10: Introduction to Asymptotic Theory - Revised
ECON 837

Brian Krauth (adapted from notes by Simon Woodcock), Spring 2010

To this point, our discussion has focused on the finite sample properties of estimators (primarily least squares, but also sample means, variances, and the like), and finite sample inference under exact distributional assumptions. In many situations, we are unable or unwilling to make exact distributional assumptions about unknown quantities. Without such assumptions we are usually unable to determine the (finite sample) properties of an estimator. In particular, we will be unable to derive an estimator's exact sampling distribution. This generates obvious problems for inference. Thus we turn to asymptotic theory, which is concerned with the properties of estimators as the sample size becomes infinite.

Of course an infinite sample size is a somewhat fanciful notion, but one that generates some powerful results. An example will serve to illustrate.

Example 1 *Suppose we observe a random sample of size n on some random variable X. Suppose the probability distribution of X is unknown, with mean μ and finite variance σ^2. We are interested in obtaining an estimate of μ. Recall our estimator $\bar{x} = \frac{1}{n}\sum_i x_i$, the sample mean. We know already that $\bar{x}$ is a random variable. If we index this estimator by the sample size, i.e., $\bar{x}_n$, the question is how the sampling distribution of the sample mean behaves as the sample size gets large. We know already that the sample mean is an unbiased estimator of μ, that is $E[\bar{x}_n] = \mu$ for all n. That is, unbiasedness is a finite sample property (it is independent of sample size). We see that regardless of sample size, the sampling distribution of $\bar{x}$ is centered (in an expectational sense) over μ. It seems intuitive that our estimator should get "better" as the sample increases (i.e., as we observe more information). We note this from the sampling variance of the sample mean, $\sigma_n^2 = \sigma^2/n$. The sampling distribution becomes less dispersed around μ as the sample size increases. In fact,*

$$\lim_{n\to\infty} Var[\bar{x}_n] = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0.$$

See figure 1. That is, as the sample size $n \to \infty$, the sampling distribution converges to point mass at the true value of the mean μ.

Most asymptotic results are based on two important theorems: "the" law of large numbers and "the" central limit theorem. It turns out there are numerous versions of each. We'll see (and prove!) a number of them this lecture. But first we need to develop some convergence concepts.

Convergence of Random Sequences

All of our notions of convergence will involve a sequence of random variables $\{X_n, n = 1, 2, ...\}$ converging to some random variable X. An important thing to remember is that a constant is an example a random variable, and in many applications we will be interested in showing that a particular sequence of estimators converges to the true value of a parameter we want to estimate.

We will also be interested what happens to sequences of random vectors. That will turn out to be a relatively simple extension.

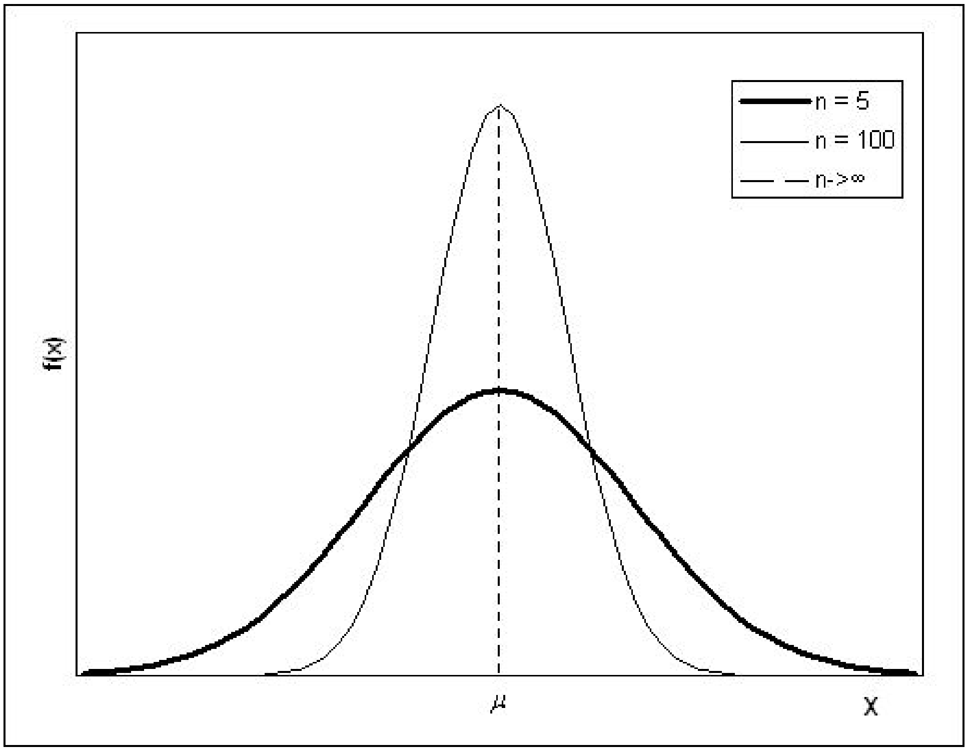

Figure 1: Sampling distribution of the sample mean $\bar{x}_n$ as n increases

Weak convergence

We'll begin with some weak notions of convergence, including the one illustrated in Example 1.

Definition 2 (Convergence in Mean Square) *Let $\{X_n, n = 1, 2, ...\}$ be a sequence of random variables, and X a random variable. Then X_n* ***converges in mean square to*** *X (or* ***converges in quadratic mean****) if*

$$\lim_{n\to\infty} E\left[(X_n - X)^2\right] = 0$$

which we write $X_n \xrightarrow{m.s.} X$.

If we think of X as a parameter of interest, and X_n as an estimator of X from a sample of size n, then convergence in mean square occurs if the mean squared error goes to zero as $n \to \infty$. If X_n is unbiased, then this is equivalent to the sampling variance of X_n converging to zero as $n \to \infty$.

Example 3 *Let's apply this definition to Example 1. Let $X_n = \bar{x}_n$ and $X = \mu$. Then*

$$\lim_{n\to\infty} E\left[(\bar{x}_n - \mu)^2\right] = \lim_{n\to\infty} Var\left[\bar{x}_n\right] = \lim_{n\to\infty} \left(\frac{\sigma^2}{n}\right) = 0$$

so $\bar{x}_n \xrightarrow{m.s.} \mu$.

Showing convergence in mean square is usually easiest using the following theorem. (try to prove for yourself, based on the intuition given above).

Theorem 4 *Let* $\{X_n, n = 1, 2, ...\}$ *be a sequence of random variables, with* $E[X_n] = \mu_n$ *and variance* $Var[X_n] = \sigma_n^2$*. Then if* $\lim_{n\to\infty} \mu_n = c$ *and* $\lim_{n\to\infty} \sigma_n^2 = 0$*, then* $X_n \xrightarrow{m.s.} c$*.*

Convergence in mean square turns out to be a special case of another type of convergence, which we call convergence in probability or weak convergence.

Definition 5 (Convergence in Probability) *Let* $\{X_n, n = 1, 2, ...\}$ *be a sequence of random variables, and* X *a random variable. Then* X_n ***converges in probability*** *to* X *(or* ***converges weakly*** *to* X*), if for every* $\varepsilon > 0$*,*

$$\lim_{n\to\infty} \Pr\left[|X_n - X| > \varepsilon\right] = 0$$

or equivalently,

$$\lim_{n\to\infty} \Pr\left[|X_n - X| \leq \varepsilon\right] = 1$$

which we write $plim\, X_n = X$ *or* $X_n \xrightarrow{p} X$*.*

We usually call the plim of a random variable its **probability limit**. The notion behind weak convergence is pretty simple. As n increases, the probability that X_n takes a value “far” from X gets closer and closer to zero.

Example 6 *Let*

$$X_n = \begin{cases} n & \text{with probability } \frac{1}{n} \\ 0 & \text{with probability } 1 - \frac{1}{n} \end{cases}$$

It is easy to show that $X_n \xrightarrow{p} 0$*. But* $E(X_n - 0)^2 = n^2/n = n$*, so* X_n *does not converge in mean square to* 0*.*

Strong forms of convergence

Definition 7 (Almost Sure Convergence) *Let* $\{X_n, n = 1, 2, ...\}$ *be a sequence of random variables, and* X *a random variable. Then* X_n ***converges almost surely*** *to* X *(or* ***converges strongly*** *to* X*, or* ***converges with probability one*** *to* X*), if for every* $\varepsilon > 0$*,*

$$\Pr\left[\lim_{n\to\infty} |X_n - X| > \varepsilon\right] = 0$$

or equivalently,

$$\Pr\left[\lim_{n\to\infty} |X_n - X| \leq \varepsilon\right] = 1$$

which we write $X_n \xrightarrow{a.s.} X$ *or* $X_n \to X$ *w.p.1.*

This appears similar to the definition of weak convergence. The notable difference is that under weak convergence the limit is outside the probability statement. Weak convergence is a statement about the limit of a sequence of probabilities. Namely, that as the sample size becomes arbitrarily large, those probabilities (the probability that X_n is "far" from X) become arbitrarily small. In contrast, strong convergence is a statement about the probabilistic behaviour of the limit of a sequence. Strong convergence tells us that as n becomes arbitrarily large, the probability that the sequence X_n doesn't converge to X vanishes. If you like, once the sequence X_n approaches X, it remains "close" to X (it doesn't diverge). Alternately, the possible realizations of the random sequence X_n such that it doesn't converge to X occur with probability zero.

Strong convergence implies weak convergence. This is most easily seen from an alternate definition of strong convergence.

Definition 8 (Almost Sure Convergence, Alternate) *Let $\{X_n, n = 1, 2, ...\}$ be a sequence of random variables, and X a random variable. Then X_n* ***converges almost surely*** *to X if for every $\varepsilon > 0$,*

$$\lim_{N\to\infty} \Pr\left[|X_n - X| > \varepsilon \textit{ for all } n > N\right] = 0$$

or equivalently,

$$\lim_{N\to\infty} \Pr\left[|X_n - X| \leq \varepsilon \textit{ for all } n > N\right] = 1.$$

The next two examples demonstrate the ideas of convergence in probability and almost sure convergence. The first example considers a random sequence that converges almost surely. The second considers a random sequence that converges in probability but not almost surely.

Example 9 *Suppose a random variable $X \sim U[0,1]$. Define the sequence $X_n = X + X^n$. When $X \in [0,1)$, $\lim_{n\to\infty} |X_n - X| = 0$. That is, for $X \in [0,1)$, the sequence X_n converges pointwise to X. However when $X = 1$, we know $X_n = 2$ for all n. But since convergence occurs on the set $X \in [0,1)$ and $\Pr[X \in [0,1)] = 1$, $X_n \xrightarrow{a.s.} X$.*

Example 10 *Suppose a random variable $X \sim U[0,1]$. Define the sequence X_n as follows*

$$\begin{aligned}
X_1 &= X + \mathbf{1}(X \in [0,1]) \\
X_2 &= X + \mathbf{1}\left(X \in \left[0, \frac{1}{2}\right]\right) \\
X_3 &= X + \mathbf{1}\left(X \in \left[\frac{1}{2}, 1\right]\right) \\
X_4 &= X + \mathbf{1}\left(X \in \left[0, \frac{1}{3}\right]\right) \\
X_5 &= X + \mathbf{1}\left(X \in \left[\frac{1}{3}, \frac{2}{3}\right]\right) \\
X_6 &= X + \mathbf{1}\left(X \in \left[\frac{2}{3}, 1\right]\right)
\end{aligned}$$

etc., where the function $\mathbf{1}(A)$ *takes value* 1 *if* A *is true, and* 0 *otherwise. It is straightforward to see* $X_n \overset{p}{\to} X$. *As* $n \to \infty$, $\Pr[|X_n - X| > \varepsilon]$ *is the probability that* X *lies in an interval whose length approaches zero. However,* X_n *does not converge to* X *almost surely. That is, there is no* $X \in [0,1]$ *such that* $\lim_{n\to\infty} X_n = X$. *For every* X *and every* N, *we see that for* $n > N$, X_n *alternates between the values* X *and* $X + 1$ *infinitely often.*

Convergence in distribution

The weakest form of convergence we usually work with is convergence in distribution.

Definition 11 (Convergence in Distribution) *Let* $\{X_n, n = 1, 2, ...\}$ *be a sequence of random variables with cdfs* $\{F_n(x), n = 1, 2, ...\}$, *and* X *a random variable with cdf* $F(x)$. *If*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at all continuity points of $F(x)$. *then we say* X_n ***converges in distribution*** *to* X, *denoted* $X_n \xrightarrow{d} X$,. *We call* $F(x)$ *the limiting distribution of* X_n.

There is a crucial difference between convergence in distribution and the other notions of convergence: note that $|X_n - X|$ never appears anywhere in the definition. In fact, X_n and X never appear together inside a probability or expectation operator.

Example 12 *Let* $X \sim N(0,1)$, *and let* $X_n = X$. *Clearly* $X_n \xrightarrow{p} X$ *and* $X_n \xrightarrow{d} X$. *In fact,* X_n *converges in distribution to any* $N(0,1)$ *random variable. In particular* $X_n \xrightarrow{d} -X$. *However it is definitely not the case that* $X_n \xrightarrow{p} -X$

Relationship between forms of convergence

The different notions of convergence form a partial hierarchy.

- Convergence in mean square implies convergence in probability.
  - We will prove this in Theorem 15 below.
  - This is a useful result, because it is frequently easier to show convergence in mean square than convergence in probability.
- Convergence in probability does not imply convergence in mean square
  - This is proved by Example 6 above.
  - Convergence in mean square usually requires the existence of certain means and variances. There are a lot of useful estimators that do not have finite means, so our primary notion of convergence is convergence in probaility.
- Almost sure convergence implies convergence in probability.
- Convergence in probability does not imply almost sure convergence.

  - This is proved by Example 10 above.

- Convergence in mean square does not imply almost sure convergence.
  - This can also be proved by Example 10 above.

- Almost sure convergence does not imply convergence in mean square.
  - This can also be proved by adapting Example 10 above.

- Convergence in probability implies convergence in distribution.

- Convergence in distribution does not imply convergence in probability.

- Convergence in distribution **to a constant** implies convergence in probability to that constant.

To prove Theorem 15, we need a few fundamental results.

Lemma 13 (Markov's Inequality) *Suppose X_n is a random variable and $Y_n = g(X_n) \geq 0$. Then for any $\varepsilon > 0$ and $p > 0$ we have*

$$\Pr[Y_n > \varepsilon] \leq \varepsilon^{-p} E[Y_n^p].$$

Proof.

$$\begin{aligned} E[Y_n^p] &= \Pr[Y_n^p \leq \varepsilon^p] E[Y_n^p | Y_n^p \leq \varepsilon^p] + \Pr[Y_n^p > \varepsilon^p] E[Y_n^p | Y_n^p > \varepsilon^p] \\ &= \Pr[Y_n \leq \varepsilon] E[Y_n^p | Y_n^p \leq \varepsilon^p] + \Pr[Y_n > \varepsilon] E[Y_n^p | Y_n^p > \varepsilon^p] \end{aligned}$$

since $Y_n \geq 0$, $\varepsilon > 0$ and $p > 0$. Furthermore, since $Y_n \geq 0$, both terms in the sum must be non-negative. Thus,

$$E[Y_n^p] \geq \Pr[Y_n > \varepsilon] E[Y_n^p | Y_n^p > \varepsilon^p].$$

Furthermore, $E[Y_n^p | Y_n^p > \varepsilon^p] > \varepsilon^p$, and hence

$$E[Y_n^p] \geq \Pr[Y_n > \varepsilon] \varepsilon^p$$

which is the result (the central inequality is weak to account for the possibility that $\Pr[Y_n > \varepsilon] = 0$). ■

This is a useful result, and it generalizes an inequality you've probably seen before.

Lemma 14 (Chebychev's Inequality) *If X_n is a random variable and c, $\varepsilon > 0$ are constants, then*

$$\Pr[|X_n - c| > \varepsilon] \leq E\left[(X_n - c)^2\right] / \varepsilon^2.$$

Proof. This is just a special case of Markov's inequality where $Y_n = g(X_n) = |X_n - c|$ and $p = 2$. ■

Now we're ready to prove that convergence in mean square is a special case of convergence in probability.

Theorem 15 *If $X_n \overset{m.s.}{\longrightarrow} c$, then plim $X_n = c$.*

Proof. Since $X_n \overset{m.s.}{\longrightarrow} c$, we know $\lim_{n\to\infty} E\left[(X_n - c)^2\right] = 0$. From Chebychev's Inequality, we know that for any $\varepsilon > 0$

$$\Pr\left[|X_n - c| > \varepsilon\right] \leq E\left[(X_n - c)^2\right]/\varepsilon^2.$$

Taking limits on both sides gives

$$\begin{aligned} \lim_{n\to\infty} \Pr\left[|X_n - c| > \varepsilon\right] &\leq \lim_{n\to\infty} E\left[(X_n - c)^2\right]/\varepsilon^2 \\ &= \varepsilon^{-2} \lim_{n\to\infty} E\left[(X_n - c)^2\right] \\ &= 0. \end{aligned}$$

Furthermore, we know that $\Pr\left[|X_n - c| > \varepsilon\right] \geq 0$ (it's a probability, after all) and hence $\lim_{n\to\infty} \Pr\left[|X_n - c| > \varepsilon\right] \geq 0$ also, and we're done. ■

Convergence of random vectors and matrices

Everything so far has been about convergence of random variables. But we will often be interested in the asymptotic properties of a sequence of random vectors. We will also be interested in random matrices, but since we can always "vectorize" a real matrix the extension is trivial. Advanced econometrics, especially nonparametric and time series econometrics, will often deal with sequences of random functions. That is beyond the scope of this course.

For convergence in probability and almost sure convergence, the existing definitions can easily be adapted to random vectors. Just change the word "variables" to "k-vectors", and replace $|X_n - X|$ with its generalization $||X_n - X||$ where $||b|| = (b'b)^{1/k}$. Alternatively we can simply say that the sequence of random vectors X_n converges to the random vector X if and only if each of the elements of X_n converges to the corresponding element of X.

For convergence in distribution the picture is more complex. As we discussed previously, when X_n converges in distribution to some random variable X, it also converges in distribution to any other random variable Y that has the same CDF as X.

Example 16 *Let $X \sim N(0,1)$*

$$X_n = \begin{bmatrix} X_{n1} \\ X_{n2} \end{bmatrix} = \begin{bmatrix} X \\ (-1)^n X \end{bmatrix}$$

Note that $X_{n1} \xrightarrow{d} X$ and $X_{n2} \xrightarrow{d} X$. But the (joint) distribution of X_n is $N(\mathbf{0}, \Sigma_n)$ where

$$\Sigma_n = \begin{cases} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} & \text{if } n \text{ is even} \\ \\ \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} & \text{if } n \text{ is odd} \end{cases}$$

which clearly doesn't converge in distribution to anything.

A result known as the "Cramer-Wold device" will allow us to make statements about convergence in distribution for vectors.

Theorem 17 (Cramer-Wold device) *Let $X_n, n = 1, 2, \ldots$ be a sequence of random k-vectors. Then $X_n \xrightarrow{d} X$ if and only if for every fixed k-vector c we have $c'X_n \xrightarrow{d} c'X$*

I should point out here that this theorem is stronger than the one stated in the textbook. The textbook version includes the "only if" but not the "if". But both are true, and the "if" is needed in order to actually use the thing, so I assume that's an oversight on their part.

The Law of Large Numbers

The Weak Law of Large Numbers

The law of large numbers (LLN) is at least somewhat familiar to you. It says that as long as certain conditions are satisfied the sample mean converges (in some sense) to its own expected value. You may also be familiar with the fact that there are many laws of large numbers, each with slightly different sets of conditions that must be satisfied. We will start with a simple law of large numbers that we can prove easily:

Theorem 18 (The weak law of large numbers) *Let $X_1, X_2, ..., X_n$ be an iid random sample of size n from a population with finite mean μ and variance σ^2, and let $\bar{x}_n$ be the associated sample mean. Then plim $\bar{x}_n = \mu$.*

Proof. $E[\bar{x}_n] = \mu$ and $Var[\bar{x}_n] = \sigma^2/n$, so $\bar{x}_n \xrightarrow{m.s.} \mu$ and hence plim $\bar{x}_n = \mu$. ■

This is one form of the weak law of large numbers (WLLN). It is called "weak" because it implies weak convergence (i.e., convergence in probability).

This law of large numbers applies to many cases of interest without modification. But notice that it imposes several strong conditions:

1. Independent.
2. Identically distributed.
3. Finite variances.

The intuition here is that in order for the sample mean to give you a good estimate of the expected value, each observation has to provide useful information (finite variances), at least some of that information has to be new information given previous information (independent), and the thing that we're getting information about can't be changing faster then we're learning about it (identically distributed).

General Forms of the WLLN

These three assumptions can be relaxed quite dramatically. But there is something of a tradeoff between how much you relax one and how much you can relax others.

Khinchine's WLLN does not require finite variance.

Theorem 19 (Khinchine's WLLN) *If $X_1, X_2, ..., X_n$ is an iid random sample from a population with finite mean μ, then*

$$plim\ \bar{x}_n = \mu.$$

Chebychev's WLLN doesn't require identically distributed observations. Note that unlike Khinchine's WLLN, Chebychev's requires bounded (not just finite) variance.

Theorem 20 (Chebychev's WLLN) *If $X_1, X_2, ..., X_n$ is an independent sample of observations from a population such that $E[X_i] = \mu_i < \infty$ and $Var[X_i] = \sigma_i^2 < c < \infty$ (that is, the variances are bounded from above), then*

$$plim\ (\bar{x}_n - \bar{\mu}_n) = 0$$

where $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \mu_i$.

Chebychev's WLLN doesn't state that $\bar{x}_n$ converges to $\bar{\mu}_n$, or even that it converges to a constant at all. In fact, it makes no statement about the behaviour of $\bar{\mu}_n$ as n increases. All it says is that as the sample size increases without bound, the probability that $\bar{x}_n$ and $\bar{\mu}_n$ are "far apart" converges to 0.

Markov's WLLN doesn't require independence or identical distribution.

Theorem 21 (Markov's WLLN) *If $X_1, X_2, ..., X_n$ is a sample of observations from a population with $E[X_i] = \mu_i < \infty$ and $\lim_{n\to\infty} Var[\bar{x}_n] = 0$, then*

$$plim\ (\bar{x}_n - \bar{\mu}_n) = 0$$

where $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \mu_i$.

There are many other versions of the WLLN, particularly in dealing with time series. We will leave those for the next course, as it takes some serious intellectual investment to understand the conditions.

The Strong Law of Large Numbers

We can obtain a stronger statement of the "the" law of large numbers that involves almost sure convergence. Of course, obtaining a stronger form of convergence requires stronger assumptions. Here are two versions of the Strong Law of Large Numbers (SLLN), proofs are omitted.

Theorem 22 (A SLLN) *If $X_1, X_2, ..., X_n$ is a sequence of iid random variables with $E[X_i] = \mu < \infty$, then $(\bar{x}_n - \mu) \overset{a.s.}{\to} 0$.*

Theorem 23 (Kolmogorov's SLLN) *If $X_1, X_2, ..., X_n$ is a sequence of independently distributed random variables such that $E[X_i] = \mu_i < \infty$ and $Var[X_i] = \sigma_i^2 < \infty$ satisfying $\lim_{n\to\infty} \sum_{i=1}^n \sigma_i^2 / i^2 < \infty$, then*

$$\bar{x}_n - \bar{\mu}_n \overset{a.s.}{\longrightarrow} 0$$

where $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^n \mu_i$.

The Central Limit Theorem

Ultimately, the most useful aspect of asymptotic theory is that it often provides an approximate sampling distribution for an estimator (or test statistic) in situations where its exact sampling distribution is unknown. The approximation is based on the distributional properties of the estimator as the sample size tends to infinity.

Often, we need to apply a **stabilizing transformation** to the statistic of interest if we're to make any progress. Why? Once again, consider the sample mean $\bar{x}_n$ for a sample size of size n. We've shown that $\bar{x}_n \xrightarrow{p} \mu$ under pretty weak conditions. In this case, the probability distribution of $\bar{x}_n$ is degenerate as $n \to \infty$, because its sampling distribution collapses to μ. But a stabilizing transformation like $\sqrt{n}\,(\bar{x}_n - \mu)$ has a non-degenerate probability distribution as $n \to \infty$.

Based on this stabilizing transformation, we might hope to make probability statements like

$$\lim_{n\to\infty} \Pr\left[\sqrt{n}\,(\bar{x}_n - \mu) < z\right] = F(z)$$

for some distribution F. Then we could use F as an approximate sampling distribution for $\sqrt{n}\,(\bar{x}_n - \mu)$. In this case, we call F the **limiting distribution** of $\sqrt{n}\,(\bar{x}_n - \mu)$. On making a simple change of variables, F implies an approximate sampling distribution for $\bar{x}_n$. We'll call that the **asymptotic distribution** of $\bar{x}_n$. In general, we'll use the term *asymptotic distribution* to refer to a probability distribution that we use to approximate the exact finite sample sampling distribution of a random variable, whenever the approximation is based on $n \to \infty$.

The following theorem is one of the most striking in asymptotic theory. It says that when we have a random sample from **any** population with finite mean and variance, the limiting distribution of the sample mean is normal.

Theorem 24 (Lindeberg-Levy Central Limit Theorem) *If $X_1, X_2, \ldots, X_n$ are an iid random sample from a population with mean $\mu < \infty$ and variance $\sigma^2 < \infty$ and*

$$Z_n = \frac{\sqrt{n}\,(\bar{x}_n - \mu)}{\sigma}$$

then,

$$Z_n \xrightarrow{d} Z$$

where $Z \sim N(0,1)$.

The point of the CLT is quite simple. For large n, the sampling distribution of the sample mean can be approximated by that of a normal with mean μ and variance σ^2/n. We'll write this in one of two equivalent ways:

$$\sqrt{n}\left(\bar{x}-\mu\right) \xrightarrow{d} N\left(0,\sigma^2\right) \tag{1}$$
$$\bar{x} \overset{a}{\sim} N\left(\mu,\sigma^2/n\right). \tag{2}$$

Expression (1) is a statement about the convergence in distribution of $\sqrt{n}\left(\bar{x}-\mu\right)$. The limiting distribution of $\sqrt{n}\left(\bar{x}-\mu\right)$ is $N\left(0,\sigma^2\right)$. Expression (2), on the other hand, is a statement about the approximate sampling distribution of $\bar{x}$ as the sample gets large. The $N\left(\mu,\sigma^2/n\right)$ distribution is what we usually call the asymptotic distribution of $\bar{x}$. It defines an approximate sampling distribution for $\bar{x}$ that we can use for inference in large samples.

There are many generalizations of the Lindeberg-Levy CLT (see the textbook for some). These relax the assumption of identical means and variances (like in Chebychev's WLLN) but require an additional assumption about the sequence of variances.

The Lindeberg-Levy CLT can also be extended to random vectors.

First we need a result known as the "Cramer-Wold device:"

Theorem 25 (Cramer-Wold device) *Let $X_n, n = 1, 2, \ldots$ be a sequence of random k-vectors. Then $X_n \xrightarrow{d} X$ if and only if for every fixed k-vector c we have $c'X_n \xrightarrow{d} c'X$*

I should point out here that this theorem is stronger than the one stated in the textbook. The textbook version includes the "only if" but not the "if". But both are true, and the "if" is needed in order to actually use the thing, so I assume that's an oversight on their part.

Theorem 26 (Multivariate Lindeberg-Levy Central Limit Theorem) *Let $X_1, X_2, \ldots, X_n$ be an IID random sample of random k-vectors from a population with finite mean μ and finite positive definite covariance matrix Σ. Then:*

$$Z_n = \sqrt{n}\left(\bar{x}_n-\mu\right) \xrightarrow{d} Z$$

where $Z \sim N\left(0,\Sigma\right)$.

Proof. To prove this we use the Cramer-Wold device. Pick any k-vector c. Then $c'X_1, c'X_2, \ldots, c'X_n$ is an IID random sample of random variables from a population with finite mean $c'\mu$ and finite variance $c'\Sigma c$. By the univariate CLT we have

$$c'Z_n = \sqrt{n}\left(c'\bar{x}_n - c'\mu\right) \xrightarrow{d} N(0, c'\Sigma c) = c'Z$$

The result then follows directly from the Cramer-Wold device. ■

Consistency

An estimator $\hat{\theta}$ of population quantity θ is **consistent** if the probability that $\hat{\theta}$ is "far" from θ becomes arbitrarily small as the sample size increases. This is just convergence in probability of $\hat{\theta}$ to θ.

Definition 27 (Consistency) *Let $\left\{\hat{\theta}_n, n = 1, 2, \ldots\right\}$ be a sequence of estimators for a parameter θ. We say $\hat{\theta}_n$ is a* ***consistent estimator*** *of θ iff plim $\hat{\theta}_n = \theta$.*

Consistency is something of a minimal requirement for an estimator.

The asymptotic distribution of a function of a random variable

At this point we have acquired many of the basic tools of asymptotic analysis. The Law of Large Numbers tells us that sample averages converge to the corresponding population mean. The Central Limit Theorem tells us that the probability distribution of a (properly rescaled) sample average converges to the normal distribution.

Many estimators and test statistics can be written as either

- Sample averages
- Functions of sample averages
  - Such estimators are usually called method of moments estimators
- The solution of a system of equations involving sample averages
  - Such estimators are usually called generalized method of moments (GMM) estimators.
  - OLS can be interpreted as a GMM estimator.
- The solution to some maximization problem involving sample averages.
  - Such estimators are usually called optimization estimators or M-estimators
  - Maximum likelihood and least squares are examples of M-estimators.

These estimators are part of a wider class of estimators known as “analog estimators.”

> [Goldberger, *A course in econometrics*, page 117] Perhaps the most natural rule for selecting an estimator is the analogy principle. A population parameter is a feature of a population. To estimate it use the corresponding feature of the sample.

Analog estimators, sometimes also called “plug-in” estimators are estimators that follow this principle.

We will now go through three powerful theorems that will allow us to construct asymptotic distributions for these statistics: Slutsky’s theorem, the continuous mapping theorem, and the delta method.

Slutsky's theorem

Recall that we showed that generally speaking $E(g(x)) \neq g(E(x))$. This makes it difficult to find unbiased estimators for many problems of interest. Slutsky's theorem makes finding consistent estimators easy.

Theorem 28 (Slutsky's Theorem) *Let $\{X_n : n = 1, 2, \ldots\}$ be a sequence of random $k-$vectors such that plim $X_n = c$, and let $g : R^k \to R^m$ be a function that is continuous at c. Then plim $g(X_n) = g(c) = g($plim $X_n)$.*

Proof. Since $g(.)$ is continuous at c, then for any $\epsilon > 0$ there exists a $\delta_\epsilon > 0$ such that $||x - c|| < \delta_\epsilon$ implies $||g(x) - g(c)|| < \epsilon$. This implies that:

$$\Pr(||X_n - c|| < \delta_\epsilon) \leq \Pr(||g(X_n) - g(c)|| < \epsilon) \leq 1$$

Since plim $X_n = c$, we have $\lim_{n\to\infty} \Pr(||X_n - c|| < \delta_\epsilon) = 1$. This means the sequence $\Pr(||g(X_n) - g(c)|| < \epsilon)$ lies between two sequences that converge to 1, so it too must converge to 1:

$$\lim_{n\to\infty} \Pr(||g(X_n) - g(c)|| < \epsilon) = 1$$

Since this is true for any $\epsilon > 0$, we have exactly the condition we need for plim $g(X_n) = g(c)$. ■

Slutsky's theorem is incredibly powerful. It implies that any estimator that can be written as a continuous function of some sample averages will usually be consistent. Note that the version of Slutsky's theorem in the textbook is only for scalars, so this version is a generalization.

Example 29 *Let $\{x_i : i = 1, 2, \ldots\}$ be an IID sequence of Bernouilli(p) random variables. Suppose we were interested in estimating $\theta = \ln p$. A natural estimator is $\hat{\theta} = \ln \bar{x}_n$. We showed several weeks ago that $E(\ln \bar{x}_n)$ does not even exist, so clearly $\hat{\theta}_n$ is a biased estimator of θ. But Slutsky's theorem implies that plim $\ln \bar{x}_n = \ln$ plim $\bar{x}_n = \ln p = \theta$, so $\hat{\theta}_n$ is a consistent estimator.*

Corollary 30 (Some properties of plim) *Let X_n and Y_n be random vectors that converge in probability to constants. Then*

1. *Like the expectation operator, the probability limit is a linear operator, i.e.*

$$\text{plim}(aX_n + bY_n + c) = a\text{plim}X_n + b\text{plim}Y_n + c$$

2. *Unlike the expectation operator, the probability limit also "goes inside" products:*

$$\text{plim}(X_nY_n) = [\text{plim}(X_n)][\text{plim}(Y_n)]$$

Note this also implies plim $X_n/Y_n = ($plim $X_n) / ($plim $Y_n)$ if plim $Y_n \neq 0$.

The continuous mapping theorem

When an estimator or test statistic is based on some (potentially complex) function of sample averages, we can usually derive its asymptotic distribution. There are two main approaches, depending on where we take the function.

Most test statistics will be of the form

$$t_n = g(\sqrt{n}(\bar{x}_n - \mu))$$

and so we will find their asymptotic distributions directly, while most estimators will be of the form:

$$\theta_n = g(\bar{x}_n)$$

So we will need to find the asymptotic distribution of

$$\sqrt{n}(g(\bar{x}_n) - g(\mu))$$

A generalization of Slutsky's theorem allows us to solve the first kind of problem. We don't have an agreed-upon name for this particular theorem. Some call it Slutsky's theorem (as it reflects a generalization of Slutsky's original theorem) and others call it the continuous mapping theorem (as it reflects a special case of a broader theorem in probability theory with that name).

Theorem 31 (Continuous mapping theorem) *Suppose that* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{p} c$, *where* c *is a constant. Then for any continuous function* g, $g(X_n, Y_n) \xrightarrow{d} g(X, c)$.

Corollary 32 *Suppose that*

$$Z_n = \sqrt{n}\frac{(\bar{x}_n - \mu)}{\sigma} \xrightarrow{d} Z \sim N(0,1)$$

and suppose that $\hat{\sigma}_n$ *is any consistent estimator of* σ. *Let*

$$t_n = \sqrt{n}\frac{(\bar{x}_n)}{\hat{\sigma}} = Z_n \frac{\sigma}{\hat{\sigma}}$$

be the usual "t-statistic" for testing the null hypothesis that $\mu = 0$. *Then since* $Z_n \xrightarrow{d} Z$ *and* $\frac{\sigma}{\hat{\sigma}} \xrightarrow{p} 1$, *we have:*

$$t_n \xrightarrow{d} Z \sim N(0,1)$$

Corollary 33 *Suppose that* $Z_n \xrightarrow{d} N(0,1)$. *Then* $Z_n^2 \xrightarrow{d} \chi_1^2$

Corollary 34 *Suppose that* $X_n \xrightarrow{d} N(0,\Sigma)$ *(where* X_n *is a* k*-vector and* Σ *is a* k*-by-*k *covariance matrix. Then for any* j*-by-*k *matrix* A,

$$AX_n \xrightarrow{d} N(0, A\Sigma A')$$

The delta method

Theorem 31 turns out to be very handy for developing the asymptotic distribution of test statistics. But it doesn't always help in deriving the asymptotic distribution of estimators. For that, we often use something called the delta method.

Theorem 35 (Delta method) *Let X_n be a random k-vector and μ a constant k-vector such that $\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \Sigma)$, and let $g : R^k \to R^j$ be a differentiable function. Then:*

$$\sqrt{n}(g(X_n) - g(X)) \xrightarrow{d} N(0, G(\mu)\Sigma G(\mu)')$$

where $G(\mu)$ is the j-by-k Jacobian matrix of $g(\mu)$

Just like Slutsky's theorem allows us to prove consistency for estimators based on some continuous function of sample averages, the delta method allows us to find the asymptotic distribution of those estimators (as long as the function is also differentiable).

Example 36 *Let x_i be an IID random sample from a Bernouilli(p) distribution We know that $\sqrt{n}(\bar{x}_n - p) \xrightarrow{d} N(0, p(1-p))$. Now $\frac{d \ln p}{dp} = \frac{1}{p}$, so $\sqrt{n}(\ln \bar{x}_n - \ln p) \xrightarrow{d} N(0, \frac{1-p}{p})$*

Appendix: The Order of a Sequence

Up to now we have been concerned with whether or not a sequence converges. In the case of convergence in probability, we are concerned with whether or not *a sequence of probabilities converges* to zero (or one). In the case of almost sure convergence, we are concerned with the *probability that a random sequence converges.* An important characteristic of a sequence is the *rate* at which it converges or diverges. We call this the order of a sequence. For example, consider the sequences $a_n = \frac{1}{n}$ and $c_n = \frac{1}{n^2}$, $n = 1, 2, 3, \ldots$. Both sequences converge to zero, but c_n does so much more quickly than a_n. Knowing the rate at which a sequence converges is useful when we take asymptotic approximations. Frequently, our approximations involve a remainder term, and we would like to know at what rate that remainder converges to a constant (or zero).

Definition 37 (big-O) *Let $\{a_n, b_n, n = 1, 2, 3, ...\}$ be a double sequence of real numbers. The sequence $\{a_n, n = 1, 2, 3, ...\}$ is said to be* ***of order at most*** *b_n, denoted $a_n = O(b_n)$ if*

$$\lim_{n\to\infty} \left| \frac{a_n}{b_n} \right| < K$$

for some $K \in (0, \infty)$.

Definition 38 (little-o) *Let $\{a_n, b_n, n = 1, 2, 3, ...\}$ be a double sequence of real numbers. The sequence $\{a_n, n = 1, 2, 3, ...\}$ is said to be* ***of order less than*** *b_n, denoted $a_n = o(b_n)$ if*

$$\lim_{n\to\infty} \frac{a_n}{b_n} = 0.$$

There are a couple of important special cases. If $a_n = O(1)$ then a_n is bounded. If $a_n = o(1)$ then a_n converges to zero. Furthermore, if $a_n = O(n^\alpha)$ then $a_n = o\left(n^{\alpha+\delta}\right)$ for $\alpha, \delta > 0$. Some examples should illustrate these concepts.

Example 39

$$\begin{aligned} a_n &= \frac{1}{n} = O\left(n^{-1}\right) = o(1) \\ c_n &= \frac{1}{n^2} = O\left(n^{-2}\right) = o\left(n^{-1}\right) \\ d_n &= \frac{1}{2n^2 - 3} = O\left(n^{-2}\right) = o\left(n^{-1}\right) \\ e_n &= n + 1 = O(n) = o\left(n^2\right) \end{aligned}$$

We can extend these concepts to real valued functions, rather than sequences. For example, if h and g are real valued functions with common domain D, then

$$\begin{aligned} h(x) &= O(g(x)) \text{ if } \lim_{x \to x_0} \left| \frac{h(x)}{g(x)} \right| \leq K \text{ for } K \in (0, \infty) \text{ and } x, x_0 \in D \\ h(x) &= o(g(x)) \text{ if } \lim_{x \to x_0} \frac{h(x)}{g(x)} = 0 \text{ for } x, x_0 \in D. \end{aligned}$$

This formulation of the big-O and little-o notation is particularly useful when we consider Taylor expansions. If $h(x)$ is differentiable of order n (i.e., the derivatives $\partial^j h / \partial x^j$ exist for $j = 1, 2, ..., n$) at $x = x_0$, then

$$h(x_0 + \delta) = h(x_0) + \delta \left. \frac{\partial h}{\partial x} \right|_{x=x_0} + \frac{\delta^2}{2!} \left. \frac{\partial^2 h}{\partial x^2} \right|_{x=x_0} + \cdots + \frac{\delta^n}{n!} \left. \frac{\partial^n h}{\partial x^n} \right|_{x=x_0} + o(\delta^n) \text{ as } \delta \to 0$$

$$\begin{aligned} \text{or } \; h(x) &= h(x_0) + (x - x_0) \left. \frac{\partial h}{\partial x} \right|_{x=x_0} + \frac{(x - x_0)^2}{2!} \left. \frac{\partial^2 h}{\partial x^2} \right|_{x=x_0} + \cdots + \frac{(x - x_0)^n}{n!} \left. \frac{\partial^n h}{\partial x^n} \right|_{x=x_0} \\ &+ \; o\left((x - x_0)^n\right) \text{ as } x \to x_0 \end{aligned}$$

We can extend the big-O and little-o notation to the case of random sequences (stochastic convergence) as follows.

Definition 40 (big-O$_p$) *Let $\{X_n, n = 1, 2, 3, ...\}$ be a sequence of random variables and $\{c_n, n = 1, 2, 3, ...\}$ be a sequence of positive real numbers. We say the sequence X_n is* ***of order at most** c_n **in probability**, denoted $X_n = O_p(c_n)$, if there exists a non-stochastic sequence $\{a_n, n = 1, 2, 3, ...\}$ such that*

$$a_n = O(1) \text{ and } \left(\frac{X_n}{c_n} - a_n \right) \xrightarrow{p} 0.$$

Definition 41 (little-o$_p$) *Let $\{X_n, n = 1, 2, 3, ...\}$ be a sequence of random variables and $\{c_n, n = 1, 2, 3, ...\}$ be a sequence of positive real numbers. We say the sequence X_n is* ***of order less than** c_n **in probability**, denoted $X_n = o_p(c_n)$, if*

$$\left(\frac{X_n}{c_n} \right) \xrightarrow{p} 0.$$

We can define $O_{a.s.}$ and $o_{a.s.}$ similarly.

Appendix: Some more difficult proofs

This section provides some proofs that are important, but somewhat difficult. It is more important for you to understand the theorem itself than to understand its proof, so I have put these proofs in an appendix. Just in case you're curious.

To prove the various weak laws of large numbers, we need one more lemma and a property of characteristic functions.

Lemma 42 *Let $b(x)$ be a function such that $\lim_{x\to 0} b(x)/x = 0$. Then*

$$\lim_{x\to 0}[1+ax+b(x)]^{1/x} = e^a.$$

Proof. Let $y = ax + b(x) = x[a + b(x)/x]$. Then

$$[1+ax+b(x)]^{1/x} = \left[(1+y)^{1/y}\right]^{y/x} = \left[(1+y)^{1/y}\right]^{a+b(x)/x} = (1+y)^{a/y}\left[(1+y)^{1/y}\right]^{b(x)/x}.$$

As $x \to 0$, then $y \to 0$ also and $(1+y)^{a/y} \to e^a$ (this is one of many ways to define the exponential function). Also as $x \to 0$, then $b(x)/x \to 0$ by hypothesis, and hence $\left[(1+y)^{1/y}\right]^{b(x)/x} \to e^0 = 1$. ■

Proposition 43 *Suppose $X_1, X_2, ..., X_n$ are independent random variables with characteristic functions $\phi_{X_j}(t) = E\left[e^{itX_j}\right]$. Then the characteristic function of $\sum_{j=1}^n X_j$ is*

$$\phi^*(t) = E\left[e^{it\sum_j X_j}\right] = \prod_j E\left[e^{itX_j}\right] = \prod_j \phi_{X_j}(t).$$

Proof of Khinchine's WLLN. Let $\phi_{X_j}(t) = E\left[e^{itX_j}\right]$ be the characteristic function for X_j, $j = 1, 2, ..., n$. Taking a Taylor series expansion of $\phi_{X_j}(t)$ around $t_0 = 0$ gives

$$\begin{aligned}\phi_{X_j}(t) &= E\left[1 + itX_j + \frac{(itX_j)^2}{2!} + \cdots\right] \\ &= E[1 + itX_j + o(t)] \\ &= 1 + it\mu + o(t).\end{aligned}$$

Since $\bar{x}$ is a sum of iid random variables X_j/n, we can write the characteristic function of $\bar{x}$ as

$$\phi^*(t) = E\left[e^{it\sum_{j=1}^n X_j/n}\right] = \prod_{j=1}^n E\left[e^{itX_j/n}\right] = \left[\phi_{X_j}\left(\frac{t}{n}\right)\right]^n.$$

Taking limits gives

$$\begin{aligned}\lim_{n\to\infty}\phi^*(t) &= \lim_{n\to\infty}\left[1 + i\frac{t}{n}\mu + o\left(\frac{t}{n}\right)\right]^n \\ &= e^{it\mu}\end{aligned}$$

by lemma 42. Note that $e^{it\mu}$ is the characteristic function of a random variable with density Δ_μ (i.e., that assigns point mass at μ). ■ Notice that we didn't directly prove that plim $\bar{x}_n = \mu$. Rather, we proved that the characteristic function of $\bar{x}$ converges to that of a random variable with a degenerate distribution that puts all mass at μ. Of course plim $\bar{x}_n = \mu$ follows immediately. The proof relied on a first-order Taylor expansion of the characteristic function of $\bar{x}_n$ around $t_0 = 0$, and then we took the limit as $n \to \infty$. We can get better approximations to the distribution of $\bar{x}$ as $n \to \infty$ if we take a higher order expansion (as we'll see in a moment).

Proof of Chebychev's WLLN. Consider the variance of the sample mean,

$$Var\left(\bar{x}_n\right) = E\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu_i\right)\right)^2\right] = \frac{1}{n^2}E\left[\sum_{i=1}^{n}\left(X_i - \mu_i\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 \leq \frac{c}{n}$$

where the second equality follows from independence of the X_i. By Chebychev's Inequality (lemma 14),

$$\Pr\left[\left|\bar{X}_n - \bar{\mu}_n\right| > \varepsilon\right] \leq \frac{Var\left(\bar{x}_n\right)}{\varepsilon^2} \leq \frac{c}{n\varepsilon^2}$$

and therefore

$$\lim_{n\to\infty}\Pr\left[\left|\bar{X}_n - \bar{\mu}_n\right| > \varepsilon\right] \leq \lim_{n\to\infty}\frac{c}{n\varepsilon^2} = 0.$$

■

Proof of Markov's WLLN. By Chebychev's inequality (lemma 14),

$$\Pr\left[\left|\bar{X}_n - \bar{\mu}_n\right| > \varepsilon\right] \leq \frac{Var\left(\bar{x}_n\right)}{\varepsilon^2}$$

and hence

$$\lim_{n\to\infty}\Pr\left[\left|\bar{X}_n - \bar{\mu}_n\right| > \varepsilon\right] \leq \lim_{n\to\infty}\frac{Var\left(\bar{x}_n\right)}{\varepsilon^2} = 0.$$

■

Proof of Lindeberg-Levy Central Limit Theorem. Let $\phi_{X_j-\mu}(t)$ be the characteristic function of $X_j - \mu$. That is, $\phi_{X_j-\mu}(t) = E\left[e^{it(X_j-\mu)}\right]$. Taking a second order Taylor expansion around $t_0 = 0$,

$$\begin{aligned}\phi_{X_j-\mu}(t) &= E\left[e^{it(X_j-\mu)}\Big|_{t=0} + it\left(X_j-\mu\right)\left(e^{it(X_j-\mu)}\Big|_{t=0}\right) + \frac{i^2t^2}{2!}\left(X_j-\mu\right)^2\left(e^{it(X_j-\mu)}\Big|_{t=0}\right) + o\left(t^2\right)\right] \\ &= 1 - \frac{t^2\sigma^2}{2} + o\left(t^2\right).\end{aligned}$$

We know that

$$Z_n = \frac{\sqrt{n}\left(\bar{x}_n - \mu\right)}{\sigma} = \frac{\sum_j\left(X_j - \mu\right)}{\sigma\sqrt{n}}.$$

Furthermore, the characteristic function of

$$Y_i = \frac{X_j - \mu}{\sigma\sqrt{n}}$$

is

$$\phi_{Y_j} = E\left[e^{itY_j}\right] = E\left[e^{it\frac{X_j-\mu}{\sigma\sqrt{n}}}\right] = \phi_{X_j-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right).$$

Therefore, the characteristic function of Z_n is

$$\begin{aligned}\phi_{Z_n}(t) &= \prod_j \phi_{Y_j} = \prod_j \phi_{X_j-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right) \\ &= \prod_j \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right] \\ &= \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n\end{aligned}$$

and therefore

$$\lim_{n\to\infty} \phi_{Z_n}(t) = \lim_{n\to\infty}\left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n = e^{-t^2/2}$$

by lemma 42. This is the characteristic function of a standard normal random variable. Therefore $Z_n \overset{d}{\to} Z \sim N(0,1)$. ■