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.

Central Limit Theorem

PSE 8.1

In the previous set of notes, we showed that a large class of estimators is consistent for their target population parameters and discussed tools for establishing the consistency of estimators. But, often, we will need to know the entire sampling distribution. For this set of notes, our aim is to understand (or at least get a reasonable approximation) of the sampling distribution of estimators, particularly using large sample approximations. We will make heavy use of knowing an entire sampling distribution very soon, but for now, let's just say that we are in interested in high-level goals like assessing the accuracy of our estimator or in testing whether or not some theory is "compatible" with the sample that we have, and that (approximately) knowing the sampling distribution of an estimator will be useful for these goals.

Let's start with an easy case. 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 just follows by properties of normal distributions, particularly that the sum of independent, normally distributed random variables is also normally distributed. This implies that we would know the entire sampling distribution of $\bar{X}$ in this case.

Unfortunately, as we discussed earlier, 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 $\bar{X}$.

In cases where it is unreasonable to assume that the X_i follow a normal distribution, 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 of (a transformed version of) an estimator as $n \to \infty$.

Convergence in Distribution

PSE 8.2

As a step in this direction, we need to think some about what it even means to think about a sampling distribution as $n \to \infty$. In my view, consistency is fairly easy to wrap your mind around. When n gets big, it is fairly intuitive that a sample average converges to it population counterpart. Its also straightforward to consider other possible behaviors of various quantities that depend on the sample as $n \to \infty$. Many of these would converge to 0 or diverge (i.e., go to $\pm\infty$) as $n \to \infty$; for example $1/n \to 0$ as $n \to \infty$ or n itself $\to \infty$. In this section, we'll introduce an alternative notion of convergence where a sequence of random variables neither converges to 0 nor diverges, but instead behaves like a draw from some distribution as $n \to \infty$.

Before defining this notion of convergence, following the textbook, let's use G_n to denote the sampling distribution of the sequence Z_n and G to denote the sampling distribution of some random variable Z.

Definition. Let Z_n be a sequence of random variables with distribution $G_n(u) := \mathrm{P}(Z_n \leq u)$. Then, Z_n is said to **converge in distribution** to Z as $n \to \infty$ as $n \to \infty$ if, for all u at which $G(u) := \mathrm{P}(Z \leq u)$ is continuous, $G_n(u) \to G(u)$.

There are couple of points worth briefly mentioning. First, in economics, we typically use the terminology "convergence in distribution", but you might sometimes here the same concept referred to as "weak convergence". Second, it is common to refer to $G(u)$ as the "asymptotic distribution" or "limiting distribution" of Z_n. Third, the caveat "for all u at which $G(u)$ is continuous" is not practically important for any of the results that we'll show this semester or next the most common limiting distributions (e.g., normal, chi-square) are continuous. Fourth, convergence in distribution is a weaker concept than convergence in probability. In particular, if $Z_n \xrightarrow{p} c$, then $Z_n \xrightarrow{d} Z$ where Z is the "degenerate" random variable that takes the value c with probability 1.

Sample Mean

PSE 8.3

Now, let's think about trying to establish the asymptotic distribution of $\bar{X}$. We know from the weak law of large numbers that $\bar{X} \xrightarrow{p} \mathbb{E}[X]$. This implies that $\bar{X} \xrightarrow{d} \mathbb{E}[X]$. However, this is not very useful for thinking about the sampling distribution of $\bar{X}$ because it is degenerate. Recall that $\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$, where $\sigma^2 = \mathrm{var}(X)$. The reason for the asymptotic distribution of $\bar{X}$ itself being degenerate is due to the n in the denominator causing the variance to go to 0 as $n \to \infty$. This suggests normalizing $\bar{X}$ by a function of n so that the variance doesn't converge to 0 as $n \to \infty$. In particular, consider

$$Z_n = \sqrt{n}(\bar{X} - \mu)$$

Notice that $\mathbb{E}[Z_n] = 0$ and $\mathrm{var}(Z_n) = n\mathrm{var}(\bar{X}) = \sigma^2$. Thus, the variance of Z_n doesn't converge to 0 here. Further, notice that multiplying by $\sqrt{n}$ is "just right" in the sense that, if you multiplied by something that grows slower than $\sqrt{n}$ (say: $n^{1/3}$), then the variance would converge to 0; alternatively, if you were to choose something that grows faster than $\sqrt{n}$ (say: n), then the variance would diverge. Also, notice that we have subtracted off μ; without subtracting μ, $\mathbb{E}[\sqrt{n}\bar{X}] = \sqrt{n}\mathbb{E}[X]$ which will diverge as $n \to \infty$. All this to say, it seems at least possible that this particular sequence Z_n may not converge to 0 nor diverge as $n \to \infty$ which suggests that it might have a useful, non-trivial limiting distribution.

Central Limit Theorem

PSE 8.4-8.6

The main tool for establishing the limiting distribution of an estimator is the central limit theorem. It will be particularly useful for recovering the limiting distribution of normalized terms like Z_n above.

Central Limit Theorem: If X_i are iid and $\mathbb{E}[X^2] < \infty$, then

$$\sqrt{n}(\bar{X} - \mathbb{E}[X]) \xrightarrow{d} N(0, \sigma^2)$$

where $\sigma^2 = \text{var}(X)$.

We had previously established that $\sqrt{n}(\bar{X} - \mu)$ had mean 0 and variance σ^2. The central limit theorem additionally implies that, in large samples, $\sqrt{n}(\bar{X} - \mathbb{E}[X])$ (approximately) follows a normal distribution. This will be an extremely useful tool for us as we will soon exploit that we know a lot about the properties of normally distributed random variables.

The central limit theorem is quite remarkable. It says that *whatever* the distribution of X_i is, the limiting distribution of $\bar{X}_n$ (recentered by μ and rescaled by $\sqrt{n}$) will be a normal distribution. This striking result will pave the way for us to perform inference on the expectation of a random variable, without knowing its full distribution.

The practical value of the CLT is that it delivers an approximation to the distribution of $\bar{X}$. For large n, we know that $\sqrt{n}(\bar{X} - \mu)$ has approximately the distribution $N(0, \sigma^2)$. An alternative, equivalent way to write the CLT is that (under the same conditions)

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0, 1)$$

This implies that, in order to get a good guess of the distribution of $\sqrt{n}(\bar{X} - \mu)/\sigma$, we only need to have estimates of μ and σ which is feasible for us to do in applications, even in applications where we are unwilling to make distributional assumptions on the X_i.

Despite the usefulness of the central limit theorem, an intuition for the central limit theorem is harder to come by, especially relative to, say, the law of large numbers; in fact, the textbook even uses the word "mysterious" to describe the central limit theorem. For example, given our earlier discussion, you can think of $\sqrt{n}$ and $(\bar{X} - \mu)$ "fighting" against each other. From the law of large numbers, we know that $(\bar{X} - \mu)$ converges to 0 as $n \to \infty$; on the other hand, multiplying it by n raised to some positive power involves a term that will go to infinity. Multiplying by $\sqrt{n}$ essentially results in a "tie" so that $\sqrt{n}(\bar{X} - \mu)$ neither converges to 0 nor diverges to $\pm\infty$. Perhaps it is reasonable to think that, in the event of a "tie", that $\sqrt{n}(\bar{X} - \mu)$ would converge in distribution to something. But, a natural question is: why does it converge to a normal distribution instead of some other distribution? This is a good question, and, our proofs below will be mostly brute force. That said, the textbook does have a number of good (though relatively mathematical) explanations of the central limit theorem; for example the "moment investigation" calculations in Section 8.4 are tedious but show that the first 6 moments of $\sqrt{n}(\bar{X} - \mu)$ as $n \to \infty$ and are the same as the first 6

moments of a random variable that follows a $N(0, \sigma^2)$ distribution. This tentatively suggests that $\sqrt{n}(\bar{X} - \mu)$ might converge in distribution to normal.

Proof of Central Limit Theorem: Preliminary Helpful Results

We will provide a proof of the central limit theorem next. We will start by collecting several helpful intermediate results/tools before the main proof.

First, we will introduce (or maybe recall as you may have seen this in math for econ or micro class before) Taylor's theorem which is useful for approximating a function by a polynomial.

Taylor's theorem: Let s be a positive integer. If $f(x)$ is s times differentiable at a, then

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots + \frac{f^{(s)}(a)}{s!}(x-a)^s + \underbrace{h(x)(x-a)^s}_{r(x)}$$

where the last term, $r(x)$, is a remainder term where $h(x) \to 0$ when $x \to a$. Often, this remainder term is written as $r(x) = o(|x-a|^s)$ where the "little-oh" notation can be read as "has lower order than", which means that, when x converges to a, then the remainder term goes to 0 faster than $|x-a|^s$ does; for us, this will often mean that we can effectively ignore the last term as it will converge to 0 faster than the other terms.

Our most common usages of Taylor's Theorem will be for $s = 1$ (which amounts to a linear approximation of the function) or $s = 2$ (which amounts to a quadratic approximation of the function). In the next box, I give two closely discuss an alternative form of Taylor's Theorem and the closely related mean value theorem. We won't use either of these in the proof of the CLT, but this is a good place to mention them as they will be useful later.

Side-Comment:

Taylor's Theorem (mean-value form): Let s be a positive integer. If $f^{(s-1)}(x)$ is continuous on $[a,b]$ and differentiable on (a,b), then there exists a point $c \in (a,b)$ such that

$$f(b) = f(a) + f'(a)(b-a) + \frac{f''(a)}{2}(b-a)^2 + \cdots + \frac{f^{(s)}(c)}{s!}(b-a)^2$$

A very useful special case of this result is for $s = 1$, in which case we get the mean value theorem, which is

Mean-value Theorem: If $f(x)$ is continuous on $[a,b]$ and differentiable on (a,b) then there exists a point $c \in (a,b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b-a}$$

In other words, there exists a point between a and b where the slope of f is equal to the slope of the line connecting a and b. It's often useful to rearrange the expression in the mean value theorem so that

$$f(b) = f(a) + f'(c)(b-a)$$

The mean value theorem also (effectively) generalizes to the case where $f : \mathbb{R}^k \to \mathbb{R}^l$. In this case, letting a and b denote $k \times 1$ vectors

$$f(b) = f(a) + \nabla f(c)'(b-a)$$

where c is "between" a and b (this is the part where it is worth being careful in that in the vector case, we need to do this sort of expansion element-wise so that c can actually vary by row of $\nabla f(c)$ though it will still be in between a and b and therefore this caveat won't matter much for us) and where

$$\nabla f(c) := \frac{\partial f(u)'}{\partial u}\Big|_{u=c}$$

which is a $k \times l$ matrix of partial derivatives of f; that is,

$$f(u) = \begin{bmatrix} f_1(u) \\ f_2(u) \\ \vdots \\ f_l(u) \end{bmatrix}_{l \times 1} \quad \text{and} \quad \frac{\partial f(u)'}{\partial u} = \begin{bmatrix} \frac{\partial f_1(u)}{\partial u_1} & \frac{\partial f_2(u)}{\partial u_1} & \cdots & \frac{\partial f_l(u)}{\partial u_1} \\ \frac{\partial f_1(u)}{\partial u_2} & \frac{\partial f_2(u)}{\partial u_2} & \cdots & \frac{\partial f_l(u)}{\partial u_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1(u)}{\partial u_k} & \frac{\partial f_2(u)}{\partial u_k} & \cdots & \frac{\partial f_l(u)}{\partial u_k} \end{bmatrix}_{k \times l}$$

Next we will cover/review some properties of moment generating functions. Recall that when we talked about the moment generating function of a random variable X, we noted the property of moment generating functions that: if two random variables have the same moment generating function then they follow the same distribution. A related result is the following

Levy's Continuity Theorem: If $\mathbb{E}[\exp(tZ_n)] \to \mathbb{E}[\exp(tZ)]$ for every $t \in \mathbb{R}$, then $Z_n \xrightarrow{d} Z$.

Levy's continuity theorem implies that, if $M_n(t) := \mathbb{E}[\exp(tZ_n)]$ converges to a limit function $M(t) = \mathbb{E}[\exp(tZ)]$, then the distribution of Z_n will converge to the distribution of Z. The intuition here is that, if all the moments of Z_n converge to all the moments of Z, then Z_n will converge in distribution to Z. This will be the strategy that we use to prove the central limit theorem.

Next, recall that if $Z \sim N(\mu, \sigma^2)$, then its moment generating function is given by

$$M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$$

Further, recall that if $X_1, \ldots, X_n$ are independent, then

$$\mathbb{E}\left[\exp\left(t\left(\sum_{i=1}^{n} X_i\right)\right)\right] = \prod_{i=1}^{n} \mathbb{E}[\exp(tX_i)]$$

Finally, recall that $\exp(x)$ (can be) defined as

$$\exp(x) = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$$

Proof of Central Limit Theorem

Now, let's write the proof. We will start with the moment generating function of $Z_n = \sqrt{n}(\bar{X} - \mu)$ (notice that $Z_n = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}}$) and we are aiming to show that this converges to the moment generating function of a normally distributed random variable with mean 0 and variance σ^2.

$$\begin{aligned}
M_{Z_n}(t) &= \mathbb{E}\left[\exp\left(tZ_n\right)\right] \\
&= \mathbb{E}\left[\exp\left(t\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}}\right)\right] \\
&= \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\frac{t}{\sqrt{n}}(X_i - \mu)\right)\right] \\
&= \left(\mathbb{E}\left[\exp\left(\frac{t}{\sqrt{n}}(X - \mu)\right)\right]\right)^n
\end{aligned}$$

where the first equality holds by the definition of mgf, the second by the definition of Z_n, the third by because X_i are iid (and by the related discussion in the previous section), and the last equality holds because X_i are iid (so that the expected value does not change with i).

For the inside term, notice that $\mathbb{E}\left[\exp\left(\frac{t}{\sqrt{n}}(X-\mu)\right)\right] = M_{(X-\mu)}\left(\frac{t}{\sqrt{n}}\right)$, and therefore, using a second order Taylor expansion, we have that

$$M_{(X-\mu)}\left(\frac{t}{\sqrt{n}}\right) = M_{(X-\mu)}(0) + M'_{(X-\mu)}(0)\left(\frac{t}{\sqrt{n}} - 0\right) + \frac{M''_{(X-\mu)}(0)}{2}\left(\frac{t}{\sqrt{n}} - 0\right)^2 + \left(\frac{t}{\sqrt{n}} - 0\right)^2 h\left(\frac{t}{\sqrt{n}}\right)$$

To be clear here, $t/\sqrt{n}$ is playing the role of x, 0 is playing the role of a, and $s = 2$ in Taylor's Theorem, and where $h(t/\sqrt{n}) \to 0$ when $t/\sqrt{n} \to 0$. Now, we need to provide the expressions for $M_{(X-\mu)}(0)$, $M'_{(X-\mu)}(0)$, and $M''_{(X-\mu)}(0)$ to arrive at

$$\begin{aligned} M_{(X-\mu)}\left(\frac{t}{\sqrt{n}}\right) &= \mathbb{E}\left[\exp\left(0(X-\mu)\right)\right] + \mathbb{E}\left[(X-\mu)\exp\left(0(X-\mu)\right)\right]\frac{t}{\sqrt{n}} + \mathbb{E}\left[(X-\mu)^2\exp\left(0(X-\mu)\right)\right]\frac{t^2}{2n} + o(t^2/n) \\ &= 1 + \sigma^2\frac{t^2}{2n} + o(t^2/n) \end{aligned}$$

Thus, from plugging this back into the mgf of Z_n above, we have that

$$M_{Z_n}(t) = \left(1 + \sigma^2\frac{t^2}{2n} + o(t^2/n)\right)^n$$

Moreover,

$$\begin{aligned} \lim_{n\to\infty} M_{Z_n}(t) &= \lim_{n\to\infty}\left(1 + \sigma^2\frac{t^2}{2n} + o(t^2/n)\right)^n \\ &= \lim_{n\to\infty}\left(1 + \sigma^2\frac{t^2}{2n}\right)^n \\ &= \exp\left(\frac{\sigma^2 t^2}{2}\right) \\ &= M_Z(t) \end{aligned}$$

where Z is a random variable that follows a normal distribution with mean 0 and variance σ^2 and where the second equality holds because the lower order terms vanish as $n \to \infty$ (see box below for a more detailed discussion), the third equality holds by the definition of $\exp(x)$ from the previous section, and the last equality holds by recalling the moment generating function of a normally distribution random variable.

Side-Comment: If we can ignore the $o(t^2/n)$ term then we are done. To show that this term indeed does not contribute in the limit, consider taking the natural logarithm of both sides of the above equation (since the log is continuous function, it preserves limits):

$$\begin{aligned}\lim_{n\to\infty} \log\left\{\left(1+\frac{\sigma^2t^2}{2n}+o(t^2/n)\right)^n\right\} &= \lim_{n\to\infty} n\cdot\log\left(1+\frac{\sigma^2t^2}{2n}+o(t^2/n)\right)\\ &= \lim_{n\to\infty} n\cdot\left(\frac{\sigma^2t^2}{2n}+o(t^2/n)\right) = \lim_{n\to\infty}\left(\frac{\sigma^2t^2}{2}+o(t^2)\right)\cdot\\ &= \frac{\sigma^2t^2}{2}\end{aligned}$$

where, for the second equality, we've used the Taylor theorem for the natural logarithm: $\log(1+z) = \log(1+0) + \frac{1}{1+0}(z-0) + o(z-0) = z + o(z)$. This implies that

$$\lim_{n\to\infty}\left(1+\sigma^2\frac{t^2}{2n}+o(t^2/n)\right)^n = \exp\left(\frac{\sigma^2t^2}{2}\right)$$

as we wanted.

Multivariate Central Limit Theorem

Also, the central limit theorem also holds in the case where X is a random vector under essentially the same conditions.

Multivariate Central Limit Theorem: If $X_i \in \mathbb{R}^k$ are iid and $\mathbb{E}||X||^2 < \infty$, then

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

where $\mathbf{V} = \text{var}(X) = \mathbb{E}[(X-\mathbb{E}[X])(X-\mathbb{E}[X])']$ (which is a $k \times k$ matrix).

(Extended) Continuous Mapping Theorem

PSE 8.9

Next, we consider a version of the continuous mapping theorem for convergence in distribution

Extended Continuous Mapping Theorem: If $Z_n \xrightarrow{d} Z$ and $h(\cdot)$ has the set of discontinuity points D_h such that $P(Z \in D_h) = 0$, then $h(Z_n) \xrightarrow{d} h(Z)$ as $n \to \infty$.

The extended continuous mapping theorem says that applying a continuous function to a sequence of random variables preserves convergence in distribution. The condition about discontinuity points allows for the function to not be strictly continuous, but that the probability of being at a discontinuous point being equal to 0 — this would typically be a discrete set of discontinuity points.

Example: The book gives the example of $h(u) = u^{-1}$. This function is discontinuous at $u = 0$ but continuous everywhere else. But, if, for example, $Z_n \xrightarrow{d} Z$, then the probability $Z = 0$ is 0, then this function satisfies the conditions of the extended continuous mapping theorem. In particular, it implies that $Z_n^{-1} \xrightarrow{d} Z^{-1}$.

Example: Let $Z_n = \sqrt{n}(\bar{X}_n - \mu)$. Then by the CLT and CMT: $Z_n^2 = n\left(\bar{X}_n - \mu\right)^2 \xrightarrow{d} \chi_1^2$, where χ_1^2 is the chi-squared distribution with one degree of freedom (this is the distribution of a standard normal $N(0, 1)$ random variable squared).

There are several special cases of the extended continuous mapping theorem that are so common that they are grouped together and called Slutsky's Theorem.

Slutsky's Theorem: If $Z_n \xrightarrow{d} Z$ and $Y_n \xrightarrow{p} c$, then

1. $Z_n + Y_n \xrightarrow{d} Z + c$
2. $Z_n Y_n \xrightarrow{d} Zc$
3. $\frac{Z_n}{Y_n} \xrightarrow{d} \frac{Z}{c}$

These cover the most common operations that we'll encounter for sequences of random variables.

The Delta Method

PSE 8.9

Now consider the case where we are interested in developing the limiting distribution of the plug-in estimator $\hat{\beta} = h(\hat{\theta})$. Earlier we showed that the continuous mapping theorem implied that $\hat{\beta} \xrightarrow{p} \beta$. For this section, we will be interested in establishing the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta)$.

Even if we have that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2)$, the extended continuous mapping theorem does not directly provide the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta)$; notice that, in general, $\sqrt{n}(\hat{\beta} - \beta) \neq h(\sqrt{n}(\hat{\theta} - \theta))$ which implies we cannot just apply the extended continuous mapping theorem.

To establish the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta)$, we will use an approach that is called the delta method.

Delta Method: If $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} Z$ and $h(u)$ is continuously differentiable in neighborhood of θ, then

$$\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n}(h(\hat{\theta}) - h(\theta)) \xrightarrow{d} h'(\theta)Z$$

where $h'(\theta) = \frac{d\,h(u)}{du}\Big|_{u=\theta}$.

The most common case for us will be when $Z \sim N(0, \sigma^2)$. In this case, given that $h'(\theta)$ is a constant, then we would have that

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, h'(\theta)^2 \sigma^2)$$

The version of the Delta method above is for the case where θ is scalar and $h : \mathbb{R} \to \mathbb{R}$. But it will generalize to the case where θ is a $k \times 1$ vector and $h : \mathbb{R}^k \to \mathbb{R}^l$. I'll just cover the most common case where we know that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathbf{V})$ for some $k \times k$ variance matrix $\mathbf{V}$. In this case,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \nabla h(\theta)' \mathbf{V} \nabla h(\theta))$$

where

$$\nabla h(\theta) := \frac{\partial h(u)'}{\partial u}\Big|_{u=\theta}$$

which is a $k \times l$ matrix. To conclude this section, let's provide a proof of the Delta method; because its not much more complicated, we will let θ be a $k \times 1$ vector and $h : \mathbb{R}^k \to \mathbb{R}^l$; and we'll focus on the case where $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathbf{V})$ which is by far the leading case.

Proof: The key step is to use the mean value theorem to write

$$h(\hat{\theta}) = h(\theta) + \nabla h(\theta^*)'(\hat{\theta} - \theta)$$

for some θ^* between $\hat{\theta}$ and θ. Thus, we can write

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta) &= \sqrt{n}(h(\hat{\theta}) - h(\theta)) \\ &= \nabla h(\theta^*)' \sqrt{n}(\hat{\theta} - \theta) \end{aligned}$$

Now, we have that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathbf{V})$ (by assumption), we also have that $\theta^* \xrightarrow{p} \theta$ because it is between $\hat{\theta}$ and θ and $\hat{\theta} \xrightarrow{p} \theta$, thus, by the continuous mapping theorem (and because h is continuously differentiable by assumption), we have that $\nabla h(\theta^*) \xrightarrow{p} \nabla h(\theta)$; finally, by the extended continuous mapping theorem, we therefore have that

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \nabla h(\theta)' Z$$

where $Z \sim N(0, \mathbf{V})$, and, therefore, $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, h(\theta)' \mathbf{V} h(\theta))$.

Covariance Matrix Estimation

PSE 8.12

In practice, the limiting distributions that we have established depend on (unknown) population quantities. For example, we just showed that $\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0, \nabla h(\theta)'\mathbf{V}\nabla h(\theta))$, but the asymptotic variance, $\nabla h(\theta)'\mathbf{V}\nabla h(\theta)$ depends on θ and $\mathbf{V}$ which are population quantities. To conserve on notation, let's define $\mathbf{V}_\beta = \nabla h(\theta)'\mathbf{V}\nabla h(\theta)$. The natural estimator of $\mathbf{V}_\beta$ is

$$\nabla h(\hat{\theta})'\hat{\mathbf{V}}\nabla h(\hat{\theta})$$

where (given $\theta = \mathbb{E}[g(X)]$),

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \quad \text{and} \quad \hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n} \big(g(X_i) - \hat{\theta}\big)\big(g(X_i) - \hat{\theta}\big)'$$

and, for example, in the simpler case when $\theta = \mathbb{E}[X]$, the expression for $\hat{\mathbf{V}}$ would simplify to

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'$$

In general, this estimator of $\mathbf{V}$ will be biased, so you could alternatively divide by $n-1$ to get an unbiased estimator, though this may not matter much when n is large.

Stochastic Order Symbols

PSE 8.14

Earlier in the notes, for some non-random sequence a_n, we used the "little oh" notation $a_n = o(f(n))$ to indicate that $f(n)a_n \to 0$ as $n \to \infty$. The most common versions of $f(n)$ are (i): $f(n) = 1$ so that $a_n = o(1)$ which just means that the sequence converges to 0 or (ii) $n^{-1/2}$ so that $a_n = o(n^{-1/2})$ which means that the sequence converges to 0 faster than $n^{-1/2}$ does.

A related notation is "big oh" notation $a_n = O(f(n))$ to indicate that $f(n)a_n$ is uniformly bounded in n. In the same two common cases as above, $a_n = O(1)$ indicates that the sequence is uniformly bounded (for this case you can think of this as meaning that the sequence behaves neither goes to 0 nor diverges as $n \to \infty$). If $a_n = O(n^{-1/2})$, it would indicate that $\sqrt{n}a_n = O(1)$ (i.e., is uniformly bounded).

Its sometimes useful to have a similar notation for random vectors. We will sometimes write

$$Z_n = o_p(1)$$

to indicate that $Z_n \xrightarrow{p} 0$, and you would say that Z_n is "little oh P one". Similarly, $Z_n = o_p(n^{-1/2})$ indicates that $\sqrt{n}Z_n \xrightarrow{p} 0$, or, equivalently, $\sqrt{n}Z_n = o_p(1)$.

Similarly, we will sometimes write

$$Z_n = O_p(1)$$

which is said "big oh P one" to indicate that Z_n is "bounded in probability" (you can see the textbook for a formal definition, but you can think of this as saying that extremely large values of Z_n are very rare). As you would expect, we would write $Z_n = O_p(n^{-1/2})$ to indicate that $\sqrt{n}Z_n = O_p(1)$. One important thing to remember here is that, if Z_n converges in distribution, then it is $O_p(1)$.

There are useful rules for working with random sequence that are $O_p(1)$ and/or $o_p(1)$. The textbook provides a number of these, but here are probably the two most useful:

$$O_p(1) + o_p(1) = O_p(1) \quad \text{and} \quad O_p(1)o_p(1) = o_p(1)$$

The intuition for the first one is that if you add something that is bounded in probability to something else that converges in probability to 0, then their sum will be bounded in probability. For the second one, if you multiply something that is bounded in probability to something that converges in probability to 0, then their product will converge to 0. These rules are implications of the continuous mapping theorem.

This sort of notation is often useful for keeping track of "leftover" terms that may be complicated by end up converging to 0.

Monte Carlo Simulations

Let's return to our previous example of flipping a coin and try to see the law of large numbers, central limit theorem, and continuous mapping theorem in practice.

First, let's consider estimating

$$p = \mathbb{E}[X] \quad \text{and} \quad \beta = p^2$$

by

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \hat{\beta} = \hat{p}^2$$

Furthermore, we have from the CLT and CMT (and because X is Bernoulli) that

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0, p(1-p)) \quad \text{and} \quad \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V)$$

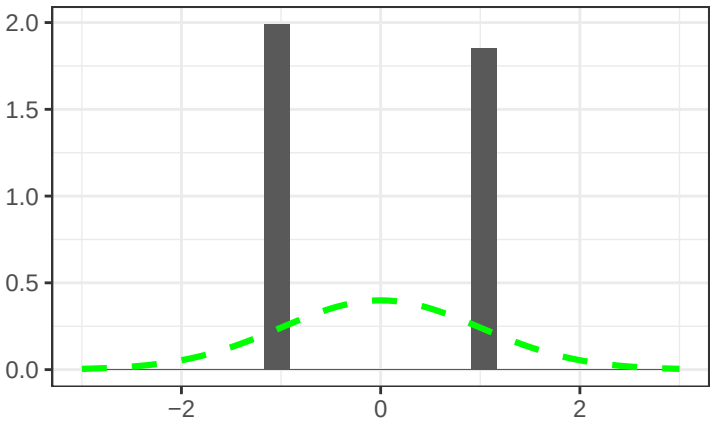
where $V = (2p)^2 p(1-p) = 4p^3(1-p)$ (this holds because $\beta = h(p) = p^2 \implies h'(p) = 2p$).

I am going to bring in some of the functions we used previously for simulating coin flips.

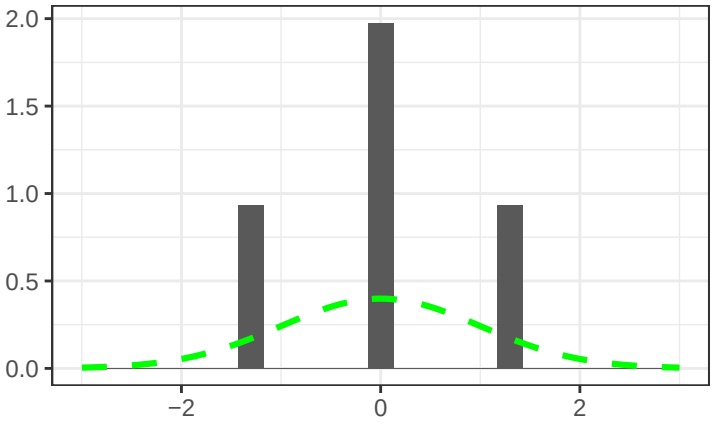
```
# load packages
library(ggplot2)
library(dplyr)

set.seed(1234)
```

```r
p <- 0.5 # prob of heads

# function to flip a coin with probability p
flip <- function(p) {
  sample(c(0,1), size=1, prob=(c(1-p,p)))
}

# function to generate a sample of size n
generate_sample <- function(n,p) {
  Y <- c()
  for (i in 1:n) {
    Y[i] <- flip(p)
  }
  Y
}

# function to carry out Monte Carlo simulations
# returns a vector of length nsims
# containing standardized versions of phat or bhat
# (depending on the argument `which_est`) from
# each simulation that come from, for example,
# sqrt(n)(phat-p)/sqrt(V)
mc <- function(n, p=0.5, nsims=1000, which_est="p") {
  phat <- c() # vector to hold estimated p
  bhat <- c() # vector to hold estimated beta
  for (i in 1:nsims) {
    Y <- generate_sample(n,p)
    phat[i] <- mean(Y)
    bhat[i] <- phat[i]^2
  }

  # subtract mean and multiply by sqrt(n)
  p_stand <- sqrt(n)*(phat-p)/sqrt(p*(1-p))
  b_stand <- sqrt(n)*(bhat-p^2)/sqrt(4*p^3*(1-p))

  if (which_est == "p") {
    return(p_stand)
  } else {
```

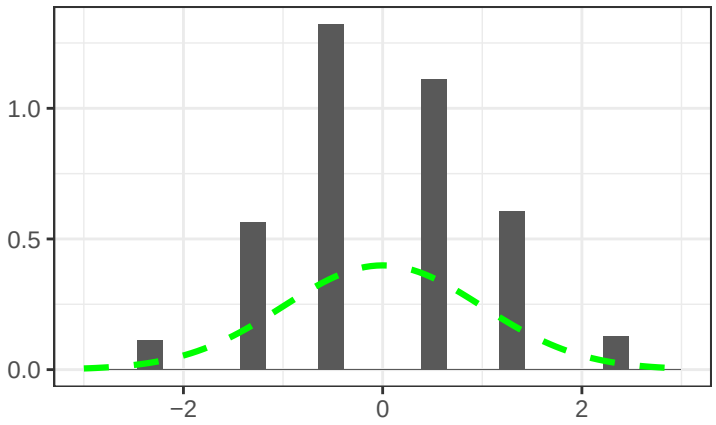
```
    return(b_stand)
  }
}

# function to make plots of our simulation results
# note: in the plot we report the fraction of
# observations in a "bin" divided by the bin
# width.  This is for approximating a pmf using
# a type of histogram; to instead get the pmf,
# you can just divide the heights of each bar by
# the bin width (in that case the bars will sum
# to 1) but what we are doing here mimics: f(x) =
# d/dx (F(x)) where the bin width plays the role of
# dx.
plot_sim_results <- function(sim_results) {
  plot_df <- data.frame(sim=sim_results)
  sep <- .26 # this is the bin width below
  ggplot(plot_df, aes(x=sim)) +
    xlim(c(-3,3)) +
    geom_histogram(aes(x=sim, y=..density..),
                   binwidth=sep) +
    stat_function(fun=dnorm,
                  n=101,
                  args=list(mean=0, sd=1),
                  color="green",
                  linetype="dashed",
                  size=1.1) +
    theme_bw() +
    ylab("") + xlab("")
}

# show results for phat for different values of n
# (plots include an overlay of N(0,p(1-p)))

# n=1
mc1 <- mc(1)
plot_sim_results(mc1)
```



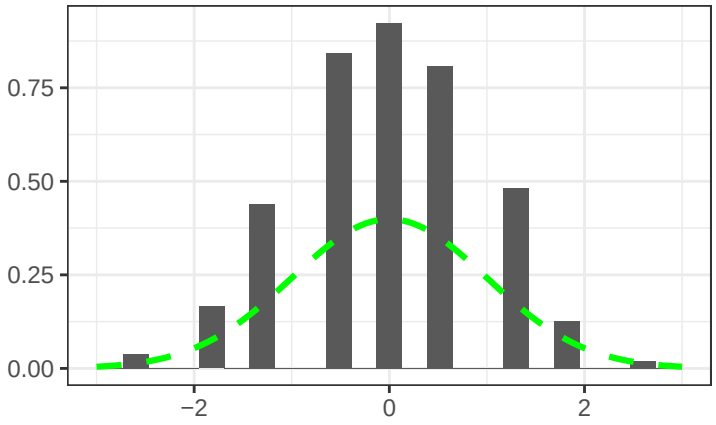

```
# n=2
mc2 <- mc(2)
plot_sim_results(mc2)
```



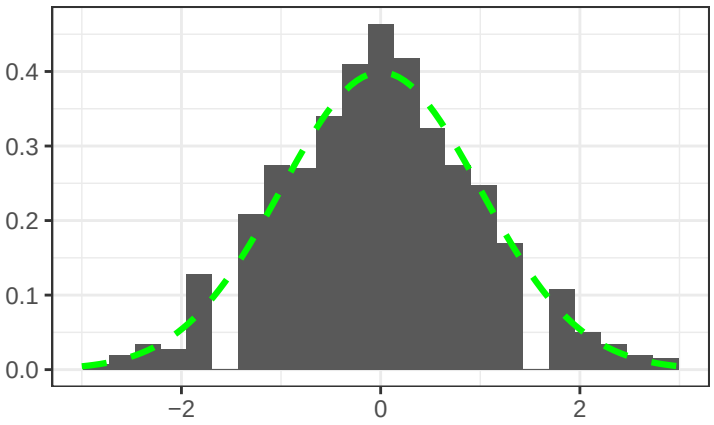

```
# n=5
mc5 <- mc(5)
plot_sim_results(mc5)
```



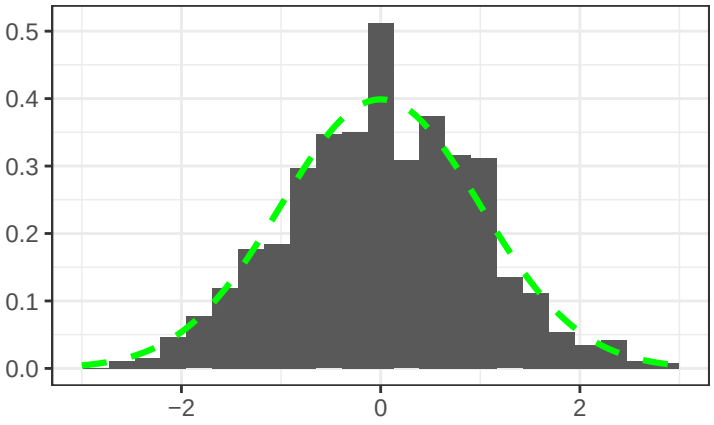

```
# n=10
mc10 <- mc(10)
plot_sim_results(mc10)
```




```
# n=50
mc50 <- mc(50)
plot_sim_results(mc50)
```



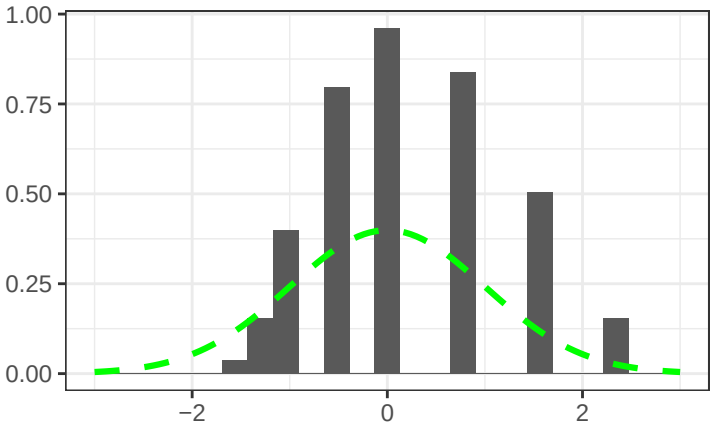

```
# n=1000
mc1000 <- mc(1000)
plot_sim_results(mc1000)
```



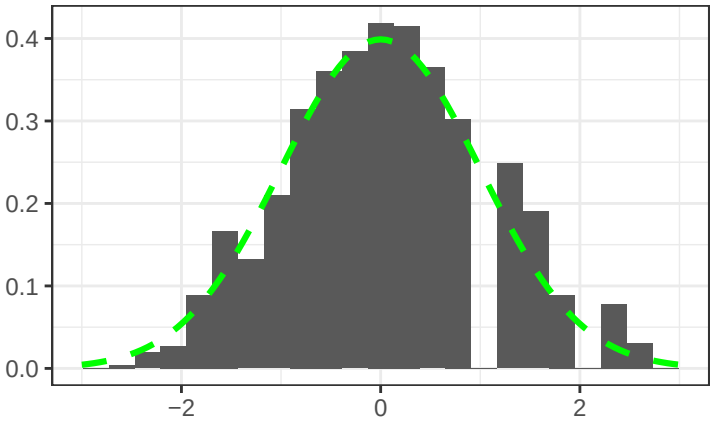
From the above results, we can see that (as we would expect) the normal approximation is not working well at all when $n = 1$; in fact, there are only two possible values that $\sqrt{n}(\hat{p} - p)$ can take in this case: 0.5 or -0.5. For larger values of n, the normal approximation starts to work better; for $n = 50$, it seems to be working reasonably well.

To conclude this section, let's do a smaller version of this for estimating β.

```
# n=10
mcb10 <- mc(10, which_est="b")
plot_sim_results(mcb10)
```




```
# n=50
mcb50 <- mc(50, which_est="b")
plot_sim_results(mcb50)
```



These are fairly interesting results. When $n = 10$ the normal approximation is not working too well; moreover, you can seemingly see some bias here (this is expected as $\beta = p^2$ is a nonlinear function of p and, therefore, $\hat{\beta}$ is consistent for β but not unbiased). However, by $n = 50$, again the normal approximation seems to work pretty well.

As a final comment: Thought experiments like this simulation experiment are useful for getting intuition about the CLT. Accoringly, you often hear descriptions of the CLT along the lines of: "the sample mean becomes normal as the sample gets bigger and bigger". This isn't wrong, but can be a little misleading. A given real-world sample never gets bigger: it always has a single finite size n! Similarly, the sample size n never "goes to infinity"–though we can get pretty close by simulating a sequence of samples on a computer! Imagining an infinite sequence of samples having means $\bar{X}_1$, $\bar{X}_2$, and so on, is just a useful abstraction.