive the full sampling distribution of $\bar{X}$. However, these sorts of distributional assumptions are often implausible in economics applications. Previously we have considered examples where X is a person's years of education or income. Neither of these variables follow a normal distribution. Years of education is discrete which implies that it does not follow a normal distribution. Income cannot be negative which implies that it doesn't follow a normal distribution (as the support of a normally distributed random variable is the entire real line); moreover, the income distribution is skewed and has a "fat" right tail (i.e., the fraction of people with "very high" incomes is too large relative to a normal distribution) which also implies non-normality.

Therefore, in this set of notes, we will start to consider "large sample" properties of estimators. These will amount to properties of estimators that hold in cases where we have a large sample and will be derived using asymptotic arguments where $n \to \infty$. The two main large sample properties of estimators are consistency and asymptotic normality. Consistency amounts to a guarantee that, given a large enough sample, our estimate should be "close" to the population parameter of interest. Asymptotic normality involves a transformed version of our estimator behaving like a draw from a normal distribution, given that we have a large sample. Asymptotic normality will be useful for (roughly) developing measures of the accuracy of our estimator and for hypothesis testing (this is essentially that we have some theory about a population quantity of interest and want to learn whether or not the data that we have is "compatible" with that theory or not). This set of notes considers consistency and the main tools for showing that an estimator is consistent. The next set of notes concern asymptotic normality.

Convergence in probability

PSE 7.1-7.4

Consider a sequence of random variables or random vectors $Z_1, Z_2, \ldots$ which we'll denote by Z_n. This is a general notion, but I think a good example to keep in mind is $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (usually I don't include the subscript n on $\bar{X}$, but I think it is helpful to include it at the moment). We are

going to think about "convergence" of sequences of random variables. Notice that the distribution of Z_n can change depending on n. For example, for $\bar{X}_n$, the variance depends on n. The first notion of convergence of the sequence Z_n that we will consider is called **convergence in probability**:

Definition. We say that Z_n converges in probability to c if for any $\epsilon > 0$:

$$\lim_{n\to\infty} \mathrm{P}(|Z_n - c| < \epsilon) = 1$$

In words, convergence in probability means that, as in $n \to \infty$, Z_n becomes arbitrarily close to c. An alternative equivalent definition is that $\lim_{n\to\infty} \mathrm{P}(|Z_n - c| \geq \epsilon) = 0$. The second version says that the probability that Z_n is "not close" to c (you can define whatever you "not close" to mean based on choosing different values of ϵ) goes to 0 as $n \to \infty$

Notation: When Z_n converges in probability to c, we write this as $Z_n \xrightarrow{p} c$, or alternatively $plim(Z_n) = c$. We say that c is the *probability limit* of the sequence Z_n.

Weak law of large numbers

PSE 7.5

The first large sample property of an estimator that we will consider is **consistency**

Definition. An estimator $\hat{\theta}$ of a parameter θ is consistent if $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$.

Consistency is a good property for an estimator to have. In fact, all the estimators that we'll consider this semester or next semester will be consistent. Consistency says that, given a large enough sample, $\hat{\theta}$ will be arbitrarily close to θ.

The main tool for showing that an estimator is consistent is the **weak law of large numbers**.

Weak Law of Large Numbers: If X_i are iid and $\mathbb{E}|X| < \infty$, then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}[X]$$

The weak law of large numbers says that, given a large enough sample (under iid sampling), sample averages should be close to population averages. In my view, this is quite intuitive and not all that surprising: if you flip lots of coins, it does not seem surprising that the fraction of heads should be very close to 0.5. The notation in the theorem above indicates that X is scalar, but the weak law of large numbers immediately applies in the case where X is a vector.

Proof: Recall Chebyshev's inequality says that, for some random variable Z,

$$\mathrm{P}(|Z - \mathbb{E}[Z]| \geq \delta) \leq \frac{\mathrm{var}(Z)}{\delta^2}$$

Then, applying Chebyshev's inequality to $\bar{X}$ (and recalling that $\mathbb{E}[\bar{X}] = \mathbb{E}[X]$ and $\text{var}(\bar{X}) = \text{var}(X)/n$), we have that

$$\begin{aligned} \text{P}(|\bar{X} - \mathbb{E}[X]| \geq \delta) &\leq \frac{\text{var}(\bar{X})}{\delta^2} \\ &= \frac{1}{n}\frac{\text{var}(X)}{\delta^2} \\ &\to 0 \text{ as } n \to \infty \end{aligned}$$

Thus, $\text{P}(|\bar{X} - \mathbb{E}[X]| \geq \delta) \to 0$ as $n \to \infty$ which implies that $\bar{X} \xrightarrow{p} \mathbb{E}[X]$.

One thing that is worth briefly mentioning is that, in the proof, we (implicity) used the slightly stronger condition that $\mathbb{E}[X^2] < \infty$ (so that $\text{var}(X)$ exists). The weak law of large numbers actually holds when only the first moment exists (as stated in the theorem) though the proof is more technical in this case. If you are interested in this case, then you can check the textbook.

So far, we have talked about three properties of estimators: bias, sampling variance, and consistency. I don't think you will get sampling variance confused with consistency, but it is worth being clear about what the differences are between unbiasedness and consistency. First, bias is a finite sample property while consistency is a large sample property. For example, our results on the unbiasedness of $\bar{X}$ for $\mathbb{E}[X]$ did not depend on the number of observations (n could be equal to 1 or 5 or 1000). Consistency, on the other hand, is a property of an estimator given that one has a large sample. Along these lines, if you know that an estimator is unbiased, given a particular sample and therefore a particular estimate, the estimate could still, in principle, be far from the target population parameter. On the other hand, a consistent estimator will be close to the target population parameter given a large enough sample. Second, unbiasedness involves the expected value of an estimator where the expectation is in the repeated sampling thought experiment that we have previously considered where one repeatedly draws random samples of size n and re-estimates $\hat{\theta}$ with each new sample. On the other hand, consistency is a property of having one large sample.

Example: Along the lines of the discussion above, I think it is helpful to have some examples of simple estimators and whether they are unbiased and/or consistent. Below, we'll consider four estimators of $\mathbb{E}[X]$.

(1) $\bar{X}$. This is unbiased (we have shown before that $\mathbb{E}[\bar{X}] = \mathbb{E}[X]$) and consistent (this follows immediately from the weak law of large numbers).

(2) $\hat{\mu}_1 = X_1$ (i.e., this is the estimator that we have talked about before that just uses the first observation). This is unbiased because $\mathbb{E}[X_1] = \mathbb{E}[X]$, but it is not consistent because, even if you have a large sample, this estimator just throws away all observations except the first one.

(3) $\hat{\mu}_\lambda = \lambda\bar{X}$ for some constant $\lambda > 0$. This is biased, unless $\lambda = 1$, because $\mathbb{E}[\lambda\bar{X}] = \lambda\mathbb{E}[X]$. It is also not consistent because $\lambda\bar{X} \xrightarrow{p} \lambda\mathbb{E}[X] \neq \mathbb{E}[X]$ unless $\lambda = 1$.

(4) $\hat{\mu}_c = \bar{X} + \frac{c}{n}$ for a constant c. This is biased because $\mathbb{E}[\bar{X} + \frac{c}{n}] = \mathbb{E}[X] + \frac{c}{n} \neq \mathbb{E}[X]$, but it is consistent because $\bar{X} + \frac{c}{n} \xrightarrow{p} \mathbb{E}[X] + 0 = \mathbb{E}[X]$ which holds because $c/n \to 0$ as $n \to \infty$ because c is constant.

Although some of these examples are obviously strange/poor ways to estimate $\mathbb{E}[X]$ (particularly (2) and (4)), they do show that its possible for an estimator to be both, either, or neither unbiased and consistent.

A couple of last things to point out. First, note that we stated the weak law of numbers for X_i, but it immediately applies to functions of random variables. That is, supposing that X_i are iid and $\mathbb{E}|g(X)| < \infty$, then $\frac{1}{n}\sum_{i=1}^n g(X_i) \xrightarrow{p} \mathbb{E}[g(X)]$. This immediately implies that the class of analogue estimators are consistent. Finally, for the weak law of large numbers, the weak law of large numbers held under the conditions that X_i are iid and that $\mathbb{E}|X| < \infty$. There exist other versions of the law of large numbers that can hold under certain violations of iid sampling. I am not planning on covering these in any detail, but it is worth noting that they exist and this can be relevant under alternative sampling schemes. The condition about the first moment of X existing is a "regularity condition" — that is, it is not really something that you can test in a particular application, but it should not be considered a strong assumption and one can reasonably expect that the first moment of the vast majority of variables in economics will exist.

Continuous Mapping Theorem

PSE 7.10

The weak law of large numbers provides a direct way to show that sample averages are consistent for population means. We'll often want to consider more complicated estimators, particularly, the class of plug-in estimators; i.e., for the case where we are interested in $\beta = h(\theta)$, θ is a population mean such as $\mathbb{E}[X]$, and where we estimate β by $\hat{\beta} = h(\hat{\theta})$. Next, we'll consider a main tool for

establishing consistency of more complicated functionals than just sample averages.

Definition. A function $h(x)$ is continuous at $x = c$ if for all $\epsilon > 0$ there exists a $\delta > 0$ such that $||x - c|| \leq \delta$ implies that $||h(x) - h(c)|| \leq \epsilon$

This definition probably corresponds to your intuition of what it means for a function to be continuous. For values of x that are very close to c, then $h(x)$ should be close to $h(c)$. It may also be helpful think about a discontinuous function and how it would violate this sort of condition.

Continuous Mapping Theorem: If $Z_n \xrightarrow{p} c$ and $h(\cdot)$ is continuous at c, then $h(Z_n) \xrightarrow{p} h(c)$.

Proof: For some $\epsilon > 0$, since $h(x)$ is continuous at c it means that there exists a $\delta > 0$ such that $||x - c|| \leq \delta \implies ||h(x) - h(c)|| \leq \epsilon$. Taking $x = Z_n$, this implies that

$$\begin{aligned} \mathrm{P}\Big(||h(Z_n) - h(c)|| \leq \epsilon\Big) &\geq \mathrm{P}\Big(||Z_n - c|| \leq \delta\Big) \\ &\to 1 \text{ as } n \to \infty \end{aligned}$$

where this last line holds because $Z_n \xrightarrow{p} c$. Thus, $\mathrm{P}\Big(||h(Z_n) - h(c)|| \leq \epsilon\Big) \to 1$ as $n \to \infty$ which implies that $h(Z_n) \xrightarrow{p} h(c)$. Also note that the direction of the inequality in the first line holds because, if $||Z_n - c|| \leq \delta$, then it must be the case that $||h(Z_n) - h(c)|| \leq \epsilon$. A longer explanation is provided in the box below.

Side-Comment: Here is a longer explanation of the direction of the inequality in the proof of the continuous mapping theorem. First, note that we can write

$$\mathrm{P}\Big(||Z_n - c|| \leq \delta\Big) = \mathrm{P}\Big(||Z_n - c|| \leq \delta, ||h(Z_n) - h(c)|| \leq \epsilon\Big) + \underbrace{\mathrm{P}\Big(||Z_n - c|| \leq \delta, ||h(Z_n) - h(c)|| > \epsilon\Big)}_{=0 \text{ because } h \text{ continuous}} \quad (1)$$

Similarly, notice that

$$\begin{aligned} \mathrm{P}\Big(||h(Z_n) - h(c)|| \leq \epsilon\Big) &= \mathrm{P}\Big(||h(Z_n) - h(c)|| \leq \epsilon, ||Z_n - c|| \leq \delta\Big) + \mathrm{P}\Big(||h(Z_n) - h(c)|| \leq \epsilon, ||Z_n - c|| > \delta\Big) \\ &= \mathrm{P}\Big(||Z_n - c|| \leq \delta\Big) + \mathrm{P}\Big(||h(Z_n) - h(c)|| \leq \epsilon, ||Z_n - c|| > \delta\Big) \\ &\geq \mathrm{P}\Big(||Z_n - c|| \leq \delta\Big) \end{aligned}$$

where the second line holds by plugging in from Equation 1 above, and the inequality holds because both the probabilities in the previous line are positive.

Here are some important examples of the continuous mapping theorem. If $Z_n \xrightarrow{p} c$ and for some constant a, then

- $Z_n + a \xrightarrow{p} c + a$

- $aZ_n \xrightarrow{p} ac$
- $Z_n^2 \xrightarrow{p} c^2$

To conclude, let's use the continuous mapping theorem to establish the consistency of the class of plugin estimators.

Theorem: If X_i are iid, $\mathbb{E}||g(X)|| < \infty$ and $h(u)$ is continuous at $u = \theta = \mathbb{E}[g(X)]$, then for $\beta = h(\theta)$, $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$, and $\hat{\beta} = h(\hat{\theta})$, $\hat{\beta} \xrightarrow{p} \beta$

This result holds immediately by the continuous mapping theorem and weak law of large numbers. In particular, the weak law of large numbers implies that $\hat{\theta} \xrightarrow{p} \theta$, and then the continuous mapping theorem implies (due to h being continuous at θ) that $h(\hat{\theta}) \xrightarrow{p} h(\theta)$ which is equivalent to $\hat{\beta} \xrightarrow{p} \beta$.