Chapter 4
Convergence in Distribution and the Central Limit Theorem

4.1 Motivation: the central limit problem.

Consider a sequence of i.i.d. random variables $X_1, X_2, \ldots$ with finite mean μ and finite variance σ^2, and let $S_n = \sum_1^n X_i$, $n \geq 1$, denote the partial sum process. We want to address the question: what kind of distribution does S_n have for large n ? Clearly the distributions of S_n will not converge to some nice limit as n tends to ∞. While the law of large numbers shows that the empirical mean process $(1/n)S_n$ converges almost-surely to the actual mean μ, it is not correct to say that somehow $S_n \approx n\mu$. The problem is that the variance of $S_n - n\mu$, which is $n\sigma^2$, is large for large n. We know also from Kolmogorov's three series theorem that $\lim_{n\to\infty} S_n - n\mu$ will fail to exist almost-surely. However, suppose we form the sequence of scaled and centered partial sums

$$Z_n := \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Clearly Z_n has zero mean for all n; the scaling factor $1/\sigma\sqrt{n}$ is chosen precisely to make $\mathrm{Var}(Z_n) = 1$ for all n. Now it is reasonable to hope that as n increases Z_n will exhibit a simple asymptotic behavior.

Historically, the distribution of Z_n was first analyzed for i.i.d. Bernoulli random variables. Let $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = 1 - p$; then S_n is a binomial random variable which counts the number of successes ($X_i = 1$ is a success) in n trials. DeMoivre and Laplace showed in this case that

$$\lim_{n\to\infty} \mathbb{P}\Big(a < \frac{S_n - n\mu}{\sqrt{np(1-p)}} \leq b\Big) = \int_a^b (2\pi)^{-1/2} e^{-z^2/2}\, dz, \tag{1}$$

for any real numbers $a < b$. Observe that the integrand of the right-hand side is the density of a normal random variable Z with mean 0 and variance 1, which means that the right-hand side is just $\mathbb{P}(a < Z \leq b)$. Thus, (1) is saying that, for large n, Z_n is approximately a normal random variable with mean 0 and variance 1. However, the limit in (1) describes only what happens to the distributions of Z_n; there is no implication in (1) that Z_n converges to a normal random variable in probability. Indeed, it can be shown that $\{Z_n\}$ does not converge in probability, although we omit the proof. We shall describe (1) in words by saying that Z_n converges *in distribution* to a standard unit normal random variable. We treat the general definition of *convergence in distribution* in section 4.2.

The central limit problem, broadly speaking, is to find conditions under which sums of random variables, rescaled to have a fixed variance, converge in distribution to a normal. It is a fundamental discovery of probability theory that such convergence holds in a wide variety of circumstances, not just for i.i.d. Bernoulli random variables as in (1). The simplest general result of this sort dispenses with all distributional assumptions on an i.i.d. sequence, except that of finite variance.

Theorem 4.1.1 (Central Limit Theorem for i.i.d. sequences) If $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables with mean μ and variance σ^2, then

$$(2) \qquad \lim_{n\to\infty} I\!P(a < Z_n \le b) = \int_a^b (2\pi)^{-1/2} e^{-z^2/2}\, dz,$$

The natural question to ask next is what happens when the X_i random variables are independent, but not necessarily identically distributed. This issue is addressed by the Lindeberg-Feller theorem; it says, in effect, that if no one X_i can make the dominant contribution to S_n, for any n, the limit will be normal. The precise statement is as follows.

Theorem 4.1.2 Let $X_1, X_2 \ldots$ be independent random variables with finite means $\mu_i = E[X_i]$ and variances $\sigma_i^2 = \mathrm{Var}(X_i)$, $i \ge 1$. Let $B_n^2 = \sum_1^n \sigma_i^2$, and set

$$Z_n = \frac{S_n - \sum_1^n \mu_i}{B_n}, \qquad n \ge 1.$$

If

$$(3) \qquad \lim_{n\to\infty} \frac{1}{B_n^2} \sum_{k=1}^n E\left[(X_k - \mu_k)^2 \mathbf{1}_{\{|X_k - \mu_k| > \epsilon B_n\}}\right] = 0, \qquad \forall \epsilon > 0,$$

then

$$(4) \qquad \lim_{n\to\infty} I\!P(a < Z_n \le b) = \int_a^b (2\pi)^{-1/2} e^{-z^2/2}\, dz.$$

Conversely, if (4) holds and if

$$(5) \qquad \lim_{n\to\infty} \max_{1\le j\le n} I\!P(|X_j - \mu_j| \ge B_n \epsilon) = 0 \qquad \text{for positive } \epsilon.$$

then (3) must hold.

Theorem 4.2 is actually a special case of the general Lindeberg-Feller theorem. The condition (5) is called Lindeberg's condition. The converse statement was proved originally by Feller. Theorem 4.1 is clearly a Corollary of this more general statement.

To get a feeling for what the Lindeberg condition means, we observe that

$$\text{condition (3) implies} \quad \max_{1\le k\le n} \frac{\sigma_k^2}{B_n^2} \to 0 \quad \text{as } n \to \infty.$$

Thus the variance of each term in the sum is small relative to the variance of the sum. We shall prove this implication in a later lecture.

The aim in this chapter is to prove Theorems 4.1 and 4.2. The purpose of the discussion is not just to satisy ourselves that the results are true, but to develop fundamental tools and concepts that apply beyond the central limit problem. The first step is to define a notion of convergence that includes, as a particular case, the type of statement made in (1) or (4). A sequence of random variables Y_n, $n \geq 1$, is said to *converge in distribution* to a random variable Y if

$$(6) \qquad \lim_{n\to\infty} E[g(Y_n)] = E[g(Y)] \quad \text{for every bounded continuous } g.$$

The basic properties of convergence in distribution are explored in section 4.2. In particular, it is proved that (6) is equivalent to

$$(7) \quad \lim_{N\to\infty} \mathbb{P}(Y_n \leq b) = \mathbb{P}(Y \leq b) = \mathbb{F}_Y(b) \quad \text{for all continuity points } b \text{ of } \mathbb{F}_Y(b).$$

Hence the conclusion (4) of the central limit theorem may be re-expressed by saying Z_n converges in distribution to a standard, unit normal random variable.

For the proof of the central limit theorem, we shall employ characteristic functions. Recall that in Chapter 2 we defined the characteristic function of a random variable X as

$$\phi_X(\lambda) = E\left[e^{i\lambda X}\right] \qquad \lambda \in \mathbb{R}.$$

In section 4.4, we derive the so-called Continuity Theorem.

Theorem 4.1.3 Suppose that $\lim_{n\to\infty} \phi_{Y_n}(\lambda) = \phi_Y(\lambda)$ for all $\lambda \in \mathbb{R}$. Then Y_n converges to Y in distribution.

In section 4.5 we shall then prove the sufficiency part of Lindeberg's theorem by verifying convergence of characteristic functions.

4.2 Weak convergence and convergence in distribution

Let $\{X_n\}$ be a sequence of random variables with corresponding probability distributions $\{\mathbb{F}_{X_n}\}$. Convergence in distribution is concerned not with the way in which $\{X_n\}$ converges, as a sequence of functions on a probability space, but with the convergence of the associated sequence of probability distributions $\{\mathbb{F}_{X_n}\}$.

Definition. The sequence $\{\mathbb{F}_n\}$ of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is said to *converge weakly* to the probability measure $\mathbb{F}$ if

(1)

$$\int g(x)\, \mathbb{F}_n(dx) \to \int g(x)\, \mathbb{F}(dx) \quad \text{for all bounded, continuous functions } g \text{ on } \mathbb{R}.$$

In this case we write $\mathbb{F}_n \to \mathbb{F}$ (weakly).

We say that the sequence of random variables $\{X_n\}$ converges in distribution to the random variable X if $\mathbb{F}_{X_n} \to \mathbb{F}_X$ (weakly). Alternatively, $X_n \to X$ (in distribution), if $\lim_{n\to\infty} E[g(X_n)] = E[g(X)]$ for every bounded, continuous g. $\diamond$

The following remarks explain the scope and meaning of these definitions.

Remark 1. Weak convergence is associated to a topology on $\mathcal{P}(I\!\!R)$, the space of probability measures on the Borel sets of $I\!\!R$. What is this topology? Let C_0 denote the set of continuous functions on $I\!\!R$ vanishing at infinity and make C_0 a Banach space by endowing it with the supremum norm. Then Riesz's theorem says that the topological dual of C_0 is the space M of all bounded Borel measures on $I\!\!R$. In functional analysis, one defines the weak-* topology on M to be the smallest topology with respect to which the maps $M \ni \mu \to \int_{I\!\!R} g(x)\,\mu(dx)$ are continuous for every $g \in C_0$. In other words, the sequence of bounded measures μ_n converges to μ in the weak-* topology if and only if

$$\int g(x)\,\mu_n(dx) \to \int g(x)\,\mu(dx) \quad \text{for every } g \in C_0. \tag{2}$$

Now the set of Borel probability measures on $I\!\!R$ is a subset of M. It turns out that $I\!\!F_n \to I\!\!F$ (weakly) if and only if $I\!\!F_n \to I\!\!F$ in the weak-*topology. Certainly (2) implies weak-* convergence because C_0 is a subset of the set of all bounded, continuous functions. We leave the converse as an exercise. Hence, what probabilists call weak convergence is the same as what analysts call weak-* convergence.

Since analysts reserve the use of weak convergence to describe a different topology on a linear topological space, the probabilists' use of *weak* rather than *weak-** to describe the convergence in (2) is a mild source of confusion. It is an historical accident now sanctified by too long and widespread a usage to be changed. So you should just remember that in probability, weak really means weak-*.

Remark 2. In the more advanced part of probability theory, it is important to consider convergence in distribution of random vectors, and even more generally, of random variables taking values in a metric space. In fact, in the definition of weak convergence, nothing specifically is used about the structure of $I\!\!R$ beyond its usual topology, which provides a set of continuous functions. Hence we could frame the definition of weak convergence in the general context of probability measures on topological spaces. However, extensions of the convergence theorems we state work mostly only for metric spaces.

Weak Convergence (General Definition). Let μ_n, $n \geq 1$ and μ be probability measures on the Borel sets of a metric space S. We say that $\mu_N \to \mu$ (weakly) if $\int g(x)\,\mu_n(dx) \to \int g(x)\,\mu(dx)$ for every bounded continuous function g on S. $\diamond$.

Remark 3. Why do we adopt this *weak* notion of convergence?. After all, why not use a total variation distance between probability measures, which requires no topology on the underlying space? It turns out that this would be too strong. For example, let $\delta_{1/n}$ be the atomic measure which assigns mass one to the point $1/n$. Think of $\delta_{1/n}$ as the probability distribution of the degenerate random variable which is almost surely equal to $1/n$. Let $\mathbf{0}$ denote the random variable identically equal to 0. Then X_n will converge to $\mathbf{0}$ in distribution: if g is continuous,

$\int g(x)\delta_{1/n}(dx) = g(1/n)$ and $\lim g(1/n) = g(0) = \int g(x)\,\delta_0(dx)$. However $\delta_{1/n}$ does not converge to δ_0 in total variation. Indeed, since $\delta_{1/n}(\{0\}) = 0$ for every n, and $\delta_0(\{0\}) = 1$, the total variation between them is 2 for every n.

The first result we prove gives a number of equivalent characterizations of weak convergence, including that of convergence of the distribution function at continuity points, as stated in equation (7) of section 1. In this theorem and in the remainder of the section, $F_n(b) := I\!\!F_n((-\infty, b])$, $F(b) := I\!\!F((-\infty, b])$, etc.

Theorem 4.2.1 The following are equivalent:
(a) $I\!\!F_n \to I\!\!F$ (weakly);
(b) $\liminf I\!\!F_n(G) \geq I\!\!F(G)$ for every open set G;
(c) $\limsup I\!\!F_n(C) \leq I\!\!F(C)$ for every closed set C;
(d) $F_n(b) \to F(b)$ for every continuity point b of F.

Proof: ((a) implies (b)) Given any open set G, there is an increasing sequence $\{g_k\}$ of continuous functions such that $0 \leq g_k(x) \leq \mathbf{1}_G(x)$ for all x and $g_k \uparrow \mathbf{1}_G$ as $k \to \infty$. For example, let $g_k(x) = \min\{1, k\mathrm{dist}(x, G^c)\}$. Thus, using (a),

$$\liminf_{n\to\infty} I\!\!F_n(G) \geq \liminf_{n\to\infty} \int g_k(x)\, I\!\!F_n(dx) = \int g_k(x)\, I\!\!F(dx)$$

for every k. Now take limits as $k \to \infty$ and use the dominated convergence theorem to prove

$$\liminf_{n\to\infty}\ I\!\!F_n(G) \geq \int \mathbf{1}_G(x)\, I\!\!F(dx) = I\!\!F(G).$$

((b) iff (c)) Observe that $I\!\!F_n(C) = 1 - I\!\!F(_nC^c)$. Applying (b) to $I\!\!F_n(C^c)$ gives (c). Likewise (c) implies (b).

((c) implies (d)) For every b, (c) implies

$$\limsup F_n(b) = \limsup I\!\!F_n((-\infty, b]) \leq I\!\!F((-\infty, b]) = F(b). \tag{3}$$

On the other hand (c) implies (b) and so

$$\liminf F_n(b) \geq \liminf I\!\!F_n((-\infty, b)) \geq I\!\!F((-\infty, b)). \tag{4}$$

But if b is a continuity point of F, we know $F(b) = I\!\!F((-\infty, b))$, and so combining (3) and (4) yields

$$\limsup F_n(b) \leq F(b) \leq \liminf F_n(b),$$

and (d) follows.

((d) implies (b)) Since any open set in $I\!\!R$ is a countable union of disjoint open intervals, it suffices to show that (d) implies that $\liminf I\!\!F_n((a, b)) \geq I\!\!F((a, b))$ for any $-\infty \leq a < b$. Let (a_k, b_k) be a sequence of open intervals such that

$(a_k, b_k) \subset (a, b)$ for every k, $a_k \downarrow a$, $b_k \uparrow b$, and a_k and b_k are continuity points of $I\!\!F$ for all k. Then, using (d),

$$
\begin{aligned}
\liminf I\!\!F_n((a,b)) &\geq \liminf I\!\!F_n((a_k, b_k)) \\
&= \liminf F_n(b_k) - F_n(a_k) \\
&= F(b_k) - F(a_k) = I\!\!F((a_k, b_k))
\end{aligned} \tag{5}
$$

for every k. Since $I\!\!F((a_k, b_k)) \uparrow I\!\!F((a,b))$ as $k \to \infty$, taking $k \to \infty$ in (5) gives $\liminf I\!\!F_n((a,b)) \geq I\!\!F((a,b))$ as desired.

((b) implies (a)) (b) implies (c) and so for any A,

$$I\!\!F(\text{int}(A)) \leq \liminf I\!\!F_n(A) \leq \limsup I\!\!F_n(\bar{A}) \leq I\!\!F(\bar{A}).$$

It follows that if $I\!\!F(\text{int}(A)) = I\!\!F(\bar{A})$, or, equivalently $I\!\!F(\bar{A} - \text{int}(A)) = 0$, then $\lim_{n\to\infty} I\!\!F_n(A) = F(A)$. This is a useful preliminary.

Now let g be a bounded, continuous function. We claim that for every positive integer k there is a sequence of simple functions $\{g_k\}$ such that $\|g_k - g\| := \sup_x |g_k(x) - g(x)| \leq 1/k$ for each k and such that

$$\lim_{n\to\infty} \int g_k(x)\, I\!\!F_n(dx) = \int g_k(x)\, I\!\!F(dx) \quad \text{for each } x. \tag{6}$$

This will suffice to complete the proof because

$$
\begin{aligned}
|\int g(x)\, I\!\!F_n(dx) - \int g(x)\, I\!\!F(dx)| &\leq \int |g_k(x) - g(x)|\, I\!\!F_n(dx) \\
&+ \int |g_k(x) - g(x)|\, I\!\!F(dx) + |\int g_k(x)\, I\!\!F_n(dx) - \int g_k(x)\, I\!\!F(dx)| \\
&\leq \frac{2}{k} + |\int g_k(x)\, I\!\!F_n(dx) - \int g_k(x)\, I\!\!F(dx)|.
\end{aligned}
$$

Thus, from (6),

$$\limsup_{n\to\infty} |\int g(x)\, I\!\!F_n(dx) - \int g(x)\, I\!\!F(dx)| \leq \frac{2}{k}.$$

Since k is arbitrary, we can take $k \to \infty$ and recover (a).

It remains to construct the sequence g_k. Without loss of generality, assume $-1 < g(x) < 1$ for all x. Choose a partition $a_0 = -1 < a_1 < \cdots < a_{m_k} = 1$ so that $I\!\!F(g^{-1}(\{a_i\})) = 0$ for every i and $a_1 - a_{i-1} < 1/k$ for every i. (We can do this because the set of values $N = \{y|\ I\!\!F(g^{-1}(\{y\}) > 0\}$ is at most countable since the sets $g^{-1}(\{y\})$ are disjoint for different y. For any countable subset $J \subset N$, $\sum_{y\in J} I\!\!F(g^{-1}(\{y\}) \leq 1$ and so $\sup\{\sum_{y\in J} I\!\!F(g^{-1}(\{y\}) \mid J \subset N \text{ is countable}\} \leq 1$. This can only occur if N is countable.) For each i, set $A_i = \{x|\ a_i < g(x) \leq a_{i+1}\}$ and define $g_k(x) = \sum a_i \mathbf{1}_{A_i}(x)$. It is clear that $\|g_k - g\| \leq 1/k$. Hence we need only to show (6). For this, note that $\bar{A}_i - \text{int}(A_i) \subset g^{-1}(\{a_{i+1}\}) \cup g^{-1}(\{a_i\})$ and so $I\!\!F(\bar{A}_i - \text{int}(A_i)) = 0$ and hence $I\!\!F_n(A_i) \to I\!\!F(A_i)$ for all k. Thus

$$\int g_k(x)\, F_n(dx) = \sum a_i I\!\!F_n(A_i) \to \sum a_i I\!\!F(A_i) = \int g_k(x)\, I\!\!F(dx)$$

as $n \to \infty$, proving (6). $\diamond$

Exercise. Show that $I\!\!F_n \to I\!\!F$ (weakly) iff

$$\int g(x)\, I\!\!F_n(dx) \to \int g(x)\, I\!\!F(dx) \quad \text{for every } G \in C_0.$$

Exercise. Verify that the equivalence of (a),(b), and (c) of Theorem 4.2.1 continue to be true for weak convergence of probability measures on a metric space S. (If this is too abstract, first consider $S = I\!\!R^n$ with Euclidean metric.)

Theorem 4.2.1 has a useful consequence for absolutely continuous distribution measures. Suppose that each $I\!\!F_n$ admits a density f_n; that is, $I\!\!F_n(A) = \int_A f_n(x)\, dx$ for all Borels sets A. Similarly, assume $I\!\!F$ admits the density f.

Theorem 4.2.2 For probability distributions with densities, if $\lim_{n\to\infty} f_n(x) = f(x)$ Lebesgue-almost-everywhere, then $I\!\!F_n \to I\!\!F$ (weakly).

Proof: Let G be any open set. By Fatou's lemma,

$$\begin{aligned} I\!\!F(G) = \int_G f(x)\, dx &= \int_G \liminf f_n(x)\, dx \\ &\le \liminf \int_G f_n(x)\, dx = \liminf I\!\!F_n(G). \end{aligned}$$

Hence by Theorem 4.2.1, part (b), $I\!\!F_n \to I\!\!F$ (weakly). $\diamond$

Exercise. Prove Scheffe's theorem: If $\{I\!\!F_n\}$ is a sequence of probability distributions on $I\!\!R$ with densities $\{f_n\}$ and $\lim_{n\to\infty} f_n(x) = f(x)$ Lebesgue-almost-everywhere, where f is the density of probability distribution $I\!\!F$, then $I\!\!F_n \to I\!\!F$ in total variation.
(Hint: Convergence of $I\!\!F_n$ to $I\!\!F$ in total variation is equivalent to convergence of f_n to f in $L^1(I\!\!R)$. If $A_n = \{x\,;\, f(x) > f_n(x)\}$, show that $\int |f - f_n|\, dx = 2\int_{A_n} (f - f_n)\, dx$.)
Note that Sheffe's theorem implies Theorem 4.2.2. $\diamond$

What relation does convergence in distribution have to other types of convergence? First note that to say $X_n \to X$ (in distribution) is really to make a statement about the associated probability distribution measures $I\!\!F_{X_n}$. There is even no need to have all the X_n and X defined on the same probability space. But when the random variables are all on one probability space and there is convergence in probability, convergence in distribution follows.

Theorem 4.2.3 If $X_n \to X$ in probability, then $X_n \to X$ in distribution.

Proof: Assume first that $X_n \to X$ almost-surely. Then, if g is any continuous bounded function, $\lim_{n\to\infty} E[g(X_n)] = E[g(X)]$, by dominated convergence. Hence, almost sure convergence implies convergence in distribution.

Now suppose $X_n \to X$ in probability. We shall use the following fact, whose easy proof-by-contradiction we leave to the reader. Let a_n be a sequence of real numbers such that each subsequence of $\{a_n\}$ contains a sub-subsequence converging to a finite a; then $\lim_{n\to\infty} a_n = a$. Let g be any continuous, bounded function. Consider a subsequence $|E[g(X_{n_k})]$ of $E[g(X_n)]$. Since X_{n_k} converges to X in probability, we may extract from $\{X_{n_k}\}$ a sub-subsequence, call it $\{X_{n'_k}\}$, converging to X almost surely. Then $E[g(X_{n'_k})] \to E[g(X)]$ because we have shown that almost sure convergence implies convergence in distribution. Thus we have shown that every subsequence of $E[g(X_n)]$ contains a sub-subsequence converging to $E[g(X)]$, which implies $E[g(X_n)]$ itself converges to $E[g(X)]$. Since g was an arbitray, bounded, continuous function, the proof is done. $\diamond$

4.3. Characteristic functions; uniqueness and smoothing

In Chapter 2 we defined the characteristic function

$$\phi_X(\lambda) = E[e^{i\lambda X}] \qquad \lambda \in I\!R$$

of a random variable of X. We proved in Corollary D.4 of Chapter 2 that for independent random variables $X_1, \ldots, X_n$ and their sum $S = X_1 + \cdots + X_n$,

$$\phi(\lambda) = \phi_{X_1}(\lambda) \cdots \phi_{X_n}(\lambda). \tag{1}$$

For the work of this chapter, the characteristic function of a normal random variable Z with mean μ and variance σ^2 is essential. It is:

$$\phi_Z(\lambda) = e^{i\mu\lambda - \sigma^2\lambda^2/2}. \tag{2}$$

We also pointed out in Chapter 2 that the characteristic function of X is essentially the Fourier transform of the distribution measure $I\!F_X$. In elementary Fourier analysis, we are used to thinking of Fourier transforms of *functions* rather than distributions. For example if X has density f_X, its characteristic function is

$$\phi_X(\lambda) = \int e^{i\lambda x} f_X(x)\, dx,$$

which is the Fourier transform of f_X. But the theory of Fourier transforms of measures is not too much more involved than that for functions. In fact, we shall use a smoothing operation to approximate Fourier transforms of probability measures by Fourier transforms of smooth density functions, and apply the usual theory. We shall derive in this way the following Fourier inversion formula.

Theorem 4.3.1 Let X be any random variable. Let $a < b$ be continuity points of F_X, the distribution function of X. Then

$$F_X(b) - F_X(a) = \lim_{\epsilon \downarrow 0} \frac{1}{2\pi} \int e^{-\epsilon^2 \lambda^2/2} \phi_X(\lambda) \frac{e^{-i\lambda b} - e^{-i\lambda a}}{-i\lambda} \, d\lambda. \tag{3}$$

As a consequence we prove that characteristic functions uniquely characterize probability distributions.

Corollary 4.3.2 Let X and Y be two random variables. If $\phi_X(\lambda) = \phi_Y(\lambda)$ for all λ, then $F_X = F_Y$.

Proof: The set of points N at which either F_X or F_Y is not continuous is countable. By Theorem 1, $F_X(b) - F_X(a) = F_Y(b) - F_Y(a)$ for any points a and b not in N for which $a < b$. Fix a point b not in N and choose a sequence of points $\{a_n\}$ such that $a_n \downarrow -\infty$ and $a_n \notin N$ for every n. Then taking $n \to \infty$ in $F_X(b) - F_X(a_n) = F_Y(b) - F_Y(a_n)$, shows that $\mathbb{F}_X(b) = \mathbb{F}_Y(b)$ for all $b \notin N$. Two right continuous functions which agree except on a countable set must in fact be equal to each other everywhere. Hence $F_X = F_Y$. $\diamond$

Here is the main technical trick we shall use in the proof. Let X be a random variable with an arbitrary distribution and let Z denote a standard normal random variable (i.e. $E[Z] = 0$ and $\mathrm{Var}(Z) = 1$) independent of X. We shall see that adding ϵZ to X for positive numbers ϵ has a smoothing effect on the distribution. We can study this effect either in the domain of distribution functions or in the domain of characteristic functions. Consider distribution functions first.

Let X have an arbitrary distribution $\mathbb{F}_X$. Observe that if Z is a standard, unit normal, ϵZ is a with mean 0 and variance ϵ^2, and so its density function is

$$f_\epsilon(x) = (2\pi\epsilon^2)^{-1/2} e^{-x^2/2\epsilon^2}. \tag{4}$$

Then, since we assume X and Z are independent, we can apply Proposition D.5 of Chapter 2 to show

$$\begin{aligned} F_{X+\epsilon Z}(x) &= \int_{-\infty}^{\infty} \int_{-\infty}^{x-y} f_\epsilon(z) \, dz \, \mathbb{F}_X(dy) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{x} f_\epsilon(z-y) \, dz \mathbb{F}_X(dy) \; . \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_\epsilon(z-y) \, \mathbb{F}_X(dy) \, dz \end{aligned}$$

To derive the last equality we used Tonelli's theorem. It follows that $X + \epsilon Z$ is a continuous random variable with probability density

$$f_{X+\epsilon Z}(x) = F'_{X+\epsilon Z}(x) = \int_{-\infty}^{\infty} f_\epsilon(x-y)\, I\!F(dy). \tag{5}$$

Since f_ϵ is infinitely differentiable with bounded derivatives of all orders it is easy to see that the density in (5) is infinitely differentiable as well. These observations are simple extensions of properties that are well-known in real analysis. The family $\{f_\epsilon;\ \epsilon > 0\}$ constitutes what is known as a smooth approximation of the identity; it can be shown that if $1 \le p < \infty$ and if g is a function in $L^p(I\!R)$, then

$$f_\epsilon * g \in C^\infty, \quad \forall \epsilon > 0, \qquad \text{and} \qquad \lim_{\epsilon\downarrow 0} ||f_\epsilon * g - g||_{L^p} = 0.$$

In analogy, we may think of the density $f_{X+\epsilon Z}$ as $f_\epsilon * I\!F_X$. It makes no sense to study L^p convergence of $f_{X+\epsilon Z}$, but we easily see that

$$I\!F_{X+\epsilon Z} \to I\!F \text{ (weakly) } \text{ as } \epsilon \downarrow 0. \tag{6}$$

Indeed, (6) follows directly from Theorem 4.2.3, because $X + \epsilon Z \to Z$ a.s. as $\epsilon \downarrow 0$.

Consider now what happens with the characteristic functions. Because X and Z are independent, the characteristic function of $X + \epsilon Z$ is

$$\phi_{X+\epsilon Z}(\lambda) = \phi_{\epsilon Z}(\lambda)\phi_X(\lambda) = e^{-\epsilon^2\lambda^2/2}\phi_X(\lambda).$$

We know that $|\phi_X(\lambda)| \le 1$ for all real λ. Hence $\phi_{X+\epsilon Z}$ decays rapidly at ∞ and is in fact an L^p function for any $p \ge 1$. It is then easy to establish a Fourier inversion theorem for $\phi_{X+\epsilon Z}$. First observe that the characteristic function $e^{-\epsilon^2\lambda^2/2}$ of the normal is, modulo a factor of $(\epsilon^2/2\pi)^{-1/2}$, the density of a normal with mean 0 and variance ϵ^{-2}. Thus

$$\frac{1}{2\pi}\int e^{-i\lambda z} e^{-\epsilon^2\lambda^2/2}\, d\lambda = (2\pi\epsilon^2)^{-1/2} e^{-z^2/2\epsilon^2}, \tag{7}$$

which verifies the Fourier inversion formula for normal densities. Next notice that for any fixed x, $e^{-i\lambda x}e^{-\epsilon^2\lambda^2/2}e^{i\lambda y}$ is integrable on $I\!R^2$ as a function of (λ, y) with respect to the measure $m \times I\!F_X$, where m denotes Lebesgue measure. Indeed,

$$\begin{aligned}\int\int |e^{-i\lambda x}e^{-\epsilon^2\lambda^2/2}e^{i\lambda y}|\, I\!F_X(dy)\, d\lambda &\le \int\int e^{-\epsilon^2\lambda^2/2}\, I\!F_X(dy)\, d\lambda \\ &= \int e^{-\epsilon^2\lambda^2/2}\, d\lambda < \infty.\end{aligned}$$

Fubini's theorem therefore applies:

$$\begin{aligned}\frac{1}{2\pi}\int e^{-i\lambda x}e^{-\epsilon^2\lambda^2/2}\phi_X(\lambda)\, d\lambda &= \frac{1}{2\pi}\int \Big[\int e^{-i\lambda x}e^{-\epsilon^2\lambda^2/2}e^{i\lambda y}\, I\!F_X(dy)\Big]\, d\lambda \\ &= \frac{1}{2\pi}\int \Big[\int e^{-i\lambda(x-y)}e^{-\epsilon^2\lambda^2/2}\, d\lambda\Big]\, I\!F_X(dy).\end{aligned}$$

Using the inversion formula (7) in the inner integral of the last expression and then formula (5) for $F'_{X+\epsilon Z}$ gives

$$\begin{aligned}\frac{1}{2\pi}\int e^{-i\lambda x}e^{-\epsilon^2\lambda^2/2}\phi_X(\lambda)\,d\lambda &= \int \frac{e^{-(y-x)^2/2\epsilon^2}}{\sqrt{2\pi\epsilon^2}}\,dF_X(y)\\ &= \quad F'_{X+\epsilon Z}(x)\end{aligned}$$

As a result, with another application of Fubini's theorem,

$$\begin{aligned}F_{X+\epsilon Z}(b)-F_{X+\epsilon Z}(a) &= \int_a^b \frac{1}{2\pi}\int e^{-i\lambda x}e^{-\epsilon^2\lambda^2/2}\phi_X(\lambda)\,d\lambda\,dx\\ &= \frac{1}{2\pi}\int e^{-\epsilon^2\lambda^2/2}\phi_X(\lambda)\frac{e^{-i\lambda b}-e^{-i\lambda a}}{-i\lambda}\,d\lambda\end{aligned} \tag{8}$$

Proof of Theorem 4.3.1: The proof follows immediately from equations (6) and (8). $\diamond$

Remark. We have been using ϕ_X to denote the characteristic function associated to the random variable X. However this notation might obscure somewhat the point that the characteristic function depends only on the probability distribution function $I\!F_X$ and not on any other feature of X. It would be a little more consistent to define the characteristic function of a probability measure $I\!F$ as $\phi_{\boldsymbol{F}}(\lambda) = \int e^{i\lambda x} I\!F(dx)$. Then Φ_X of our previous notation becomes $\phi_{\boldsymbol{F}_X}$; we stick with the old notation as it saves us some subscripts. Keep in mind however that the theory of characteristic functions is really about Fourier transforms of probability measures and not directly about random variables.

4.4 The Continuity Theorem.

In this section we prove the Continuity Theorem, stated in Theorem 4.1.3; if $\phi_{X_n}(\lambda) \to \phi_X(\lambda)$ for all λ, then $I\!F_{X_n} \to I\!F_X$ (weakly). We first establish this for regularized random variables.

Lemma 4.4.1 Let Z be a standard, unit normal independent of every X_n. If $\phi_{X_n}(\lambda) \to \phi_X(\lambda)$ for every λ, then for any $\epsilon > 0$

$$I\!F_{X_n+\epsilon Z} \to I\!F_{X+\epsilon Z} \text{ (weakly)}.$$

Proof: Fix $\epsilon > 0$. In the derivation of the inversion formula (8) in section 3, we showed that the density

$$f_{X_n+\epsilon}(x) = \frac{1}{2\pi}\int e^{-i\lambda x}e^{-\epsilon^2\lambda^2/2}\phi_{X_n}(\lambda)\,d\lambda. \tag{1}$$

Take limits in (1) as $n \to \infty$; because $e^{-\epsilon^2\lambda^2/2}$ is integrable and because $|\phi_{X_n}(\lambda)| \leq 1$ for all $\lambda \in \mathbb{R}$ and all n, the dominated convergence theorem allows us to conclude,

$$\lim_{n\to\infty} f_{X_n+\epsilon}(x) = \frac{1}{2\pi}\int e^{-i\lambda x} e^{-\epsilon^2\lambda^2/2}\phi_X(\lambda)\,d\lambda = f_X(x),$$

for every x. It follows now from Theorem 4.3.2 that $\mathbb{F}_{X_n+\epsilon Z} \to \mathbb{F}_{X+\epsilon Z}$ (weakly). $\diamond$

To finish the proof of Theorem 4.1.3, we shall use the following simple inequalities

$$\begin{aligned} F_{X_n}(b) &= \mathbb{P}(X_n \leq b) \\ &\leq \mathbb{P}\left(X_n \leq b, \epsilon|Z| < \delta\right) + \mathbb{P}\left(\epsilon|Z| \geq \delta\right) \\ &\leq F_{X_n+\epsilon}(b+\delta) + \mathbb{P}\left(\epsilon|Z| \geq \delta\right), \end{aligned} \tag{2}$$

and

$$\begin{aligned} F_{X_n+\epsilon Z}(b-\delta) &= \mathbb{P}\left(X_n + \epsilon Z \leq b-\delta, \epsilon|Z| < \delta\right) + \mathbb{P}\left(\epsilon|Z| \geq \delta\right) \\ &\leq F_{X_n}(b) + \mathbb{P}\left(\epsilon|Z| \geq \delta\right). \end{aligned} \tag{3}$$

Apply Lemma 4.4.1 to (2) and (3), noting that for any $\epsilon > 0$, $F_{X+\epsilon Z}$ is continuous; then for any b, and $\epsilon > 0$, and any $\delta > 0$,

$$\limsup_{n\to\infty} F_{X_n}(b) \leq F_{X+\epsilon Z}(b+\delta) + \mathbb{P}\left(\epsilon|Z| \geq \delta\right), \tag{4}$$

and

$$\liminf_{n\to\infty} F_{X_n}(b) \geq F_{X+\epsilon Z}(b-\delta) - \mathbb{P}\left(\epsilon|Z| \geq \delta\right). \tag{5}$$

Since $X + \epsilon Z$ converges a.s. to X as $\epsilon \downarrow 0$, we know that $\mathbb{F}_{X+\epsilon Z} \to \mathbb{F}_X$ (weakly). Because the interval $(-\infty, b]$ is closed, Theorem 4.2.1 (c) then implies

$$\lim_{\epsilon\to 0} F_{X+\epsilon Z}(b) = \lim_{\epsilon\to 0} \mathbb{P}\left(X + \epsilon Z \leq b\right) \leq F_X(b),$$

for any b. Let $\delta > 0$ be such that $b - \delta$ is a continuity point of $\mathbb{F}_X$. Then, taking $\epsilon \downarrow 0$ in (4) and (5),

$$F_X(b-\delta) \leq \liminf_{n\to\infty} F_{X_n}(b) \leq \limsup_{n\to\infty} F_{X_n}(b) \leq F_X(b+\delta), \tag{6}$$

The discontinuity points of any distribution function are at most countable in number, and hence there is a sequence $\delta_k \to 0$ of positive numbers such that $b - \delta_k$ is a continuity point of F_X for every k. Letting $\delta \downarrow 0$ in (6) along the sequence $\{\delta_k\}$, it follows that for any b

$$F_X(b-) \leq \liminf_{n\to\infty} F_{X_n}(b) \leq \limsup_{n\to\infty} F_{X_n}(b) \leq F_X(b).$$

When b is a continuity point, $F_X(b-) = F_X(b)$, and so, at all continuity points, $\lim_{n\to\infty} F_{X_n}(b) = F_X(b)$. By part (d) of Theorem 4.2.1, $\mathbb{F}_{X_n} \to \mathbb{F}_X$ (weakly). $\diamond$

There is a stronger version of the continuity theorem that we state now.

Theorem 4.4.2 Let $\{\mathbb{F}_n\}$ be a sequence of probability distribution measures with corresponding characteristic functions $\phi_{\mathbb{F}_n}$. Assume that $\psi(\lambda) = \lim_{n\to\infty} \phi_{\mathbb{F}_n}(\lambda)$ exists for every λ and ψ is continuous at λ. Then there is a probability distribution measure $\mathbb{F}$ such that $\psi(\lambda) = \phi_{\mathbb{F}}(\lambda)$, and $\mathbb{F}_n \to \mathbb{F}$ (weakly).

The novel feature in this theorem is that the limit will be a characteristic function of a probability measure if it is continuous at zero. Once this is proved, the weak convergence, is a consequence of Theorem 4.1.3. We omit the proof for now.

4.5 Proof of Lindeberg's Central Limit Theorem

In this section we shall prove part of the Central Limit Theorem stated in Theorem 4.1.2. For convenience, we restate exactly what we intend to prove.

Theorem 4.5.1 Let $X_1, X_2 \ldots$ be independent random variables with finite means $\mu_i = E[X_i]$ and variances $\sigma_i^2 = \mathrm{Var}(X_i)$, $i \geq 1$. Let $B_n^2 = \sum_1^n \sigma_i^2$, and set

$$Z_n = \frac{\sum_1^n X_i - \mu_i}{B_n}, \qquad n \geq 1.$$

If

$$\text{(1)} \qquad \lim_{n\to\infty} \frac{1}{B_n^2} \sum_{k=1}^{n} E\left[(X_k - \mu_k)^2 \mathbf{1}_{\{|X_k-\mu_k|>\epsilon B_n\}}\right] = 0, \qquad \forall \epsilon > 0,$$

then

$$\text{(2)} \qquad \lim_{n\to\infty} \mathbb{P}(a < Z_n \leq b) = \int_a^b (2\pi)^{-1/2} e^{-z^2/2}\, dz.$$

Clearly, there is no loss of generality in the proof if we assume that the means $\mu_k = 0$ to begin with. This assumption will be in force for the remainder of the section.

As we mentioned in section 1, the Lindeberg condition has a simple consequence for the magnitude of the variances of the individual summands in the numerator of Z_n.

Lemma 4.5.2 If condition (1) holds,

$$\max_{1\le k\le n} \frac{\sigma_k^2}{B_n^2} \to 0 \qquad \text{as } n\to\infty. \tag{3}$$

Proof: For any $\epsilon > 0$, and

$$\begin{aligned}\max_{1\le k\le n} \frac{\sigma_k^2}{B_n^2} &\le \frac{1}{B_n^2} \max_{1\le k\le n} \left(E\left[X_k^2 \mathbf{1}_{\{|X_k|\le \epsilon B_n\}}\right] + E\left[X_k^2 \mathbf{1}_{\{|X_k|>\epsilon B_n\}}\right]\right) \\ &\le \epsilon^2 + \frac{1}{B_n^2}\sum_1^n E\left[X_k^2 \mathbf{1}_{\{|X_k|>\epsilon B_n\}}\right].\end{aligned}$$

Letting $n\to\infty$, using (1), and then taking $\epsilon \downarrow 0$, completes the proof. $\diamond$

The proof of Theorem 4.5.1 will also use a technical result on the Taylor expansion of the characteristic function. It is convenient to present this next.

Lemma 4.5.3 Let X be a mean 0 random variable with finite variance σ^2. For any $\delta > 0$, its characteristic function $\phi_X(\lambda)$ satisfies

$$|\phi_X(\lambda) - (1 - \frac{\lambda^2}{2}\sigma^2)| \le \lambda^2 E\left[X^2 \mathbf{1}_{\{|X|>\delta\}}\right] + \frac{\lambda^3 \delta}{6} E\left[X^2 \mathbf{1}_{\{|X|\le\delta\}}\right]. \tag{4}$$

Proof: The Taylor remainder bound implies

$$\begin{aligned}|e^{i\lambda x} - (1 + i\lambda x)| &\le \frac{\lambda^2}{2}x^2, \qquad \text{and} \\ |e^{i\lambda x} - (1 + i\lambda x - \frac{\lambda^2}{2}x^2)| &\le \frac{\lambda^3}{6}|x|^3.\end{aligned}$$

Notice that $\phi_X(\lambda) - (1 - (\lambda^2/2)\sigma^2) = E\left[e^{i\lambda X} - \left(1 + i\lambda X - (\lambda^2/2)X^2\right)\right]$. It follows that

$$\begin{aligned}|\phi_X(\lambda) - (1 - \frac{\lambda^2}{2}\sigma^2)| &\le E\left[|e^{i\lambda X} - (1+i\lambda X)|\mathbf{1}_{\{|X|>\delta\}}\right] + E\left[\lambda^2 X^2 \mathbf{1}_{\{|X|>\delta\}}\right] \\ &\quad + E\left[|e^{i\lambda X} - \left(1 + i\lambda X - (\lambda^2/2)X^2\right)|\mathbf{1}_{\{|X|\le\delta\}}\right] \\ &\le \lambda^2 E\left[X^2 \mathbf{1}_{\{|X|>\delta\}}\right] + \frac{\lambda^3}{6} E\left[|X|^3 \mathbf{1}_{\{|X|\le\delta\}}\right]\end{aligned}$$

Since $E[|X|^3 \mathbf{1}_{\{|X|\le\delta\}}] \le \delta E[|X|^2 \mathbf{1}_{\{|X|\le\delta\}}]$, inequality (4) follows. $\diamond$

Finally, we need a purely analytic lemma.

Lemma 4.5.4 Let $\{s_{kn}\}$, $1 \le k \le n$, $n \ge 1$ be complex numbers satisfying

$$(5) \qquad \lim_{n\to\infty} \max_{1\le k\le n} |s_{kn}| = 0, \qquad \sup_{n\ge 1} \sum_{k=1}^{n} |s_{kn}| < \infty, \quad \text{and} \qquad \lim_{n\to\infty} \sum_{k=1}^{n} s_{kn} = \alpha.$$

Then $$\prod_{k=1}^{n}(1+s_{kn}) \to e^{\alpha} \quad \text{as } n \to \infty.$$

Proof: Let $\ln z$ denote the principal branch of the logarithm in the complex plane; if $z = re^{i\theta}$, where $r > 0$ and $-\pi/2 < \theta \le \pi/2$, $\ln(re^{i\theta}) = r + i\theta$. Using Taylor's remainder bound,

$$(6) \qquad |\ln(1+z) - z| \le 2|z|^2 \qquad \text{if } |z| \le 1/2.$$

Assumption (5) assures us that for all large enough n and $1 \le k \le n$, $|s_{kn}| \le 1/2$. We may write

$$(7) \qquad \prod_{k=1}^{n}(1+s_{kn}) = \exp\left\{\sum_{k=1}^{n} s_{kn} + \sum_{k=1}^{n}(\ln(1+s_{kn}) - s_{kn})\right\}.$$

Since, by (6),

$$|\sum_{k=1}^{n}(\ln(1+s_{kn}) - s_{kn})| \le 2\sum_{k=1}^{n}|s_{kn}|^2 \le 2\left(\max_{1\le k\le n}|s_{kn}|\right)\sum_{k=1}^{n}|s_{kn}|,$$

the assumptions made in (5) imply that this term tends to 0 as n increases. Thus, the limit as $n \to \infty$ in (7) equals e^{α}. $\diamond$

Proof of Theorem 4.5.1. Since the X_k are independent, the characteristic function of Z_n is

$$\phi_{Z_n}(\lambda) = \prod_{k=1}^{n} \phi_{X_k}\left(\frac{\lambda}{B_n}\right).$$

Define the remainder terms $R_k(\lambda) = \phi_{X_n}(\lambda) - \big(1 - (\lambda^2/2)\sigma_k^2\big)$. Then we rewrite

$$\phi_{Z_n}(\lambda) = \prod_{k=1}^{n}\left(1 - \frac{\sigma_k^2}{2B_n^2} + R_k(\frac{\lambda}{B_n})\right).$$

We shall prove,

$$(8) \qquad \lim_{n\to\infty} \sum_{k=1} |R_k(\frac{\lambda}{B_n})| = 0.$$

This will suffice to complete the proof, as we now show. Indeed, (8) implies

$$\sup_{n\geq 1}\sum_{k=1}^{n}|-\frac{\sigma_k^2}{2B_n^2}+R_k(\frac{\lambda}{B_n})|\leq \frac{\lambda^2}{2}+\sup_{n\geq 1}\sum_{k=1}^{n}|R_k(\frac{\lambda}{B_n})|<\infty,$$

$$\lim_{n\to\infty}\left(\sum_{k=1}^{n}-\frac{\sigma_k^2}{2B_n^2}+R_k(\frac{\lambda}{B_n})\right)=-\frac{\lambda^2}{2}+\lim_{n\to\infty}\sum_{k=1}^{n}R_k(\frac{\lambda}{B_n})=-\frac{\lambda^2}{2}.$$

Moreover, because of Lemma 4.5.2, (8) also implies

$$\max_{1\leq k\leq n}|-\frac{\sigma_k^2}{2B_n^2}+R_k(\frac{\lambda}{B_n})|\leq \max_{1\leq k\leq n}\frac{\sigma_k^2}{2B_n^2}+\sum_{k=1}^{n}|R_k(\frac{\lambda}{B_n})|\to 0 \ \text{ as } n\to\infty.$$

We have thus shown that the hypotheses of Lemma 4.5.4 are satisfied in the case when $s_{kn}=-\frac{\sigma_k^2}{2B_n^2}+R_k(\frac{\lambda}{B_n})$ and $\alpha=-\lambda^2/2$. By applying the lemma we find that $\lim_{n\to\infty}\phi_{Z_n}(\lambda)=e^{-\lambda^2/2}$, which is the characteristic function of the standard, unit normal. The continuity theorem, 4.1.3, now shows that the sequence $\{Z_n\}$ converges in distribution to a standard, unit normal.

It remains to show (8). For this, apply Lemma 4.5.3 with $\delta=\epsilon B_n$. Then, for every $\epsilon>0$,

$$\sum_{k=1}^{n}|R_k(\frac{\lambda}{B_n})|\leq \frac{\lambda^2}{B_n^2}\sum_{k=1}^{n}E\left[|X_k|^2\mathbf{1}\{|X_k|>\epsilon B_n\}\right]+\frac{|\lambda|^3}{6}\epsilon.$$

The Lindeberg condition says that as $n\to\infty$, the coefficient of λ^2 in the last expression tends to 0. Hence $\lim_{n\to\infty}\sum_{k=1}^{n}|R_k(\frac{\lambda}{B_n})|<|\lambda|^3\epsilon/2$. Taking $\epsilon\downarrow 0$ proves (8).
$\diamond$