

Chapter 6

Asymptotics

In this chapter we cover some fundamental results from probability theory concerning asymptotics of random sequences.

6.1 LLN and CLT

The law of large numbers and central limit theorem are two of the pillars of econometrics and statistics. Below we review both theorems, starting with the necessary concepts of convergence in probability and distribution.

6.1.1 Convergence of Random Vectors

A sequence of random vectors $\{\mathbf{x}_n\}$ is said to **converge in probability** to a random vector $\mathbf{x}$ if,

$$\text{for all } \delta > 0, \quad \mathbb{P}\{\|\mathbf{x}_n - \mathbf{x}\| > \delta\} \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad (6.1)$$

In symbols, we write $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$. In the scalar case $\|\mathbf{x}_n - \mathbf{x}\|$ reduces to $|x_n - x|$.

Example 6.1.1 If $\mathcal{L}(\mathbf{x}_n) = N(\mathbf{0}, \sigma_n \mathbf{I})$ and $\sigma_n \rightarrow 0$, then $\mathbf{x}_n \xrightarrow{p} \mathbf{0}$ as $n \rightarrow \infty$. Figure 6.1 illustrates this visually for the scalar case. The variance is $\sigma_n = 1/n$. With fixed $\delta > 0$, the probability $\mathbb{P}\{|x_n| > \delta\}$ is shown for different values of n . This probability collapses to zero as $n \rightarrow \infty$. If we now fix δ at a smaller positive value, $\mathbb{P}\{|x_n| > \delta\}$ can again be made arbitrarily small by increasing n . In other words, (6.1) holds.

Fact 6.1.1 The following statements are true:

$$(i) \quad \mathbf{x}_n \xrightarrow{p} \mathbf{x} \iff \|\mathbf{x}_n - \mathbf{x}\| \xrightarrow{p} 0.$$

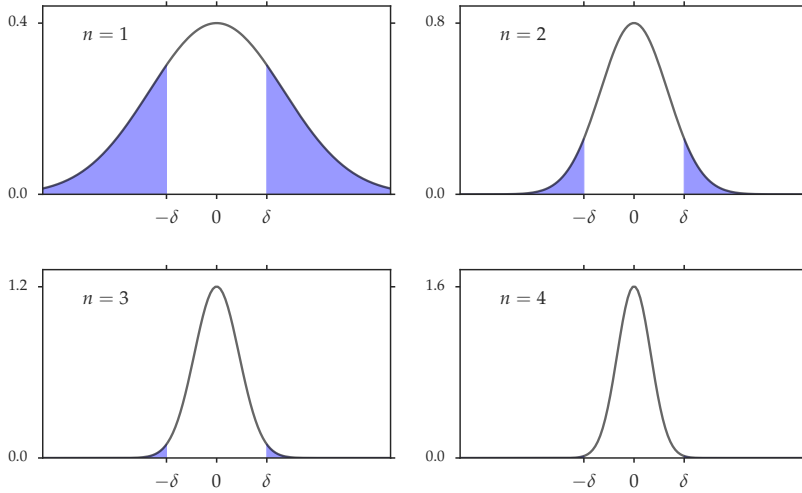


Figure 6.1 $\mathbb{P}\{|x_n| > \delta\} \rightarrow 0$ when $\mathcal{L}(x_n) = \mathcal{N}(0, 1/n)$

- (ii) $\mathbf{x}_n \xrightarrow{p} \mathbf{x} \implies g(\mathbf{x}_n) \xrightarrow{p} g(\mathbf{x})$ whenever g is continuous at $\mathbf{x}$.
- (iii) $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$ and $\mathbf{y}_n \xrightarrow{p} \mathbf{y} \implies \mathbf{x}_n + \mathbf{y}_n \xrightarrow{p} \mathbf{x} + \mathbf{y}$ and $\mathbf{x}_n^\top \mathbf{y}_n \xrightarrow{p} \mathbf{x}^\top \mathbf{y}$.
- (iv) $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$ and $\mathbf{a}_n \rightarrow \mathbf{a} \implies \mathbf{x}_n + \mathbf{a}_n \xrightarrow{p} \mathbf{x} + \mathbf{a}$ and $\mathbf{x}_n^\top \mathbf{a}_n \xrightarrow{p} \mathbf{x}^\top \mathbf{a}$.
- (v) $\mathbf{x}_n \xrightarrow{p} \mathbf{x} \iff \mathbf{a}^\top \mathbf{x}_n \xrightarrow{p} \mathbf{a}^\top \mathbf{x}$ for any $\mathbf{a} \in \mathbb{R}^K$.

Part (i) is an immediate consequence of the definition. Other proofs are touched on in the exercises. For more details, see, for example, Çinlar (2011).

The scalar sequence $\{x_n\}$ is said to converge to x **in mean square** if

$$\mathbb{E}(x_n - x)^2 \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad (6.2)$$

and we write $x_n \xrightarrow{ms} x$. Unlike convergence in probability, for convergence in mean square to be defined we require that our variables have finite second moments.

Fact 6.1.2 Let $\{x_n\}$ and x have finite second moments and let α be any constant. The following statements are true:

- (i) $x_n \xrightarrow{ms} x \implies x_n \xrightarrow{p} x$.
- (ii) $x_n \xrightarrow{ms} \alpha \iff \mathbb{E}x_n \rightarrow \alpha$ and $\text{var}[x_n] \rightarrow 0$.

Part (i) follows from Chebyshev's inequality (page 97). In particular, from monotonicity of $\mathbb{P}$ and (4.19) we get

$$\mathbb{P}\{|x_n - x| > \delta\} \leq \mathbb{P}\{|x_n - x| \geq \delta\} \leq \frac{\mathbb{E}(x_n - x)^2}{\delta^2}$$

Part (ii) is implied by the next fact.

Fact 6.1.3 For any $x \in L_2$ and any constant α we have

$$\mathbb{E}[(x - \alpha)^2] = \text{var}[x] + (\mathbb{E}[x] - \alpha)^2 \quad (6.3)$$

As we'll see later, (6.3) plays an important role in statistics. Its verification is an exercise. (The proof is straightforward. If you want more of a challenge, try to obtain it as a consequence of (iv) of fact 5.2.4 on page 146.)

Example 6.1.2 In figure 6.1 we showed visually how $\mathcal{L}(x_n) = \mathcal{N}(0, 1/n)$ implies $x_n \xrightarrow{p} 0$. This also follows from (i)–(ii) of fact 6.1.2, since $\mathbb{E}x_n = 0$ and $\text{var}[x_n] = 1/n \rightarrow 0$.

6.1.2 The Law of Large Numbers

The law of large numbers (LLN) is one of the most fundamental theorems of both probability and statistics. As a prelude to the LLN, let's consider the effects of averaging over independent random quantities. We'll take the perspective of investors, who often diversify their portfolios in order to decrease risk. To formalize this idea let

- x_n be the payoff from holding one dollar of asset n ,
- $\mathbb{E}x_n = \mu$ and $\text{var}[x_n] = \sigma^2$ for all n , and
- $\text{cov}[x_j, x_k] = 0$ when $j \neq k$.

If we hold just asset 1, then the payoff is x_1 , the expected payoff is μ and the variance is σ^2 . If we diversify by spreading one dollar evenly over N of these assets, our payoff is

$$\bar{x}_N := \frac{1}{N} \sum_{n=1}^N x_n$$

The expected payoff is unchanged at

$$\mathbb{E}\bar{x}_N = \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^N x_n\right] = \frac{1}{N} \sum_{n=1}^N \mathbb{E}x_n = \mu \quad (6.4)$$

But the variance declines at rate $\frac{1}{N}$ because

$$\begin{aligned}\mathbb{E}[(\bar{x}_N - \mu)^2] &= \mathbb{E}\left\{\left[\frac{1}{N}\sum_{i=1}^N(x_i - \mu)\right]^2\right\} \\ &= \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N\mathbb{E}(x_i - \mu)(x_j - \mu) = \frac{1}{N^2}\sum_{i=1}^N\mathbb{E}(x_i - \mu)^2 = \frac{\sigma^2}{N}\end{aligned}$$

The important equality here is the third one, which holds because of the zero covariance between assets. In this step N^2 terms become N terms. To summarize,

$$\mathbb{E}\bar{x}_N = \mu \quad \text{and} \quad \text{var}[\bar{x}_N] = \frac{\sigma^2}{N} \quad \text{for all } N \quad (6.5)$$

Now imagine someone is kind enough to sell us an unlimited quantity of such assets, so we can diversify without limit. By taking $N \rightarrow \infty$ and combining (6.5) with fact 6.1.2 on page 162 we obtain a proof of the **law of large numbers**:

Theorem 6.1.1 *Let $\{x_n\}$ be IID copies of x . If x is integrable, then*

$$\frac{1}{N}\sum_{n=1}^N x_n \xrightarrow{p} \mathbb{E}x \quad \text{as } N \rightarrow \infty \quad (6.6)$$

Actually we haven't fully proved the LLN because we assumed existence of second moments above, while theorem 6.1.1 assumes only existence of the first moment. However, the proof above captures the spirit of the LLN most cleanly. (See theorem 8.3.5 of Dudley 2002 for the general case.)

At first glance, the law of large numbers (6.6) appears to only be a statement about the sample mean but in fact it's self-improving: We can go from (6.6) to arbitrary functions of random variables, or even random vectors. In particular, if $\mathbf{x}$ is any random vector, $\{\mathbf{x}_n\}$ are IID copies and $h: \mathbb{R}^N \rightarrow \mathbb{R}$ is any $\mathcal{B}$ -measurable function such that $h(\mathbf{x})$ is integrable, then

$$\frac{1}{N}\sum_{n=1}^N h(\mathbf{x}_n) \xrightarrow{p} \mathbb{E}h(\mathbf{x}) \quad \text{as } N \rightarrow \infty \quad (6.7)$$

This can be confirmed by letting $y_n := h(\mathbf{x}_n)$ and then applying theorem 6.1.1. The independence of $\{y_n\}$ is confirmed by fact 5.1.10 on page 137. That each y_n has the same distribution as $y := h(\mathbf{x})$ follows from

$$\mathbb{P}\{y_n \in B\} = \mathbb{P}\{h(\mathbf{x}_n) \in B\} = \mathbb{P}\{\mathbf{x}_n \in h^{-1}(B)\} = \mathbb{P}\{\mathbf{x} \in h^{-1}(B)\}$$

The law of large numbers applies to probabilities as well as expectations. To see this, fix $B \subset \mathcal{B}(\mathbb{R}^N)$, and consider the probability $\mathbb{P}\{\mathbf{x} \in B\}$. Letting h be the indicator function $h(\mathbf{s}) = \mathbb{1}_B(\mathbf{s}) = \mathbb{1}\{\mathbf{s} \in B\}$ and using the principle that expectations of indicators equal probabilities of events (page 95), we have

$$\mathbb{E}h(\mathbf{x}) = \mathbb{E}\mathbb{1}\{\mathbf{x} \in B\} = \mathbb{P}\{\mathbf{x} \in B\}$$

Combining this equality with (6.7), we see that if $\{\mathbf{x}_n\}$ is IID with distribution P , then

$$\frac{1}{N} \sum_{n=1}^N \mathbb{1}\{\mathbf{x}_n \in B\} \xrightarrow{p} P(B) \quad (6.8)$$

In words, the fraction of the sample that falls in B converges to the probability that the distribution assigns to B .

6.1.2.1 An Illustration

To illustrate the law of large numbers, consider flipping a coin until 10 heads have occurred. The coin is not fair: The probability of heads is 0.4. Let x be the number of tails observed in the process. This random variable is known to have the **negative binomial distribution**, with $\mathbb{E}x = 15$. The LLN predicts that if we simulate a large number of observations of x and take the average, we get a value close to 15. Running the Julia code in listing 1 gives outcomes consistent with this claim.¹

Figure 6.2 shows what can happen when the finite first moment condition in the LLN is not enforced. The draws are from the standard Cauchy distribution, which fails to have a finite first moment. The top panel shows observations and the bottom shows a running sample mean of observations $\bar{x}_N$, with N on the horizontal axis. A few extreme observations dominate the average and the sample mean does not converge.²

6.1.3 Convergence in Distribution

As a prelude to the central limit theorem we need to review convergence in distribution. This is conceptually quite different to convergence of random variables, whether in probability or mean square. For example, the distributions $N(0, 1)$ and $N(10^{-10}, 1)$ are very close, but independent draws from these respective distributions need not be.

1. See johnstachurski.net/emet.html for code. A routine to draw from the negative binomial distribution is available from the *Distributions* package, but we've rolled our own for extra transparency.

2. In fact it can be shown that the sample mean is itself standard Cauchy for all N .

Listing 1 Illustrates the LLN (Julia)

```

num_reps = 10^6
outcomes = Array{Float64, num_reps}

for i in 1:num_reps
    num_tails = num_heads = 0
    while num_heads < 10
        b = rand()
        num_heads = num_heads + (b < 0.4) # +1 with prob 0.4
        num_tails = num_tails + (b >= 0.4) # +1 with prob 0.6
    end
    outcomes[i] = num_tails
end

println(mean(outcomes))

```

Let $\{P_n\}$ and P be distributions on $\mathbb{R}^K$ in the sense of §5.1.2. To represent the notion that P_n converges to P we could require that $P_n(B) \rightarrow P(B)$ for all $B \in \mathcal{B}(\mathbb{R}^K)$. However, this turns out to be rather strict and also indifferent to closeness in the sense of Euclidean norm. For example, suppose that P_n is a distribution on $\mathbb{R}$ that puts all mass on $\frac{1}{n}$, while P puts all mass on 0. Intuition suggests that $P_n \rightarrow P$, but if we take $B = (0, \infty)$, we have $P_n(B) = 1$ for all n , while $P(B) = 0$.

For this reason the common notion of convergence of distributions, which is called weak convergence, requires only that $P_n(B) \rightarrow P(B)$ for all “continuity sets” in $\mathbb{R}^K$. There’s an equivalent way to phrase this that’s easier to work with in applications: We say that $\{P_n\}$ **converges weakly** to P if

$$\int h(\mathbf{s}) P_n(d\mathbf{s}) \rightarrow \int h(\mathbf{s}) P(d\mathbf{s}) \quad \text{for all continuous bounded } h: \mathbb{R}^K \rightarrow \mathbb{R}$$

and write $P_n \xrightarrow{w} P$. For the definition of these integrals see §5.1.3.3. Continuity of h enforces respect for the topology on $\mathbb{R}^K$, while boundedness is imposed only to ensure that the integrals are finite.

Fact 6.1.4 Let F_n be the CDF of P_n and let F be the CDF of P . In the univariate case ($K = 1$) we have

$$P_n \xrightarrow{w} P \iff F_n(s) \rightarrow F(s) \text{ for all } s \text{ at which } F \text{ is continuous}$$

Example 6.1.3 It can be shown that the t -distribution with k degrees of freedom con-

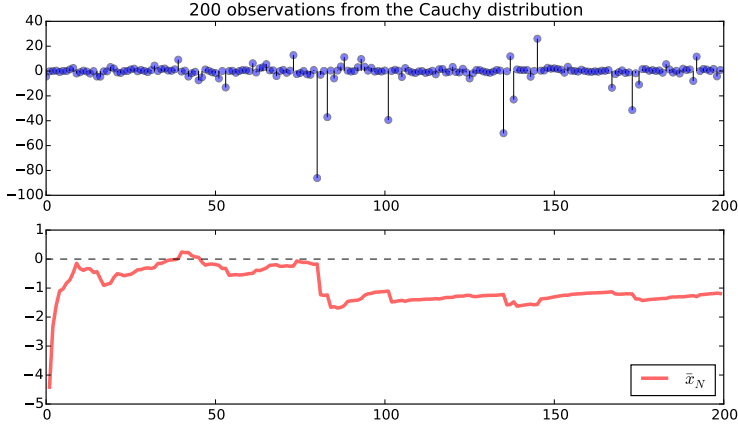


Figure 6.2 Samples from the Cauchy distribution and sample mean

verges weakly to the standard normal distribution as $k \rightarrow \infty$; see figure 6.3.

Sometimes densities are easier to work with than CDFs. In this connection, the following sufficient condition can be useful.

Fact 6.1.5 Let $\{P_n\}$ and P be absolutely continuous probability measures on $\mathbb{R}^K$, with densities p_n and p . If $p_n(\mathbf{s}) \rightarrow p(\mathbf{s})$ for all $\mathbf{s} \in \mathbb{R}^K$, then $P_n \xrightarrow{w} P$.

Let $\{\mathbf{x}_n\}$ and $\mathbf{x}$ be random vectors. We say that $\mathbf{x}_n \rightarrow \mathbf{x}$ **in distribution** if their respective distributions converge weakly. This convergence is symbolized by $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$. Thus

$$\mathbf{x}_n \xrightarrow{d} \mathbf{x} \iff \mathcal{L}(\mathbf{x}_n) \xrightarrow{w} \mathcal{L}(\mathbf{x})$$

In view of fact 5.1.7 on page 134, this is equivalent to

$$\mathbb{E}[h(\mathbf{x}_n)] \rightarrow \mathbb{E}[h(\mathbf{x})] \quad \text{for all continuous bounded } h: \mathbb{R}^K \rightarrow \mathbb{R}$$

Fact 6.1.6 The following statements are true:

- (i) If $g: \mathbb{R}^K \rightarrow \mathbb{R}^J$ is continuous and $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$, then $g(\mathbf{x}_n) \xrightarrow{d} g(\mathbf{x})$.
- (ii) If $\mathbf{a}^\top \mathbf{x}_n \xrightarrow{d} \mathbf{a}^\top \mathbf{x}$ for any $\mathbf{a} \in \mathbb{R}^K$, then $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$.
- (iii) $\mathbf{x}_n \xrightarrow{p} \mathbf{x} \implies \mathbf{x}_n \xrightarrow{d} \mathbf{x}$.
- (iv) If $\mathbf{a}$ is a constant vector and $\mathbf{x}_n \xrightarrow{d} \mathbf{a}$, then $\mathbf{x}_n \xrightarrow{p} \mathbf{a}$.

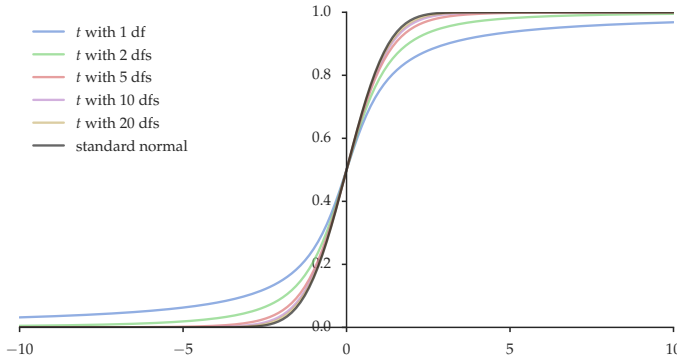


Figure 6.3 t -Distribution with k df converges to $N(0, 1)$ as $k \rightarrow \infty$

Part (i) of fact 6.1.6 is called the **continuous mapping theorem**. Part (ii) is called the Cramér–Wold theorem, or the **Cramér–Wold device**. The next result is sometimes known as **Slutsky’s theorem**.

Fact 6.1.7 If α is constant, $x_n \xrightarrow{p} \alpha$ and $y_n \xrightarrow{d} y$, then $x_n + y_n \xrightarrow{d} \alpha + y$ and $x_n y_n \xrightarrow{d} \alpha y$.

An immediate but useful consequence is that

Fact 6.1.8 $x_n \xrightarrow{p} 0$ and $y_n \xrightarrow{d} y \implies x_n y_n \xrightarrow{p} 0$.

6.1.4 The Central Limit Theorem

The **central limit theorem** is among the most striking and important results in all of mathematics. Relative to the LLN, it requires an additional second moment condition. Here’s a statement of the theorem in the scalar IID case.

Theorem 6.1.2 Let x have finite second moment and let $\{x_n\}$ be IID copies of x . If $\mu := \mathbb{E}x$ and $\sigma^2 := \text{var } x$, then

$$\sqrt{N}(\bar{x}_N - \mu) \xrightarrow{d} N(0, \sigma^2) \quad \text{as } N \rightarrow \infty \quad (6.9)$$

The meaning of $\xrightarrow{d}$ here is that the distribution of the left-hand side $\xrightarrow{w} N(0, \sigma^2)$. We can interpret (6.9) as follows: On one hand, $(\bar{x}_N - \mu) \xrightarrow{p} 0$ by the LLN; on the other, $\sqrt{N} \rightarrow \infty$. If we take the product, these two competing terms just balance.

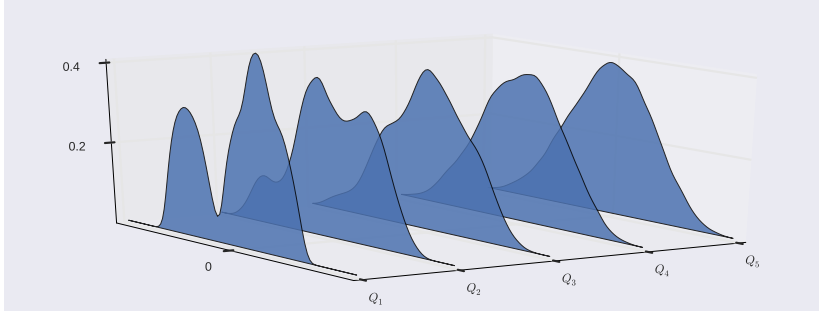


Figure 6.4 CLT in action, starting from a beta mixture

Remarkably, the distribution of the product approaches a zero-mean Gaussian as $N \rightarrow \infty$, regardless of the distribution of x .

We'll skip the proof of the CLT (which provides little intuition) and turn to simulation. For starters, see figure 6.4, where $Q_N :=$ the distribution of the left-hand side of (6.9) for $N = 1, \dots, 5$. The initial distribution $Q = Q_1$ is multi-modal, and was constructed as a convex combination of three beta distributions. The CLT predicts that $Q_N \xrightarrow{w} N(0, \sigma^2)$, where $\sigma^2 = \int s^2 Q_1(ds)$. Note how quickly the bell shape emerges, despite the irregularity of the initial distribution.

Another common statement of the central limit theorem is as follows: If all the conditions of theorem 6.1.2 are satisfied, then

$$z_N := \sqrt{N} \left\{ \frac{\bar{x}_N - \mu}{\sigma} \right\} \xrightarrow{d} N(0, 1) \quad \text{as } N \rightarrow \infty \quad (6.10)$$

Exercise 6.4.6 asks you to confirm this.

The convergence in (6.10) is illustrated by listing 2, the output of which is given in figure 6.5. The listing generates 5,000 observations of the random variable z_N defined in (6.10), where each x_n is $\chi^2(5)$. (The mean of this distribution is 5, and the variance is $2 \times 5 = 10$.) The observations of z_N are stored in the vector `outcomes`. The figure histograms this vector and then superimposes the density of the $N(0, 1)$ distribution.

6.2 Extensions

In this section we cover some extensions to the classical LLN and CLT.

Listing 2 Illustrates the CLT (Python)

```

import numpy as np
import scipy.stats as st

num_reps = 5000
outcomes = np.empty(num_reps)
N, k = 1000, 5      # k = degrees of freedom
chi = st.chi2(k)

for i in range(num_reps):
    xvec = chi.rvs(N)
    outcomes[i] = np.sqrt(N / (2 * k)) * (xvec.mean() - k)

# code to histogram outcomes and plot st.norm.pdf
# see johnstachurski.net/emet.html

```

6.2.1 Convergence of Random Matrices

The notion of convergence in probability discussed in §6.1.1 extends naturally to matrices. Let $\{\mathbf{X}_n\}_{n=1}^{\infty}$ be a sequence of random $N \times K$ matrices. We say that $\mathbf{X}_n$ converges to a random $N \times K$ matrix $\mathbf{X}$ **in probability** and write $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ if

$$\|\mathbf{X}_n - \mathbf{X}\| \xrightarrow{p} 0 \quad \text{as } n \rightarrow \infty$$

where $\|\cdot\|$ is the matrix norm defined in §3.3.5.

Fact 6.2.1 Assuming conformability, the following statements are true:

(i) If $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ and $\mathbf{X}_n$ and $\mathbf{X}$ are nonsingular, then $\mathbf{X}_n^{-1} \xrightarrow{p} \mathbf{X}^{-1}$.

(ii) If $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ and $\mathbf{Y}_n \xrightarrow{p} \mathbf{Y}$, then

$$\mathbf{X}_n + \mathbf{Y}_n \xrightarrow{p} \mathbf{X} + \mathbf{Y}, \quad \mathbf{X}_n \mathbf{Y}_n \xrightarrow{p} \mathbf{X} \mathbf{Y}, \quad \text{and} \quad \mathbf{Y}_n \mathbf{X}_n \xrightarrow{p} \mathbf{Y} \mathbf{X}$$

(iii) If $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ and $\mathbf{A}_n \rightarrow \mathbf{A}$, then

$$\mathbf{X}_n + \mathbf{A}_n \xrightarrow{p} \mathbf{X} + \mathbf{A}, \quad \mathbf{X}_n \mathbf{A}_n \xrightarrow{p} \mathbf{X} \mathbf{A}, \quad \text{and} \quad \mathbf{A}_n \mathbf{X}_n \xrightarrow{p} \mathbf{A} \mathbf{X}$$

(iv) $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ if and only if $\mathbf{X}_n \mathbf{a} \xrightarrow{p} \mathbf{X} \mathbf{a}$ for any conformable vector $\mathbf{a}$.

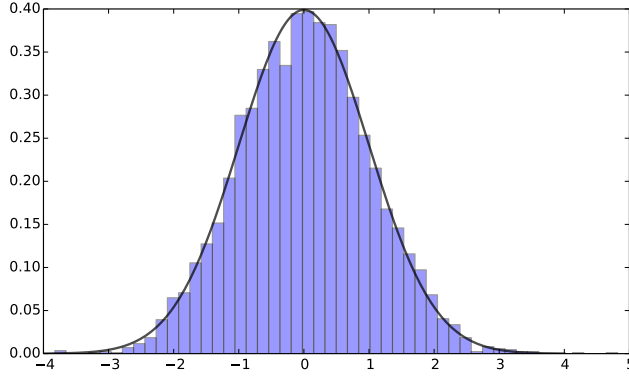


Figure 6.5 Observations of z_N in (6.10) when the underlying distribution is $\chi^2(5)$

(v) $\mathbf{a}^\top \mathbf{X}_n \mathbf{a} \xrightarrow{p} \mathbf{a}^\top \mathbf{X} \mathbf{a}$ whenever $\mathbf{a}$ is a conformable constant vector and $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$.

In part (iii) of fact 6.2.1, the matrices $\mathbf{A}_n$ and $\mathbf{A}$ are nonrandom and convergence is in matrix norm. Note that part (v) can be obtained from two applications of part (iii).

In econometrics we often use the vector version of Slutsky's theorem:

Fact 6.2.2 Let $\mathbf{x}_n$ and $\mathbf{x}$ be random vectors in $\mathbb{R}^K$, let $\mathbf{Y}_n$ be random matrices, and let $\mathbf{C}$ be a constant matrix. Assuming conformability, we have

$$\mathbf{Y}_n \xrightarrow{p} \mathbf{C} \text{ and } \mathbf{x}_n \xrightarrow{d} \mathbf{x} \implies \mathbf{Y}_n \mathbf{x}_n \xrightarrow{d} \mathbf{C} \mathbf{x} \text{ and } \mathbf{Y}_n + \mathbf{x}_n \xrightarrow{d} \mathbf{C} + \mathbf{x}$$

6.2.2 Vector-Valued LLNs and CLTs

The scalar LLN and CLT that we discussed above extend to the vector case in a natural way. For example,

Theorem 6.2.1 Let $\mathbf{x}$ be a random vector in $\mathbb{R}^K$ and let $\{\mathbf{x}_n\}$ be IID copies of $\mathbf{x}$. If $\boldsymbol{\mu} := \mathbb{E} \mathbf{x}$ is finite, then

$$\bar{\mathbf{x}}_N := \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \xrightarrow{p} \boldsymbol{\mu} \text{ as } N \rightarrow \infty \quad (6.11)$$

If, in addition, $\mathbb{E} \|\mathbf{x}\|^2 < \infty$, then

$$\sqrt{N} (\bar{\mathbf{x}}_N - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \Sigma) \text{ where } \Sigma := \text{var } \mathbf{x} \quad (6.12)$$

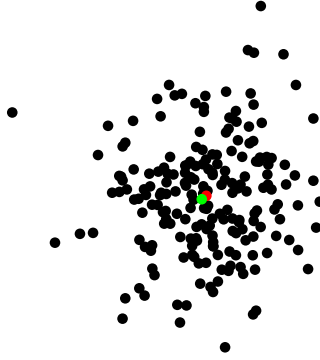


Figure 6.6 LLN, vector case

Here $\frac{1}{N} \sum_{n=1}^N \mathbf{x}_n$ should be understood in terms of vector addition and scalar multiplication. Thus $\bar{\mathbf{x}}_N$ is a convex combination of the sample points in $\mathbb{R}^K$. Figure 6.6 illustrates this in two dimensions. The green dot is $\boldsymbol{\mu}$. The black dots are realizations of $\mathbf{x}_1, \dots, \mathbf{x}_N$. The red dot is the sample mean $\frac{1}{N} \sum_{n=1}^N \mathbf{x}_n$. By (6.11), the red dot converges to the green dot.

The vector LLN in theorem 6.2.1 follows from the scalar LLN. To see this, let $\mathbf{x}_n$ be as in theorem 6.2.1, let $\mathbf{a}$ be any constant vector in $\mathbb{R}^K$, let $y_n := \mathbf{a}^\top \mathbf{x}_n$ and let $y := \mathbf{a}^\top \boldsymbol{\mu}$. The sequence $\{y_n\}$ is IID (see fact 5.1.10 on page 137) with the same distribution as y . By the scalar LLN (theorem 6.1.1), we have

$$\frac{1}{N} \sum_{n=1}^N y_n \xrightarrow{p} \mathbb{E} y = \mathbb{E} [\mathbf{a}^\top \mathbf{x}] = \mathbf{a}^\top \mathbb{E} [\mathbf{x}] = \mathbf{a}^\top \boldsymbol{\mu}$$

At the same time,

$$\frac{1}{N} \sum_{n=1}^N y_n = \frac{1}{N} \sum_{n=1}^N \mathbf{a}^\top \mathbf{x}_n = \mathbf{a}^\top \left[\frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \right] = \mathbf{a}^\top \bar{\mathbf{x}}_N$$

We have shown that

$$\mathbf{a}^\top \bar{\mathbf{x}}_N \xrightarrow{p} \mathbf{a}^\top \boldsymbol{\mu} \text{ for any } \mathbf{a} \in \mathbb{R}^K$$

The claim $\bar{\mathbf{x}}_N \xrightarrow{p} \boldsymbol{\mu}$ now follows from fact 6.1.1.

The vector CLT in theorem 6.2.1 also follows from the scalar case and the proof is quite similar. See exercise 6.4.10.

These results can be further extended to random matrices. In fact we will need a

matrix LLN in our investigation of linear regression. Here's a suitable result:

Fact 6.2.3 Let $\mathbf{X}$ be a random matrix and let $\{\mathbf{X}_n\}$ be IID copies of $\mathbf{X}$. If $\mathbb{E}\|\mathbf{X}\| < \infty$, then

$$\frac{1}{N} \sum_{n=1}^N \mathbf{X}_n \xrightarrow{p} \mathbb{E}\mathbf{X} \quad \text{as } N \rightarrow \infty \quad (6.13)$$

The proof is straightforward: By fact 6.2.1, to prove (6.13), it suffices to show that $\frac{1}{N} \sum_{n=1}^N \mathbf{X}_n \mathbf{a} \xrightarrow{p} \mathbb{E}[\mathbf{X}] \mathbf{a}$ for any conformable vector $\mathbf{a}$. Since $\mathbf{X}_n \mathbf{a}$ is a vector with expectation $\mathbb{E}[\mathbf{X}] \mathbf{a}$, this is immediate from the vector LLN in theorem 6.2.1.

6.2.3 The Delta Method

The asymptotic normality result in the central limit theorem is preserved under linear transformations, as follows from fact 6.2.2. It turns out that the same thing is true even for functions that are only locally almost linear—that is, for differentiable functions. This greatly extends the power and reach of the central limit theorem.

Theorem 6.2.2 Let $g: \mathbb{R}^K \rightarrow \mathbb{R}$, let $\boldsymbol{\theta}$ be a point in the domain of g , and let $\{\mathbf{t}_n\}$ be a sequence of random vectors in $\mathbb{R}^K$. If

- (i) $\sqrt{n}(\mathbf{t}_n - \boldsymbol{\theta}) \xrightarrow{d} N(0, \boldsymbol{\Sigma})$ for some positive definite $\boldsymbol{\Sigma}$ and
- (ii) $\nabla g(\boldsymbol{\theta})$ exists, is continuous, and each element is nonzero,

then

$$\sqrt{n}\{g(\mathbf{t}_n) - g(\boldsymbol{\theta})\} \xrightarrow{d} N(0, \nabla g(\boldsymbol{\theta})^\top \boldsymbol{\Sigma} \nabla g(\boldsymbol{\theta})) \quad \text{as } n \rightarrow \infty \quad (6.14)$$

Here $\nabla g(\boldsymbol{\theta})$ is the **gradient vector** of g at $\boldsymbol{\theta}$, which is defined as

$$\nabla g(\boldsymbol{\theta}) := \begin{pmatrix} g'_1(\boldsymbol{\theta}) \\ \vdots \\ g'_K(\boldsymbol{\theta}) \end{pmatrix} \quad \text{where } g'_k(\boldsymbol{\theta}) := \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_k}$$

In the scalar case, (6.14) translates to

$$\sqrt{n}\{g(t_n) - g(\theta)\} \xrightarrow{d} N(0, g'(\theta)^2 \sigma^2) \quad \text{as } n \rightarrow \infty \quad (6.15)$$

In statistical settings, the technique of obtaining asymptotic distributions by appealing to theorem 6.2.2 is called the **delta method**. We will see its importance later on. The proof of theorem 6.2.2 is based on a Taylor expansion of g around the point $\boldsymbol{\theta}$. See theorem 3.1 of Van der Vaart (2000). Exercise 6.4.9 walks you through some of the key ideas.

6.3 Further Reading

For high quality treatments of the law of large numbers and the central limit theorem see, for example, Dudley (2002), Taylor (1997), Durrett (2010), Pollard (2002), DasGupta (2008), or Çinlar (2011). An advanced but thorough treatment of the delta method can be found in Van der Vaart (2000).

6.4 Exercises

Ex. 6.4.1 Let $\{x_n\}$ be a sequence of random variables satisfying $x_n = y$ for all n , where y is a single random variable. Show that if $\mathbb{P}\{y = -1\} = \mathbb{P}\{y = 1\} = 0.5$, then $x_n \xrightarrow{p} 0$ fails. Show that if $\mathbb{P}\{y = 0\} = 1$, then $x_n \xrightarrow{p} 0$ holds.

Ex. 6.4.2 Prove the following part of fact 6.1.1: $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$ and $\mathbf{y}_n \xrightarrow{p} \mathbf{y} \implies \mathbf{x}_n + \mathbf{y}_n \xrightarrow{p} \mathbf{x} + \mathbf{y}$. (The proof is nontrivial if you are still getting used to these kinds of arguments.)

Ex. 6.4.3 Show that $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$ implies $\mathbf{a}^\top \mathbf{x}_n \xrightarrow{p} \mathbf{a}^\top \mathbf{x}$ for every $\mathbf{a} \in \mathbb{R}^K$.

Ex. 6.4.4 We saw in fact 6.1.6 that if $x_n \xrightarrow{p} x$, then $x_n \xrightarrow{d} x$. Show that the converse is not in general true.

Ex. 6.4.5 Show that $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ if and only if every element of $\mathbf{X}_n$ converges in probability to the corresponding element of $\mathbf{X}$.³

Ex. 6.4.6 Confirm (6.10).⁴

Ex. 6.4.7 Let $\mathcal{L}(u) = U[0, 1]$ and $x_n := n\mathbb{1}\{0 \leq u \leq 1/n\}$ for $n \in \mathbb{N}$.

(i) Calculate the expectation of x_n^2 .

(ii) Show that $x_n \xrightarrow{p} 0$ as $n \rightarrow \infty$.

Conclude that $x_n \xrightarrow{p} x$ does not imply $x_n \xrightarrow{ms} x$.

Ex. 6.4.8 Let x be any random variable with $\mathbb{E}x = \mu$ and $\text{var } x = \sigma^2 < \infty$. Show that $x_n := x/n$ converges to zero in probability as $n \rightarrow \infty$.

Ex. 6.4.9 Suppose that $\{t_n\}$ is a sequence of random variables, θ is a constant, and

$$\sqrt{n}(t_n - \theta) \xrightarrow{d} N(0, \sigma^2) \quad \text{as } n \rightarrow \infty$$

3. Hint: To prove the "if" part use fact 3.3.13 on page 68.

4. Hint: Combine theorem 6.1.2 and fact 6.1.6.

Let $g: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable at θ with $g'(\theta) \neq 0$. Taking a first-order Taylor expansion of g around θ , we can write $g(t_n) = g(\theta) + g'(\theta)(t_n - \theta) + R(t_n - \theta)$, where $R(t_n - \theta)$ is a remainder term. As shown in theorem 3.1 of Van der Vaart (2000), under these conditions we have $\sqrt{n}R(t_n - \theta) \xrightarrow{p} 0$. Using this fact, prove that $\sqrt{n}\{g(t_n) - g(\theta)\} \xrightarrow{d} N(0, g'(\theta)^2 \sigma^2)$.

Ex. 6.4.10 Confirm the claim $\sqrt{N}(\bar{\mathbf{x}}_N - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \Sigma)$ in theorem 6.2.1.

Ex. 6.4.11 Let $\{\mathbf{x}_n\}$ be an IID sequence of random vectors in $\mathbb{R}^K$ with $\mathbb{E}\mathbf{x}_n = \mathbf{0}$ and $\text{var } \mathbf{x}_n = \mathbf{I}$. Let

$$\bar{\mathbf{x}}_N := \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \quad \text{and} \quad y_N := N \cdot \|\bar{\mathbf{x}}_N\|^2$$

What is the asymptotic distribution of $\{y_N\}$?

6.4.1 Solutions to Selected Exercises

Solution to Ex. 6.4.1. From the definition, $x_n \xrightarrow{p} 0$ means that, given any $\delta > 0$, we have $\mathbb{P}\{|x_n| > \delta\} \rightarrow 0$. If $x_n = y$ for all n and $\mathbb{P}\{y = -1\} = \mathbb{P}\{y = 1\} = 0.5$, then this fails for $\delta := 1/2$, say, since $\mathbb{P}\{|x_n| > 1/2\} = 1$ for all n .

However, if $\mathbb{P}\{y = 0\} = 1$, then for any $\delta > 0$ we have $\mathbb{P}\{|x_n| > \delta\} = \mathbb{P}\{|y| > \delta\} = 0$. This sequence is constant at zero and hence $x_n \xrightarrow{p} 0$ holds. $\square$

Solution to Ex. 6.4.2. Let $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$ and $\mathbf{y}_n \xrightarrow{p} \mathbf{y}$. Fix $\delta > 0$. Using the triangle inequality and elementary reasoning we can establish the following ordering of events:

$$\begin{aligned} \{\|\mathbf{x}_n + \mathbf{y}_n - (\mathbf{x} + \mathbf{y})\| > \delta\} &\subset \{\|\mathbf{x}_n - \mathbf{x}\| + \|\mathbf{y}_n - \mathbf{y}\| > \delta\} \\ &\subset \{\|\mathbf{x}_n - \mathbf{x}\| > \delta/2\} \cup \{\|\mathbf{y}_n - \mathbf{y}\| > \delta/2\} \end{aligned}$$

By monotonicity and subadditivity of $\mathbb{P}$ (see page 86), we then have

$$\mathbb{P}\{\|\mathbf{x}_n + \mathbf{y}_n - (\mathbf{x} + \mathbf{y})\| > \delta\} \leq \mathbb{P}\{\|\mathbf{x}_n - \mathbf{x}\| > \delta/2\} + \mathbb{P}\{\|\mathbf{y}_n - \mathbf{y}\| > \delta/2\}$$

By hypothesis, both terms on the right-hand side converge to zero in n . Hence $\mathbf{x}_n + \mathbf{y}_n \xrightarrow{p} \mathbf{x} + \mathbf{y}$ as claimed. $\square$

Solution to Ex. 6.4.3. Let $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$ in $\mathbb{R}^K$ and fix any $\mathbf{a} \in \mathbb{R}^K$. If $\mathbf{a} = \mathbf{0}$ the claim is trivial, so suppose instead that $\mathbf{a}$ is nonzero. Let $\delta > 0$. By the Cauchy–Schwarz inequality,

$$\{|\mathbf{a}^\top \mathbf{x}_n - \mathbf{a}^\top \mathbf{x}| > \delta\} \subset \{\|\mathbf{a}\| \|\mathbf{x}_n - \mathbf{x}\| > \delta\}$$

Using monotonicity of $\mathbb{P}$ now gives

$$\mathbb{P}\{|\mathbf{a}^\top \mathbf{x}_n - \mathbf{a}^\top \mathbf{x}| > \delta\} \leq \mathbb{P}\{\|\mathbf{x}_n - \mathbf{x}\| > \delta/\|\mathbf{a}\|\}$$

Since $\mathbf{x}_n \xrightarrow{p} \mathbf{x}$, the right-hand side converges to zero. $\square$

Solution to Ex. 6.4.4. On one hand, if $\{x_n\}$ and x are IID and $N(0, 1)$, then $\mathcal{L}(x_n) = \mathcal{L}(x)$ for all n , and hence $x_n \xrightarrow{d} x$. On the other hand, $\mathcal{L}(x_n - x) = N(0, 2)$ for all n . Hence $\mathbb{P}\{|x_n - x| \geq \delta\}$ is a positive constant for any $\delta > 0$, and $x_n \xrightarrow{p} x$ fails. $\square$

Solution to Ex. 6.4.10. Define

$$\mathbf{z}_N := \sqrt{N}(\bar{\mathbf{x}}_N - \boldsymbol{\mu}) \quad \text{and} \quad \mathcal{L}(\mathbf{z}) = N(\mathbf{0}, \Sigma)$$

We need to show that $\mathbf{z}_N \xrightarrow{d} \mathbf{z}$. To do this, we apply the Cramér–Wold device (fact 6.1.6, page 167) and the scalar CLT (theorem 6.1.2, page 168). To begin, fix $\mathbf{a} \in \mathbb{R}^K$. Observe that

$$\mathbf{a}^\top \mathbf{z}_N := \sqrt{N}(\bar{y}_n - \mathbb{E}[y_n])$$

where $y_n := \mathbf{a}^\top \mathbf{x}_n$. Since y_n is IID and

$$\text{var } y_n = \text{var}[\mathbf{a}^\top \mathbf{x}_n] = \mathbf{a}^\top \text{var}[\mathbf{x}_n] \mathbf{a} = \mathbf{a}^\top \Sigma \mathbf{a}$$

the scalar CLT yields

$$\mathbf{a}^\top \mathbf{z}_N \xrightarrow{d} N(0, \mathbf{a}^\top \Sigma \mathbf{a})$$

Since $\mathcal{L}(\mathbf{a}^\top \mathbf{z}) = N(0, \mathbf{a}^\top \Sigma \mathbf{a})$, we have shown that $\mathbf{a}^\top \mathbf{z}_N \xrightarrow{d} \mathbf{a}^\top \mathbf{z}$. Since $\mathbf{a}$ was arbitrary, the Cramér–Wold device tells us that $\mathbf{z}_N$ converges in distribution to $\mathbf{z}$. $\square$

Solution to Ex. 6.4.11. By assumption, $\{\mathbf{x}_n\}$ is an IID sequence in $\mathbb{R}^K$ with $\mathbb{E} \mathbf{x}_n = \mathbf{0}$ and $\text{var } \mathbf{x}_n = \mathbf{I}$. It follows from the vector central limit theorem that

$$\sqrt{N} \bar{\mathbf{x}}_N \xrightarrow{d} \mathbf{z} \quad \text{where} \quad \mathcal{L}(\mathbf{z}) = N(\mathbf{0}, \mathbf{I})$$

Letting $g(\mathbf{s}) := \|\mathbf{s}\|^2$ and applying the continuous mapping theorem (fact 6.1.6 on page 167), we obtain

$$y_N = \|\sqrt{N} \bar{\mathbf{x}}_N\|^2 \xrightarrow{d} \|\mathbf{z}\|^2 = \sum_{k=1}^K z_k^2$$

From fact 5.1.18 on page 140, we conclude that $y_N \xrightarrow{d} \chi^2(K)$. $\square$