Law of Large Numbers

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.

So far, we have primarily focused on the mean and variance of estimators, particularly $\bar{X}$. 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, often, we will need to know the entire sampling distribution. Suppose that we somehow knew that X_i were normally distributed. In this case, it immediately follows that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. This implies that we would know the entire sampling distribution of $\bar{X}$ in this case.

Unfortunately, it is not reasonable to assume that many variables in economics follow a normal distribution. In this case, it is generally much harder (or impossible) to derive the sampling distribution of $\bar{X}$.

That said, the most common way to derive a sampling distribution for estimators is for the case where the researcher has a "large' ' sample. These arguments are called **asymptotic approximations** and amount to deriving properties of the sampling distribution as $n \to \infty$. In this set of notes, we'll cover the law of large numbers and continuous mapping theorem. In the next set of notes, we'll cover the central limit theorem.

Convergence in probability

PSE 7.1-7.4

Given a sequence of random variables or random vectors $Z_1, Z_2, \ldots$, let us now define a notion of convergence of the sequence Z_n called **convergence in probability**:

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

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

An alternative equivalent definition is that $\lim_{n\to\infty} \mathrm{P}(|Z_n - c| > \delta) = 0$.

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

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 minimal 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. That said, consistency does not give us a precise definition of what a `large enough'' sample is --- in fact, how quickly the law of large number`kicks in'' depends on the distribution of X.

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

Theorem: [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.

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 $\bar{X}$, we have that

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

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

Introduction: the law of large numbers

Consider an *i.i.d.* sample $\{X_1, \ldots X_n\}$ of some random variable X_i. The *sample average* of X_i in our data simply takes the arithmetic mean across these n observations:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

The law of large numbers (LLN) states the deep and useful fact that for very large n, it becomes very unlikely that $\bar{X}_n$ is very far from $\mu = \mathbf{E}[X_i]$, the "population mean'' of X_i.

Theorem: [(law of large numbers)] If X_i are $i.i.d$ random variables and $E[X_i]$ is finite, then for any $\epsilon > 0$:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

Note: The LLN is stated above for a random variable, but the result generalizes easily to random vectors. In that case, $\lim_{n\to\infty} P(||\bar{\mathbf{X}}_n - \mu||_2 > \epsilon) = 0$ where $||\cdot||_2$ denotes the Euclidean norm, i.e.: $|\bar{\mathbf{X}}_n - \mu| = (|\bar{\mathbf{X}}_n - \mu|)'(|\bar{\mathbf{X}}_n - \mu|)$, where $\bar{\mathbf{X}}_n$ is a vector of sample means for each component of X_i, and similarly for μ.\
Note: the version of the law of large numbers above is called the *weak* law of large numbers. There exists another version called the strong LLN.\%The *strong* law of large numbers says that not only does $\bar{x}_n$ converge in probability to μ, but that along a sequence of random samples indexed by n, the limit of $\bar{x}_n$ as n approaches infinity is equal to μ, with probability one.
Let us now prove the LLN. We will do so using a tool called *Chebyshev's inequality*. This proof assumes that $Var(X_i)$ is finite, but the LLN holds even if $Var(X_i) = \infty$. Chebyshev's inequality allows us to use the variance of a random variable to put an upper bound on the probability that the random variable is far from its mean. In particular, for any random variable Z with finite mean and variance:

$$P(|Z - \mathbb{E}[Z]| \geq \epsilon) \leq \frac{Var(Z)}{\epsilon^2}$$

To see that this holds, use the law of iterated expectations to write out the variance as

$$\begin{aligned} Var(Z) = \mathbb{E}\left[Z - \mathbb{E}[Z])^2\right] &= P(|Z - \mathbb{E}[Z]| \geq \epsilon) \cdot \mathbb{E}\left[(Z - \mathbb{E}[Z])^2 | (Z - \mathbb{E}[Z])^2 \geq \epsilon^2\right] \\ &\quad + P(|Z - \mathbb{E}[Z]| < \epsilon) \cdot \mathbb{E}\left[(Z - \mathbb{E}[Z])^2 | (Z - \mathbb{E}[Z])^2 < \epsilon^2\right] \\ &\geq P(|Z - \mathbb{E}[Z]| \geq \epsilon) \cdot \epsilon^2 + P(|Z - \mathbb{E}[Z]| < \epsilon) \cdot 0, \end{aligned}$$

noting that $|Z - \mathbb{E}[Z]| \geq \epsilon$ iff $(Z - \mathbb{E}[Z])^2 \geq \epsilon^2$.\
Now, we will show that as $n \to \infty$, $Var(\bar{X}_n) \to 0$. This along with Chebyshev's inequality implies the LLN, by letting $Z = \bar{X}_n$.
To see that $Var(\bar{X}_n) \overset{n}{\to} 0$, note first that

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

The first equality is simply the definition of $\bar{X}_n$, while the second uses linearity of the expectation

operator. Now consider

$$\begin{aligned} Var(\bar{X}_n) &= \mathbb{E}\left[(\bar{X}_n - \mathbb{E}[\bar{X}_n])^2\right] = \mathbb{E}\left[(\bar{X}_n - \mu)^2\right] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right)^2\right] \\ &= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - \mu)(X_j - \mu)\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[(X_i - \mu)(X_j - \mu)\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(X_i - \mu)[(X_i - \mu)\right] = \frac{1}{n^2}\cdot nVar(X_i) = \frac{Var(X_i)}{n} \end{aligned}$$

where the first equality in the third line follows because when $i \neq j$, $X_i \perp X_j$ implies that $\mathbb{E}[(X_i - \mu)] \cdot \mathbb{E}[(X_j - \mu)] = 0 \cdot 0$. Thus, the only terms that remain are when $j = i$.

Another way to see that $Var(\bar{X}_n) = \frac{Var(X_i)}{n}$ is to notice that when Y and Z are independent, $Var(Y + Z) = Var(Y) + Var(Z)$. Thus:

$$Var\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n+\right) = n \cdot Var\left(\frac{1}{n}X_i\right) = n \cdot \frac{1}{n^2} \cdot Var(X_i) = \frac{Var(X_i)}{n}$$

Asymptotic sequences

The law of large numbers provides a way to justify the claim that when n is large, $\bar{X}_n$ will be close to μ with high probability. The approximation $\bar{X}_n \approx \mu$ lies at the heart of our claims to be learning about an underlying population when we have a large sample.

In the next section, we'll see that there is more than one way to develop a large-n approximation to the distribution of a random variable. To talk about such approximations, it is useful to introduce the idea of a *sequence* of random variables Z_n, where $n = 1, 2, \ldots \infty$. For example, we can consider the sample mean $\bar{X}_n$—which is a random variable for any given n—across various possible sample sizes n.

The general problem

The primary motivation for considering such *asymptotic sequences* of random variables Z_n is when Z_n represents a statistic $\hat{\theta}$—something that depends upon my data (see Definition **??**). Since $\hat{\theta}$ is random (it depends on the sample that I drew), I'd like to know something about its distribution. For example, how likely is it that my sample mean is far from the population mean?

Definition. The **sampling distribution** of an statistic $\hat{\theta}$ is its CDF: $F_{\hat{\theta}}(t) = P(\hat{\theta} \leq t)$.

When our statistic is computed as $\hat{\theta} = g(X_1, X_2, \ldots X_n)$ from an *i.i.d* sample of X_i, $F_{\hat{\theta}}$ depends upon three things: the function g, the population distribution of X_i, and the sample size n.% Our notation F_n for the sampling distribution highlights the dependence on n, which is our primary focus here.

Knowing the sampling distribution of a statistic is typically a hard problem. We know g and n, but in a research setting we don't generally know the CDF F that describes the underlying population. However, if we view $\hat{\theta}$ as a point along a sequence of random variables Z_n, it is often possible to say something about the limiting behavior of F_{Z_n} as $n \to \infty$. *Asymptotic theory* is a set of tools for describing this limiting behavior. The law of large numbers is one such tool. If we believe that are actual sample size n is large enough that $F_{Z_n} \approx F_{Z_\infty}$, then tools like the LLN can be extremely useful. For the sample mean for example, we might, on the basis of the LLN, be prepared to believe that $\bar{X}_n$ is close to μ with very high probability.

Conceptually, we can think of what we're doing as follows. Suppose our sample size is $n = 10,576$, and we calculate a statistic $\hat{\theta} = g(X_1, X_2, \ldots X_{10,576})$ from our sample. Now imagine applying the same function g to various samples of size $1, 2, \ldots$ and so on, and defining a sequence Z_1, Z_2, Z_n of the corresponding values. Each Z along this sequence is itself a random variable: let $F_{Z_1}, F_{Z_2}, \ldots$ be their corresponding CDFs. Our statistic $\hat{\theta}$ can be seen as a specific point along this sequence: $\hat{\theta} = Z_{10,576}$ (circled in red in Figure 1). Since we don't know $F_{Z_{10,576}}$, but we can say something about F_{Z_∞}, we use the latter as an approximation for the former. Figure 1 depicts this logic.
Of course, the above technique only works if we can say something definite about F_∞. The law of large numbers says that we can when our statistic is the sample mean. In Section **??**, we'll see that the *central limit theorem* provides even more information about the limiting distribution of the sample mean: that it will become approximately normal, regardless of F. %We close this section by thinking through the case of the LLN in more detail.

> *Note:* The logic of Figure 1 is the "classical" approach to approximating the sampling distribution of $\hat{\theta}$, but it is certainly not the only one. An increasingly popular alternative involves *bootstrap* methods. These methods still appeal to n being "large enough", but they do so in a different way. They also require computing power, because bootstrapping involves resampling new datasets from our original dataset $\mathbf{X}$. This has become increasingly feasible, and boostrap-based methods have become increasingly popular.

Example: LLN and the sample mean

Let's go through the logic of Figure 1 in more detail in the case of the the law of large numbers. The LLN tells us that when we let the sample mean $\bar{X}_n$ define our asymptotic sequence Z_n, the resulting distributions F_{Z_n} eventually cluster all of their probability mass around the point μ, the sample mean. Figure 2 illustrates this point, through a simulation in `R`. I drew 1,000 $i.i.d$ samples of size n of a random variable X_i for which $P(X_i = 0) = 1/2$ and $P(X_i = 1) = 1/2$, representing a coin flip. Then, I plot a histogram of $\bar{X}_n$ across the 1,000 samples. This process is repeated for $n = 2$, $n = 10$, $n = 100$ and $n = 1,000$. You can think of this as illustrating Figure 1 for the specific population distribution F that describes a coin-flip. With $n = 2$, we see that we have a 50% chance of getting $\bar{X}_n$ of 0.5, which is the true "population mean'' of X_i: $\mu = \mathbb{E}[X_i] = 0.5$. Then 25% of the time we get $\bar{X}_n = 0$ (two flips of tails), and 25% of the time we get $\bar{X}_n = 1$ (two flips of heads).

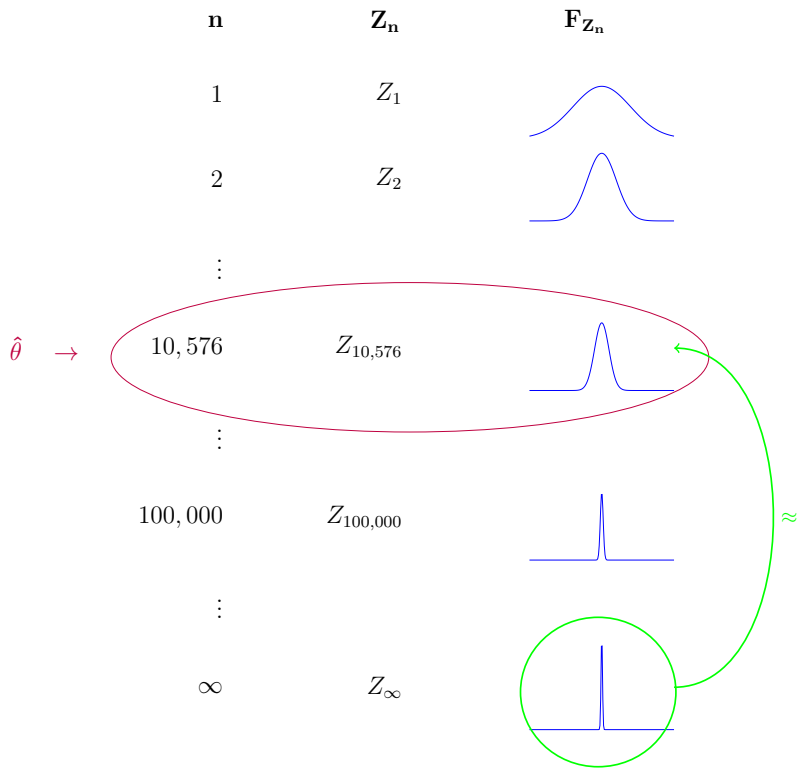

Figure 1: We are interested in the sampling distribution of some statistic $\hat{\theta}$, computed on our sample of $10,576$ observations. This is in general hard to compute. As a tool, we imagine a sequence of random variables $Z_1, Z_2, \ldots$ in which $\hat{\theta} = Z_{10,576}$. Asymptotic theory allows us to derive properties of F_{Z_∞}, the limiting distribution of Z_n as $n \to \infty$ (circled in green). Then we use F_{Z_∞} as an approximation to $F_{Z_{10,576}}$, which we justify by n being "large". The above figure depicts a situation in which $Z_n = \bar{X}_n$, so that the distribution of Z_n narrows to a point as $n \to \infty$ (by the LLN).

Thus, the distribution of $\bar{X}_n$ is not very well concentrated around $\mu = 0.5$.

The red vertical lines in Figure 2 illustrate the law of large numbers in action. They mark the points 0.45 and 0.55, which represent a $\epsilon = .05$ in Theorem . We can see that by the time $n = 100$, $P(|\bar{X}_n - 1/2| > 0.05)$ starts to become reasonably small; roughly $1/3$ of the mass of $\bar{X}_n$ is outside of $[0.45, 0.55]$. When $n = 1000$, there is an imperceptible chance of obtaining an $\bar{X}_n$ outside of the vertical red lines. If we continued this process for larger and larger n, we would see the mass of $\bar{X}_n$ continue to cluster closer and closer to $\mu = 1/2$. Regardless of how small a ϵ we choose, we can always find an n that fits as much of the mass as we want inside the corresponding red lines.

Note that the law of large numbers does *not* say that $P(|\bar{X}_n - \mu| > \epsilon)$ will necessarily monotonically decrease with n, for each n. For example, we can see that for $\epsilon = .05$, we have that $P(|\bar{X}_1 - \mu| > \epsilon)$ is 0.5 and $P(|\bar{X}_2 - \mu| > \epsilon)$ is about 0.25. All that the LLN says is that $P(|\bar{X}_1 - \mu| > \epsilon)$ will get (arbitrarily) small with n, for any value of ϵ.

%The next section will define two versions of the idea that a sequence of random variables Z_n

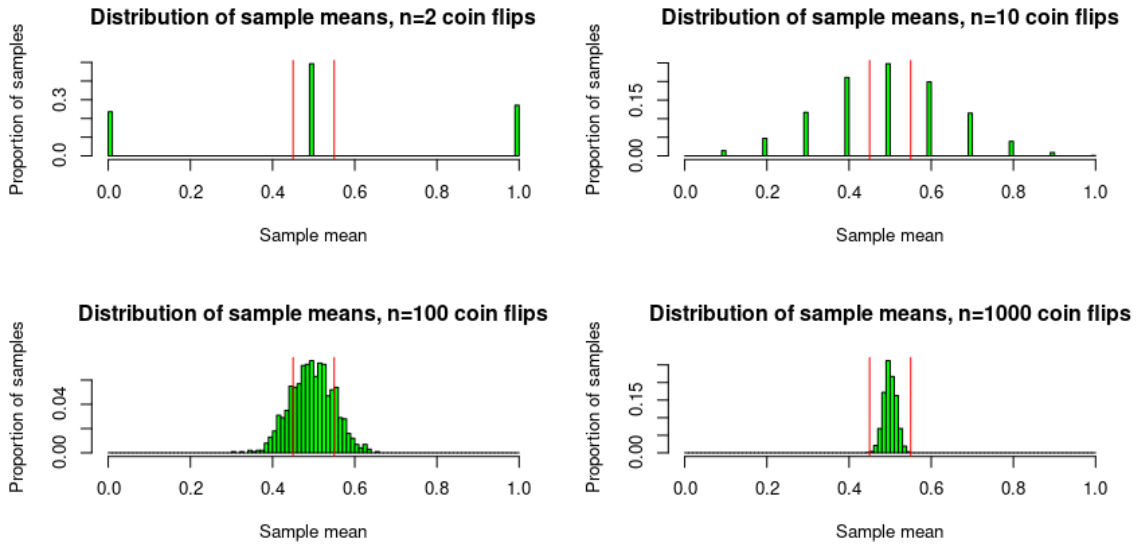

Figure 2: Distributions along the sequence $\bar{X}_n$ for a set of n i.i.d. coin flips. Red liness illustrate the mass of the distribution $\bar{X}_n$ that is more than .05 away from 1/2.

converges to something. When conceptualizing these ideas, keep Figure 2 in mind. The following is the R code I used to generate this figure, if you'd like to copy-paste it and experiment: %in mind how Figure 2 was generated. We fix the distribution of X_i, and

```
numsims<-1000
par(mfrow=c(2,2), main="Title")
for (n in c(2,10,100,1000)){
    results<-data.frame(simulation_num=integer(), sample_mean=double())
    for (x in 1:numsims) {
        thissample<-sample(c(0, 1), size = n, replace=TRUE)
        samplemean<-mean(thissample)
        results[x,] = c(x,samplemean)
    }

    h<-hist(results$sample_mean, plot=FALSE, breaks = seq(from=0, to=1, by
        =.01))
    h$density = h$density/100
    plot(h, freq=FALSE, main=paste0("Distribution of sample means, n=",n,"
        coin flips"), xlab="Sample mean", ylab="Proportion of samples", col="
        green")
    abline(v=c(.45,.55), col=c("red", "red"))
}
```