Chapter 3

Central limit theorems

Probability theory around 1700 was basically of a combinatorial nature. The main monograph of the period was Abraham de Moivre's *The Doctrine of Chances; or, a Method for Calculating the Probabilities of Events in Play* from 1718, which solved a large number of combinatorial problems relating to games with cards or dice. As a typical example, de Moivre computed the probability of obtaining at least two aces in eight rolls with a dice.

In this type of problem the answers will inevitably involve binomial coefficients. The answer can be completely digestible as long as the problem size is small. But binomial coefficients have an extremely rapid growth with the problem size, and the available tables will soon be exhausted. It can be very difficult to get a feeling for the magnitude of a certain combination of binomial coefficients - does it correspond to a sizeable probability or is is for all practical purposes equal to zero?

For many problems the natural answer comes in the form of a sum of point-probabilities, and such problems become even less tractable. To get a feeling for a sum with many terms, we must be able to identify the major contributions, but if each term is a product of a huge factor (say, a binomial coefficient) and a tiny factor (say, a reasonable probability raised to a large power) it is very difficult to know which terms that are noteworthy and which terms that can be ignored.

De Moivre worked for many years (partly in collaboration with and partly in competition with Stirling) on how to find approximate expressions for sums of probabilities.

The main result was what we today know as Stirling's formula. We tend to think of this formula as an asymptotic expansion of the Gamma-function (see section 8.3 of MT, where this viewpoint is expanded), but the main motivation for de Moivre and Stirling was to control binomial coefficients. De Moivre established that if $X_1, X_2, \ldots$ are independent Bernoulli variables with

$$P(X_i = 1) = \frac{1}{2}, \quad P(X_i = 0) = \frac{1}{2},$$

and if we let

$$k_n(x) = \left\lfloor n/2 + x\sqrt{n/4} \right\rfloor$$

then it holds that

$$P\left(\sum_{i=1}^{n} X_i = k_n(x)\right) \approx \sqrt{\frac{2}{n\pi}}\, e^{-x^2/2} .$$

This claim (or at least a claim closely related to it) appears in the 1738-edition of *The Doctrine of Chances*, side by side with simular formulas for sums of asymmetric Bernoulli variables. After a bit of handwaving, this leads to the first version of what we today call a **central limit theorem (CLT)**: for any $a < b$ it holds that

$$P\left(a \leq \frac{\sum_{i=1}^{n} X_i - n/2}{\sqrt{n/4}} \leq b\right) \approx \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx . \tag{3.1}$$

In general, a CLT is a limit theorem with a normal distribution in the limit. The use of the word 'central' in this connection is originally due to the fact that it is a theorem about the central parts of the distribution (as opposed to the tail-behavior), but it could equally well be attributed to the paramount importance of these theorems - they are the cornerstones of applied probability and statistics.

De Moivre did probably not see anything metaphysical in (3.1). To him, it was just another approximation formula. He was probably the first man in history to consider the integral on the right hand side of (3.1), and he did not relate it to anything. The term 'normal distribution' is a much later development, and de Moivre did not even realize that there is a probability distribution involved - this observation is due to Laplace some 50 years later. To get a proper feeling for the difficulties facing de Moivre, please note that the symbols e and π were not in use in his days, and he had to represent these magnitudes using his own notation - which of course tended to shift over time. Furthermore, the modern argument leading to (3.1) is to consider the left hand side as a Riemann-sum for the integral on the right - but the foundations of integration theory was a mess at the time, and a clear grasp of the concept of a Riemann-sum was more than a century away.

Quite a lot of people tried to advance the ideas of de Moivre, but for a long time without any real progress: the approximation arguments seemed to work only for the binomial distribution. But in 1782 Laplace had a breakthrough. He was able to approximate sums of of variables that could assume **three** values (instead of the two values of a Bernoulli variable). The key to success was an extraordinary tricky substitution argument, where he translated everything to integrals involving the complex exponential function. His ideas are in fact quite similar to what we today call *characteristic functions*, an essential part of advanced probability theory, closely related to the concept of Fourier series. The big surprise was that he encountered the same approximating integral as in the binomial case. And here the metaphysical principle begin to take shape: it appears that the distribution of a sum of stochastic variables is almost unaffected by the distribution of the individual term!

Laplace succeeded in giving quite general proofs of CLTs for averages of independent, identically distributed variables by using his trick with the complex exponential. A few years later Gauss devised a theory of 'measurement errors' based on the normal distributions, and the normal distribution has ever since been a key ingredient in every course of statistics.

In the late 19th century, Chebyshev and his students (in particular Markov and Lyapounov) started contemplating on what really happens in a CLT. It turns out that there are some subtle differences between convergence of distribution functions (the type of convergence appearing in de Moivre's theorem) and convergence of characteristics functions. Chebyshev developed a theory for convergence of abstract probability measures to explore this. We shall retrace his steps in this chapter, even though we shall not introduce characteristic functions at all.

3.1 Convergence of measures

Let $C_b(\mathbb{R}^k)$ be the class functions $\mathbb{R}^k \to \mathbb{R}$ which are continuous and bounded. Clearly every $C_b(\mathbb{R}^k)$-function is Borel measurable, and as they are bounded, they are integrable with respect to any probability measure on $(\mathbb{R}^k, \mathbb{B}_k)$. For a function $f \in C_b(\mathbb{R}^k)$ we will consider the **uniform norm**

$$\|f\| = \sup_{x \in \mathbb{R}^k} |f(x)| \, .$$

Definition 3.1 *A sequence of probability measures $\nu_1, \nu_2, \dots$ on $(\mathbb{R}^k, \mathbb{B}_k)$ is said to* **converge weakly** *to a limit probability measure ν if*

$$\lim_{n\to\infty} \int f(x)\, d\nu_n(x) = \int f(x)\, d\nu(x) \quad \textit{for every } f \in C_b(\mathbb{R}^k)\,. \tag{3.2}$$

We will write $\nu_n \overset{wk}{\to} \nu$ for $n \to \infty$ to denote weak convergence.

Note that a limit for weak convergence is necessarily unique: if $\nu_n \overset{\text{wk}}{\to} \nu$ and $\nu_n \overset{\text{wk}}{\to} \lambda$ for $n \to \infty$, then it holds that $\int f\, d\nu = \int f\, d\lambda$ for every function $f \in C_b(\mathbb{R}^k)$. And this implies that $\nu = \lambda$, according to problem 1.11.

Example 3.2 Let $x_1, x_2, \dots$ and x be points in $\mathbb{R}^k$, and consider the corresponding one-point measures $\epsilon_{x_1}, \epsilon_{x_n} \dots$ and ϵ_x. If $x_n \to x$ for $n \to \infty$, then it is simple to see that $\epsilon_{x_n} \overset{\text{wk}}{\to} \epsilon_x$. For if f is a continuous bounded function, then we have that

$$\int f\, d\epsilon_{x_n} = f(x_n) \to f(x) = \int f\, d\epsilon_x \qquad \text{for } n \to \infty\,.$$

In fact, the opposite implication is also true. Suppose that $\epsilon_{x_n} \overset{\text{wk}}{\to} \epsilon_x$. Let $\delta > 0$ be given, and construct a continuous function f satisfying that

$$1_{\{x\}} \le f \le 1_{B(x,\delta)}\,.$$

For instance, f could be the composition of $y \mapsto |y - x|$ with the real-valued piecewise affine map $z \mapsto (1 - z/\delta)^+$. Due to the weak convergence, we see that

$$\int f\, d\epsilon_{x_n} \to \int f\, d\epsilon_x = f(x) = 1 \qquad \text{for } n \to \infty\,,$$

and as $\int f\, d\epsilon_{x_n} = f(x_n) \le 1_{(B(x,\delta)}(x_n)$ it holds that $|x_n - x| < \delta$ for n large enough.

∘

Example 3.3 On the measurable space $(\mathbb{R}, \mathbb{B})$, let ν_n be the empirical measure in the points $0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}$. If f is a continuous function, we see that

$$\int f\, d\nu_n = \frac{1}{n} \sum_{i=1}^{n} f\left(\frac{i-1}{n}\right).$$

Consider the piecewise constant functions

$$f_n(x) = \begin{cases} f(0) & \text{for } x \in \left[0, \frac{1}{n}\right) \\ f\left(\frac{1}{n}\right) & \text{for } x \in \left[\frac{1}{n}, \frac{2}{n}\right) \\ \vdots & \vdots \\ f\left(\frac{n-1}{n}\right) & \text{for } \ x \in \left[\frac{n-1}{n}, 1\right) \end{cases}$$

Observe that

$$\int_0^1 f_n(x)\, dm(x) = \frac{1}{n} \sum_{i=1}^{n} f\left(\frac{i-1}{n}\right),$$

so we actually have that $\int f \, d\nu_n = \int_0^1 f_n(x)\, dx$. The continuity of f ensures that $f_n(x) \to f(x)$ for every $x \in [0, 1)$, and this convergence is dominated by $\|f\|$, so the dominated convergence theorem implies that

$$\int_0^1 f_n(x)\, dx \to \int_0^1 f(x)\, dx \qquad \text{for } \ n \to \infty,$$

And translated into ν_n-integrals, this is exactly the statement that ν_n converges weakly to the restriction of the Lebesgue measure to the unit interval.

∘

An interesting observation can be made in example 3.3. The converging measures ν_n are discrete, while the limit measure is continuous. It is equally easy to produce examples where sequences of continuous measures converge weakly to a discrete measure. So within the theory of weak convergence, the usual dichotomy between continuous and discrete probability measures is less pronounced than in other areas of probability.

Lemma 3.4 (Scheffé) *Let $\nu_1, \nu_2, \dots$ and ν be probability measures on $(\mathbb{R}^k, \mathbb{B}_k)$. Assume that for some choice of basic measure μ it holds that $\nu_n = f_n \cdot \mu$ for every n and $\nu = f \cdot \mu$ for suitable density functions f_n and f. If*

$$f_n(x) \to f(x) \qquad \textit{for } n \to \infty \qquad \mu\textit{-almost surely}$$

then it holds that $\nu_n \xrightarrow{wk} \nu$.

Proof: We will prove that $f_n \to f$ in $\mathcal{L}_1(\mu)$. As any function $g \in C_b(\mathbb{R}^k)$ is bounded by $\|g\|$, the $\mathcal{L}_1$-convergence implies that

$$\left|\int g\,d\nu_n - \int g\,d\nu\right| = \left|\int g\,f_n - g\,f\,d\mu\right| \leq \int \|g\|\,|f_n - f|\,d\mu \to 0\,,$$

and so the desired weak convergence follows.

As we discussed in section 1.1, it is not true in general that pointwise convergence implies $\mathcal{L}_1$-convergence - some kind of domination is usually required to make the argument work. But when we know that the functions involved are probability densities, there is a trick with positive and negative parts that allows us to circumvent the need for domination. We can observe that for every x it holds that

$$(f_n - f)^-(x) = \max\Big\{-(f_n(x) - f(x)), 0\Big\} = \max\{f(x) - f_n(x), 0\} \leq f(x)\,,$$

using that $f_n(x) \geq 0$ and that $f(x) \geq 0$. As $x \mapsto \max\{-x, 0\}$ is a continuous function, we see that $(f_n - f)^-$ converges pointwise to zero μ-almost surely. As we observed, this convergence is dominated by the integrable function f, so by Lebesgues theorem on dominated convergence, we conclude that

$$\int (f_n - f)^-\,d\mu \to 0 \qquad \text{for } n \to \infty\,.$$

The trick is now to observe that since ν_n and ν have total mass 1, it holds that

$$0 = \int f_n\,d\mu - \int f\,d\mu = \int f_n - f\,d\mu = \int (f_n - f)^+\,d\mu - \int (f_n - f)^-\,d\mu\,.$$

Hence

$$\int |f_n - f|\,d\mu = \int (f_n - f)^+\,d\mu + \int (f_n - f)^-\,d\mu = 2\int (f_n - f)^-\,d\mu \to 0\,,$$

exactly as desired.

□

Example 3.5 Let $\nu_n = \mathcal{N}(\xi_n, {\sigma_n}^2)$ be a sequence of Gaussian distributions on $\mathbb{R}$. If

$$\xi_n \to \xi\,, \quad \sigma_n \to \sigma \qquad \text{for } n \to \infty\,,$$

for some $\sigma > 0$, then the ν_n-measures will converge weakly to $\mathcal{N}(\xi, \sigma^2)$. This follows from Scheffé's lemma, since

$$\frac{1}{\sigma_n}\phi\left(\frac{x - \xi_n}{\sigma_n}\right) \to \frac{1}{\sigma}\phi\left(\frac{x - \xi}{\sigma}\right) \qquad \text{for } n \to \infty\,,$$

where ϕ is the density of the standard normal distribution. The functions in this relation are precisely the densities of ν_n and of $\mathcal{N}(\xi, \sigma^2)$ with respect to Lebesgue measure.

$\circ$

Weak convergence is a notion that makes sense for a sequence of probability measures. But in applications of the theory, it is usually more convenient to express the results in terms of stochastic variables. This leads to the following terminology:

Definition 3.6 *A sequence of stochastic variables $X_1, X_2, \dots,$ defined on a common background space $(\Omega, \mathbb{F}, P)$ and with values in $\mathbb{R}^k$, is said to* **converge in distribution** *to a limit variable X, symbolically written as $X_n \stackrel{\mathcal{D}}{\to} X$, if the corresponding image measures $X_1(P), X_2(P), \dots$ converge weakly to the image measure $X(P)$.*

Mathematically speaking, it is not necessary that the variables are defined on a common background space, as the definition only involves the image measures. But it is difficult to get a grip on the scenario if the background space is not considered fixed. By use of the abstract change-of-variable formula, the condition for convergence in distribution can conveniently be formulated as follows: it holds that $X_n \stackrel{\mathcal{D}}{\to} X$ if and only if

$$\int f(X_n)\, dP \to \int f(X)\, dP \quad \text{for every} \;\; f \in C_b(\mathbb{R}^k)\,.$$

We know that a limit measure for weak convergence is necessarily unique. But for convergence in distribution, the situation is somewhat more intricate. The limit variable X only appears in the definition of convergence in distribution through its distribution $X(P)$. Hence all stochastic variables with the same distribution are equally applicable. As an illustration of the surprises hidden in this observation, note that when $X_n \stackrel{\mathcal{D}}{\to} X$ where X is $\mathcal{N}(0,1)$-distributed, then it also true that $X_n \stackrel{\mathcal{D}}{\to} -X$, since $-X$ is $\mathcal{N}(0,1)$-distributed as well.

Theorem 3.7 (Continuous mapping theorem) *Let $X_1, X_2, \dots$ and X be stochastic variables with values in $\mathbb{R}^k$, and let $h : \mathbb{R}^k \to \mathbb{R}^m$ be continuous. If $X_n \stackrel{\mathcal{D}}{\to} X$, then it holds that $h(X_n) \stackrel{\mathcal{D}}{\to} h(X)$.*

PROOF: Note that if g is a $C_b(\mathbb{R}^m)$-function, then $g \circ h$ is a $C_b(\mathbb{R}^k)$-function. Hence

$$\int g\big(h(X_n)\big)\,dP = \int g \circ h(X_n)\,dP \to \int g \circ h(X)\,dP = \int g\big(h(X)\big)\,dP\,,$$

exactly as desired.

□

As a trivial application of the continuous mapping theorem, observe that if $X_n \xrightarrow{\mathcal{D}} X$ in $\mathbb{R}^k$, then for any fixed vector $v \in \mathbb{R}^k$ it holds that

$$v^T X_n \quad \xrightarrow{\mathcal{D}} \quad v^T X\,,$$

where weak convergence in $\mathbb{R}^k$ is translated into weak convergence in $\mathbb{R}$. It turns out that the opposite implication is true:

Theorem 3.8 (Cramér-Wold) *Let $X_1, X_2, \ldots$ and X be stochastic variables with values in $\mathbb{R}^k$. If*

$$v^T X_n \quad \xrightarrow{\mathcal{D}} \quad v^T X\,, \tag{3.3}$$

for any fixed vector $v \in \mathbb{R}^k$, then it holds that $X_n \xrightarrow{\mathcal{D}} X$.

We will not prove the Cramér-Wold theorem, even though we will rely heavily on it in latter parts of this chapter. It is very difficult (though not impossible) to prove the Cramér-Wold theorem without taking recourse to characteristic functions or similar Fourier-based techniques[1].

3.2 Distribution functions and weak convergence

It would be difficult to defend the definition of weak convergence of probability measures on the ground that it is an intuitive and natural concept. Certain integrals are required to converge, but what information does such a statement really contain about the measures themselves? On $(\mathbb{R}, \mathbb{B})$ one way to get a partial answer to this question is to examine what happens with the distribution functions.

[1] One of the classics of probability theory, Billingsley's book *Probability and measure*, states that Fourier-based techniques are seemingly necessary for the proof of the Cramér-Wold theorem. There are, however, direct proofs available. One such proof is given in Pollard's book *A user's guide to measure theoretic probability*.

Lemma 3.9 *Let $\nu_1, \nu_2, \dots$ and ν be probability measures on $(\mathbb{R}, \mathbb{B})$, and let $F_1, F_2, \dots$ and F be the corresponding distribution functions. If $\nu_n \stackrel{wk}{\to} \nu$ then it holds that*

$$F_n(x_0) \to F(x_0) \qquad \text{for } n \to \infty\,,$$

whenever x_0 is a point of continuity for F.

Proof: For given $\varepsilon > 0$ we can choose continuous bounded functions f and g, satisfying that

$$1_{(-\infty, x_0-\varepsilon]} \leq f \leq 1_{(-\infty, x_0]} \qquad \text{and} \qquad 1_{(-\infty, x_0]} \leq g \leq 1_{(-\infty, x_0+\varepsilon]}\,.$$

We can for instance choose f and g to be piecewise affine. We observe that

$$F(x_0 - \varepsilon) \leq \int f\, d\nu = \lim_{n\to\infty} \int f\, d\nu_n \leq \liminf F_n(x_0)\,.$$

Similarly,

$$F(x_0 + \varepsilon) \geq \int g\, d\nu = \lim_{n\to\infty} \int g\, d\nu_n \geq \limsup F_n(x_0)\,.$$

We can collect these inequalities and obtain

$$F(x_0 - \varepsilon) \leq \liminf F_n(x_0) \leq \limsup F_n(x_0) \leq F(x_0 + \varepsilon)\,.$$

Letting $\varepsilon \to 0$, we can exploit the assumption that x_0 is a point of continuity for F. We see that $F(x_0 - \varepsilon) \to F(x_0)$ and that $F(x_0 + \varepsilon) \to F(x_0)$ as $\varepsilon \to 0$. So passing to the limit we obtain the inequalities

$$F(x_0) \leq \liminf F_n(x_0) \leq \limsup F_n(x_0) \leq F(x_0)\,,$$

and this establishes that $F_n(x_0) \to F(x_0)$ for $n \to \infty$.

□

At a first encounter it may come as a surprise that lemma 3.9 only deals with the behavior of the distribution functions in points of continuity of the limit probability measure. After all, there are at most countably many points of discontinuity, and it would seem plausible that some sort of approximation-scheme using points of continuity would extend the result to pointwise convergence in every point. But this is not so - weak convergence does not allow us to make definite statements about points of discontinuity for the limit measure!

Example 3.10 Consider on $(\mathbb{B}, \mathbb{B})$ the sequence ν_n of one-point measures in the points $\frac{1}{n}$, and let ν be the one-point measure in 0. According to example 3.2 it holds that $\nu_n \overset{\text{wk}}{\to} \nu$. But looking at the value of the distribution functions in 0, we see that

$$F_n(0) = 0 \quad \text{for every } n\,, \qquad F(0) = 1\,.$$

So there is no chance that the distribution functions should converge in **every** point.

Curiously, if we repeat the exercise using the one-point measures in $-1, -\frac{1}{2}, -\frac{1}{3}, \ldots$, we will in fact obtain pointwise convergence in every point of the distribution functions, even though this situation seems extremely similar to the previous one.

So the problem with lack of pointwise convergence in every point is not just associated to discontinuity in the limit distribution. There is also an issue of a subtle left-right asymmetry in the definition of the distribution function.

$\circ$

It turns out that weak convergence on $\mathbb{R}$ can indeed be rephrased in terms of convergence properties of distribution functions. But as we have illustrated in example 3.10, some care is needed to pinpoint the essential convergence property. It has to do with convergence in 'many points' rather than in all points:

Theorem 3.11 *Let $\nu, \nu_1, \nu_2, \ldots$ be probability measures on $(\mathbb{R}, \mathbb{B})$ and let $F, F_1, F_2, \ldots$ be the corresponding distribution functions. It holds that $\nu_n \overset{wk}{\to} \nu$ if and only if there is a dense subset $A \subset \mathbb{R}$ such that*

$$F_n(x) \to F(x) \qquad \textit{for } n \to \infty \quad \textit{for every } \; x \in A\,. \tag{3.4}$$

Proof: If $\nu_n \overset{\text{wk}}{\to} \nu$, we have already seen in lemma 3.9 that (3.4) holds for every $x \in \mathbb{R} \setminus \Delta(\nu)$,where $\Delta(\nu)$ is the the set of jump points for ν. As the set of jump points is countable, the complement $\mathbb{R} \setminus \Delta(\nu)$ will necessarily be a dense subset of $\mathbb{R}$.

So let us conversely assume that there is a dense subset $A \subset \mathbb{R}$ where (3.4) holds. Let f be a $C_b(\mathbb{R})$-function, and let $\varepsilon > 0$ be given. As $F(x)$ converges to 0 and 1 for $x \to \pm\infty$, we can find a and b such that

$$F(a) < \frac{\varepsilon}{\|f\| + 1} \qquad \text{and} \qquad F(b) > 1 - \frac{\varepsilon}{\|f\| + 1}\,.$$

Without loss of generality we can assume that $a, b \in A$ since A is dense. It trivially holds that $-\|f\| \leq f(x) \leq \|f\|$ for every x, and thus in particular

$$-\|f\|\, \nu\big((-\infty, a]\big) \leq \int_{(-\infty,a]} f\, d\nu \leq \|f\|\, \nu\big((-\infty, a]\big)\,.$$

A similar estimate can be obtained for the upper tail. Hence the choice of a and b ensures that

$$-\varepsilon < \int_{(-\infty,a]} f\, d\nu < \varepsilon \qquad \text{and} \qquad -\varepsilon < \int_{(b,\infty)} f\, d\nu < \varepsilon\,.$$

Using that $a, b \in A$, we can from (3.4) choose $N \in \mathbb{N}$ such that

$$F_n(a) < \frac{\varepsilon}{\|f\| + 1} \qquad \text{and} \qquad F_n(b) > 1 - \frac{\varepsilon}{\|f\| + 1} \qquad \text{for } n \geq N\,. \tag{3.5}$$

If we repeat the arguments above, we obtain that

$$-\varepsilon < \int_{(-\infty,a]} f\, d\nu_n < \varepsilon \qquad \text{and} \qquad -\varepsilon < \int_{(b,\infty)} f\, d\nu_n < \varepsilon \qquad \text{for } n \geq N\,.$$

Collecting the informations, we conclude that

$$\left| \int_{(-\infty,a]} f\, d\nu_n - \int_{(-\infty,a]} f\, d\nu \right| < 2\varepsilon\,, \quad \left| \int_{(b,\infty)} f\, d\nu_n - \int_{(b,\infty)} f\, d\nu \right| < 2\varepsilon\,,$$

for every $n \geq N$. We will now - possibly after replacing N with a larger value - establish that

$$\left| \int_{(a,b]} f\, d\nu_n - \int_{(a,b]} f\, d\nu \right| < 3\varepsilon \qquad \text{for } n \geq N\,. \tag{3.6}$$

In that case it follows that

$$\begin{aligned} \left| \int f\, d\nu_n - \int f\, d\nu \right| &\leq \left| \int_{(-\infty,a]} f\, d\nu_n - \int_{(-\infty,a]} f\, d\nu \right| + \left| \int_{(a,b]} f\, d\nu_n - \int_{(a,b]} f\, d\nu \right| \\ &\quad + \left| \int_{(b,\infty)} f\, d\nu_n - \int_{(b,\infty)} f\, d\nu \right| \\ &< 2\varepsilon + 3\varepsilon + 2\varepsilon\,, \end{aligned}$$

for $n \geq N$. From this estimate is is clear how to establish weak convergence.

To show (3.6) we note that the continuous function f is uniformly continuous on the compact interval $[a, b]$. Hence we can find $\delta > 0$ such that

$$x, y \in [a, b]\,,\ |x - y| < \delta \qquad \Rightarrow \qquad |f(x) - f(y)| < \varepsilon\,.$$

Choose a partition of $[a, b]$,

$$a = y_0 < y_1 < \ldots < y_{M-1} < y_M = b\,,$$

such that $|y_n - y_{n-1}| < \delta$ for every n, and such that $y_n \in A$ for every n. We can for instance start out with an equidistant grid between a and b where the distance between neighboring points is $\rho < \delta/2$, and then choose an A-point in each of the grid intervals.

From this partition we construct a function g as follows:

$$g = \sum_{i=1}^{M} m_i \, 1_{(y_{i-1}, y_i]} \;\;,$$

where

$$m_i = \min\Big\{ f(y) \mid y \in [y_{i-1}, y_i] \Big\} \qquad \text{for } \; i = 1, \ldots, M\,.$$

The essential point is that we have constructed g such that

$$g(x) \leq f(x) < g(x) + \varepsilon \quad \text{for } x \in (a, b]\,.$$

Furthermore, we notice that

$$\int g\,d\nu_n = \sum_{i=1}^{M} m_i\Big(F_n(y_i) - F_n(y_{i-1})\Big) \to \sum_{i=1}^{M} m_i\Big(F(y_i) - F(y_{i-1})\Big) = \int g\,d\nu$$

for $n \to \infty$. Hence

$$\begin{aligned}
\left| \int_{(a,b]} f\,d\nu_n - \int_{(a,b]} f\,d\nu \right| &\leq \left| \int_{(a,b]} f\,d\nu_n - \int_{(a,b]} g\,d\nu_n \right| + \left| \int_{(a,b]} g\,d\nu_n - \int_{(a,b]} g\,d\nu \right| \\
&\quad + \left| \int_{(a,b]} g\,d\nu - \int_{(a,b]} f\,d\nu \right| \\
&< \varepsilon + \left| \int_{(a,b]} g\,d\nu_n - \int_{(a,b]} g\,d\nu \right| + \varepsilon
\end{aligned}$$

which again is smaller that 3ε when n is large enough.

□

It should be noted that in most applications of the weak convergence concept, the limit measure will actually be continuous - in the most common cases the limit is a Gaussian distribution or a χ^2-distribution. In these cases, weak convergence is equivalent to pointwise convergence of the distribution functions - indeed, it can be shown to be equivalent to *uniform* convergence of the distribution functions, though we will not need this here.

3.3 Tightness

The purpose of this section is to show that when we attempt to establish weak convergence, we do not have to check convergence of the integrals as in (3.2) for every single C_b-function. There is a considerable number of subclasses of C_b that will suffice.

We will need to work with an increasing sequence $K_1 \subset K_2 \subset \ldots$ of compact subsets of $\mathbb{R}^k$. To have a specific sequence in mind, we will use

$$K_m = [-m, m] \times \ldots \times [-m, m] \qquad \text{for } \; m \in \mathbb{N}\,,$$

but many other choices would be be equally applicable. One of the useful properties of this specific sequence is that is easy to see how to produce a continuous function f with the property

$$1_{K_m}(x) \leq f(x) \leq 1_{K_{m+1}}(x) \qquad \text{for } x \in \mathbb{R}^k\,. \tag{3.7}$$

In one dimension we can use a piecewise affine function (there is a suggestive picture p. 390 in MT). If we denote this one-dimensional solution by ϕ, the solution in dimension k can take the form $f(x_1, \ldots, x_k) = \phi(x_1) \cdot \ldots \cdot \phi(x_k)$. Note that a function satisfying (3.7) will automatically be bounded and have compact support (see section 1.7).

Definition 3.12 *A family $(\mu_i)_{i \in I}$ of probability measures on $(\mathbb{R}^k, \mathbb{B}_k)$ is* **tight** *if for any $\varepsilon > 0$ there is $m \in \mathbb{N}$ such that*

$$\mu_i\Big(K_m\Big) \geq 1 - \varepsilon \qquad \textit{for every } i \in I\,. \tag{3.8}$$

Note that a family consisting of a single probability measure is always tight. Indeed, any finite family of probability measures is tight. As is the union of two tight families.

Theorem 3.13 *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}^k, \mathbb{B}_k)$. If*

$$\int f\, d\mu_n \to \int f\, d\mu \qquad \textit{for every } \; f \in C_c(\mathbb{R}^k)\,,$$

then the family $\{\mu, \mu_1, \ldots\}$ is tight.

Proof: Pick $m_0 \in \mathbb{N}$ such that the limit measure μ satisfies that

$$\mu\Big(K_{m_0}\Big) \geq 1 - \frac{\varepsilon}{2}\,. \tag{3.9}$$

Find a C_b-function f such hat

$$1_{K_{m_0}}(x) \leq f(x) \leq 1_{K_{m_0+1}}(x) \qquad \text{for } x \in \mathbb{R}^k\,.$$

Note that

$$\begin{aligned}\liminf \mu_n(K_{m_0+1}) &\geq \liminf \int f\, d\mu_n = \int f\, d\mu \\ &\geq \mu(K_{m_0}) \\ &\geq 1 - \frac{\varepsilon}{2}\end{aligned}$$

In particular there is an N such that

$$\mu_n(K_{m_0+1}) \geq 1 - \varepsilon \qquad \text{for } n > N\,.$$

But as the finite family $\{\mu_1, \ldots, \mu_N\}$ is tight, we can easily find m_1 such that

$$\mu_n(K_{m_1}) \geq 1 - \varepsilon \qquad \text{for } n = 1, \ldots, N\,.$$

Taking $m = \max\{m_0 + 1, m_1\}$ we see that (3.8) is satisfied.

□

It follows from theorem 3.13 that any weakly convergent sequence of probability measure is tight. In the opposite direction it can be shown that any tight sequence has a weakly convergent subsequence. This is known as Helly's theorem. This theorem identifies the compact subsets of the space of probability measures on $\mathbb{R}^k$ (in the topology of weak convergence) as the tight subsets.

Theorem 3.14 *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}^k, \mathbb{B}_k)$. If*

$$\int f\, d\mu_n \to \int f\, d\mu \qquad \textit{for every } f \in C_c(\mathbb{R}^k)\,, \tag{3.10}$$

then it holds that $\mu_n \overset{wk}{\to} mu$.

PROOF: Consider a $C_b(\mathbb{R}^k)$-function f and $\varepsilon > 0$. Using theorem 3.13 we know that $\mu_1, \mu_2, \ldots$ is a tight sequence, hence we can find m such that

$$\mu_n(K_m) \geq 1 - \varepsilon \qquad \text{for every } n \in \mathbb{N}$$

and also for the limit measure μ. Construct a continuous function ϕ satisfying

$$1_{K_m}(x) \leq \phi(x) \leq 1_{K_{m+1}}(x) \qquad \text{for } x \in \mathbb{R}^k\,.$$

As noted earlier, such a function will automatically be $C_c(\mathbb{R}^k)$. We see that

$$\int f\,d\mu_n = \int f\phi\,d\mu_n + \int f(1-\phi)\,d\mu_n\,.$$

Observe that $f\phi$ is a C_c-function, so here (3.10) can be applied. Observe also that

$$\left|\int f(1-\phi)\,d\mu_n\right| \leq \|f\|\,\varepsilon\,,$$

and similarly for the integral of the limit measure. If we pick n_0 so large that

$$\left|\int f\phi\,d\mu_n - \int f\phi\,d\mu\right| < \varepsilon \qquad \text{for } n \geq n_0\,,$$

it follows that

$$\left|\int f d\mu_n - \int f\,d\mu\right| \leq 2\|f\|\,\varepsilon + \left|\int f\phi\,d\mu_n - \int f\phi\,d\mu\right| < (2\|f\|+1)\varepsilon\,.$$

And so we see that $\int f\,d\mu_n$ converges to $\int f\,d\mu$.

□

Corollary 3.15 *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}^k, \mathbb{B}_k)$. If*

$$\int f\,d\mu_n \to \int f\,d\mu \qquad \textit{for every } f \in C_c^\infty(\mathbb{R}^k)\,, \tag{3.11}$$

then it holds that $\mu_n \overset{wk}{\to} mu$.

PROOF: It is straightforward to show that convergence of C_c^∞-integrals imply convergence of any C_c-integral by applying theorem 1.25.

□

We shall use corollary 3.15 on numerous occasions. Not necessarily on C_c^∞, but on larger classes that come into play more gracefully. It follows for instance from corollary 3.15 that to establish weak convergence, we only need to establish convergence of integrals as in (3.2) for C_b-functions that are uniformly continuous.

3.4 Weak convergence and convergence in probability

Recall that a sequence of stochastic variables $X_1, X_2, \dots$ with values in $\mathbb{R}^k$ converges in probability to a point $x \in \mathbb{R}^k$, written $X_n \xrightarrow{P} x$, if

$$P(|X_n - x| > \varepsilon) \to 0 \quad \text{for} \ \ n \to \infty \,.$$

for every $\varepsilon > 0$. Convergence in probability to a point can be formulated in terms of convergence in distribution - or weak convergence, if you will.

Lemma 3.16 *Let $X, X_1, X_2, \dots$ be stochastic variables with values in $\mathbb{R}^k$, and assume that $P(X = x_0) = 1$. It holds that*

$$X_n \xrightarrow{P} x_0 \quad \Leftrightarrow \quad X_n \xrightarrow{\mathcal{D}} X \,.$$

Proof: Assume that $X_n \xrightarrow{\mathcal{D}} X$. Take $\varepsilon > 0$, and construct a C_b-function f satisfying

$$1_{\{x_0\}}(x) \le f(x) \le 1_{B(x_0,\varepsilon)}(x) \qquad \text{for } x \in \mathbb{R}^k \,,$$

for instance by applying the idea from the introduction of section 3.3 on products of piecewise affine functions of a single coordinate. As $1 - f$ is a C_b-function, we see that

$$P(|X_n - x_0| > \varepsilon) = \int 1_{B(x_0,\varepsilon)^c}(X_n)\, dP \le \int \Big(1 - f(X_n)\Big)\, dP \to \int \Big(1 - f(X)\Big)\, dP = 0 \,.$$

And we conclude that $X_n \xrightarrow{P} x_0$.

For the opposite implication, assume that $X_n \xrightarrow{P} x_0$. Let f be a $C_b(\mathbb{R}^k)$-function, and let $\varepsilon > 0$ be given. As f is continuous in x_0, there is a $\delta > 0$ such that

$$|x - x_0| < \delta \quad \Rightarrow \quad |f(x) - f(x_0)| < \varepsilon \,.$$

Hence

$$\begin{aligned}
\left|\int f(X_n)\, dP - \int f(X)\, dP\right| &= \left|\int f(X_n) - f(x_0)\, dP\right| \\
&\le \int_{(|X_n - x_0| < \delta)} |f(X_n) - f(x_0)|\, dP + \int_{(|X_n - x_0| \ge \delta)} |f(X_n) - f(x_0)|\, dP \\
&\le \varepsilon + 2\, \|f\|\, P\Big(|X_n - x_0| \ge \delta\Big) \,.
\end{aligned}$$

If N is chosen large enough, it holds that

$$\left|\int f(X_n)\,dP - \int f(X)\,dP\right| \leq 2\varepsilon \qquad \text{for } n \geq N\,.$$

As f and ε were arbitrary, it follows that $X_n \xrightarrow{\mathcal{D}} X$.

□

If two sequences of stochastic variables both converge in distribution,

$$X_n \xrightarrow{\mathcal{D}} X\,, \quad Y_n \xrightarrow{\mathcal{D}} Y\,,$$

it is generally **not** true that the bundle (X_n, Y_n) is converging in distribution to the 'obvious' limit (X, Y) - in fact, the bundle may not converge at all. Convergence in distribution of (X_n, Y_n) is a statement about the sequence of joint distributions, and there is not sufficient information in the behavior of the marginal distributions to answer such a question. It is for instance quite easy to construct examples where all the X_n's have the same distribution (and hence they are trivially converging in distribution) and where all the Y_n's have the same distribution, but where the joint distributions are non-fixed and not necessarily converging, see problem 3.3.

But under additional assumptions it is possible to conclude that the bundle (X_n, Y_n) is converging in distribution. A common situation where such a conclusion can be obtained relates to the special form of convergence in distribution from lemma 3.16:

Lemma 3.17 *Let $X, X_1, X_2, \ldots$ be stochastic variables with values in $\mathbb{R}^k$, let $Y_1, Y_2, \ldots$ be stochastic variables with values in $\mathbb{R}^m$, and let y be a vector in $\mathbb{R}^m$. If*

$$X_n \xrightarrow{\mathcal{D}} X\,, \quad Y_n \xrightarrow{P} y$$

then the bundle (X_n, Y_n) in $\mathbb{R}^{n+m}$ will converge in distribution to the limit variable (X, y).

Proof: Let $f \in C_b(\mathbb{R}^{k+m})$, and assume that f is uniformly continuous. For a given $\varepsilon > 0$ we can find $\delta > 0$ such that

$$|(x,z) - (x',z')| < \delta \quad \Rightarrow \quad |f(x,z) - f(x',z')| < \varepsilon\,.$$

We shall only apply this to two points with the same first coordinate. So essentially the formula we shall rely on is

$$|z - z'| < \delta \quad \Rightarrow \quad |f(x,z) - f(x,z')| < \varepsilon \qquad \text{for every } x, z, z' \,.$$

We have to estimate

$$\left| \int f(X_n, Y_n)\, dP - \int f(X, y)\, dP \right| \,. \tag{3.12}$$

A simple upper bound is

$$\left| \int f(X_n, Y_n) - f(X_n, y)\, dP \right| + \left| \int f(X_n, y) - f(X, y)\, dP \right|$$

The first integral can be estimated using the uniform continuity of f,

$$\left| \int f(X_n, Y_n) - f(X_n, y)\, dP \right| \leq \varepsilon P\Big(|Y_n - y| < \delta \Big) + 2\, \| f \|\, P\Big(|Y_n - y| \geq \delta \Big)$$

which is smaller than 2ε when n is large enough. Using theorem 3.7 on the inclusion map $x \mapsto (x, y)$, we see that

$$(X_n, y) \xrightarrow{\mathcal{D}} (X, y) \,. \tag{3.13}$$

So it follows that

$$\left| \int f(X_n, y) - f(X, y)\, dP \right| < \varepsilon \,,$$

when n is large enough. And consequently it follows that (3.12) is smaller than 3ε when n is large enough.

The remark following corollary 3.15 shows that it is sufficient to check (3.2) for uniformly continuous bounded functions to conclude weak convergence. So it does actually follow that (X_n, Y_n) converges to (X, y) in distribution.

□

When this lemma is combined with theorem 3.7, it has a wealth of consequences and makes convergence in distribution very flexible to work with. We can for instance note the following triviality:

Corollary 3.18 (Slutsky) *Let $X, X_1, X_2, \ldots,\ Y_1, Y_2, \ldots$ be stochastic variable with values in $\mathbb{R}^k$. If*

$$X_n \xrightarrow{\mathcal{D}} X \,, \quad Y_n \xrightarrow{P} 0$$

then it holds that

$$X_n + Y_n \xrightarrow{\mathcal{D}} X \,.$$

PROOF: It follows from lemma 3.17 that (X_n, Y_n) converges in distribution to $(X, 0)$. Now apply the continuous mapping theorem to this, using the continuous map $(x, y) \mapsto x + y$.

$\square$

Corollary 3.19 *Let $X_1, X_2, \ldots$ and X be stochastic variables with values in $\mathbb{R}^k$ and let $Y_1, Y_2, \ldots$ be real-valued stochastic variables. If*

$$X_n \xrightarrow{\mathcal{D}} X\,, \quad Y_n \xrightarrow{P} 1$$

then it holds that

$$Y_n X_n \xrightarrow{\mathcal{D}} X\,.$$

PROOF: It follows from lemma 3.17 hat (X_n, Y_n) converges in distribution to $(X, 1)$. Now apply the continuous mapping theorem to this, using the continuous map $(x, y) \mapsto y\,x$.

$\square$

3.5 Asymptotic normality

A certain form of weak convergence to a normal distributed limit is so common that that is has been given a name of its own:

Definition 3.20 *Let $X_1, X_2, \ldots$ be stochastic variables, defined on a background space $(\Omega, \mathbb{F}, P)$ with values in $(\mathbb{R}^k, \mathbb{B}_k)$, let $\xi \in \mathbb{R}^k$ be a vector and let Σ be a positive semidefinite $k \times k$ matrix.*

We say that X_n has an **asymptotic normal distribution** *with parameters $\left(\xi, \frac{1}{n}\Sigma\right)$, written*

$$X_n \overset{as}{\sim} \mathcal{N}\left(\xi, \frac{1}{n}\Sigma\right),$$

if

$$\sqrt{n}\,(X_n - \xi) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}(0, \Sigma) \qquad \text{for } n \to \infty.$$

The central limit theorem of de Moivre, given in formula (3.1) can be rephrased in this language as

$$\frac{1}{n}\sum_{i=1}^{n} X_i \quad \overset{\mathcal{D}}{\to} \quad \mathcal{N}\left(\frac{1}{2}, \frac{1}{4\,n}\right) \qquad \text{for } n \to \infty.$$

Many variants of the central limit theorem are best formulated using the concept of asymptotic normality.

Lemma 3.21 *Let $X_1, X_2, \ldots$ and X be stochastic variables with values in $\mathbb{R}^k$, and assume that $\sqrt{n}\,X_n \overset{\mathcal{D}}{\to} X$. It follows that $X_n \overset{P}{\to} 0$.*

Proof: As $\sqrt{n}\,X_n$ converges in distribution, the corresponding image measures form a tight family of probability measures on $\mathbb{R}^k$. Hence, for any given $\delta > 0$ we can find a constant C such that

$$P\left(\left|\sqrt{n}\,X_n\right| > C\right) < \delta \qquad \text{for } n \in \mathbb{N}\,.$$

For given $\varepsilon > 0$ we can choose N so large that $C/\sqrt{N} < \varepsilon$. And thus

$$P\left(|X_n| > \epsilon\right) \le P\left(|X_n| > C/\sqrt{n}\right) < \delta \qquad \text{for } n \ge N\,.$$

It follows that $X_n \overset{P}{\to} 0$ as desired.

□

An obvious application of this lemma shows that

$$X_n \overset{\text{as}}{\sim} \mathcal{N}\left(\xi, \frac{1}{n}\Sigma\right) \qquad \Rightarrow \qquad X_n \overset{P}{\to} \xi\,. \tag{3.14}$$

So asymptotic normality implies that the variables are concentrating around the asymptotic mean. But asymptotic normality is a much stronger statements, as it gives a very precise description of the magnitude of the deviations that must be expected - it tells us **how** the distribution of X_n concentrates around ξ for large values of n.

The fundamental source for results on asymptotic normality are the various central limit theorems, that we shall encounter later in this chapter. These theorems are concerned with sums and averages. But in applications we will frequently encounter processes that are best described as *modified* averages, and it is essential to be able to say that such processes are still asymptotically normal. Luckily, asymptotic normality is preserved under a number of operations, in particular under transformations with differentiable functions:

Theorem 3.22 *Let $X_1, X_2, \ldots$ and Y be stochastic variable with values in $\mathbb{R}^k$, and assume that $\sqrt{n}\,X_n \xrightarrow{\mathcal{D}} Y$. Let $f : \mathbb{R}^k \to \mathbb{R}^m$ be a measurable map. Assume that $f(0) = 0$ and that f is differentiable in 0 with derivative $Df(0) = A$. Then it holds that*

$$\sqrt{n}\, f(X_n) \xrightarrow{\mathcal{D}} A\,Y\,.$$

Proof: Let us define a function $\varepsilon : \mathbb{R}^k \to \mathbb{R}^m$, by the relation

$$f(x) = Ax + |x|\,\varepsilon(x)$$

for $x \neq 0$, and by letting $\varepsilon(0) = 0$. The assumption of differentiability ensures that $\varepsilon(x)$ is continuous in 0. We have that

$$\sqrt{n}\, f(X_n) = \sqrt{n}\Big(A\,X_n + |X_n|\,\epsilon(X_n)\Big) = A\,\sqrt{n}\,X_n + \Big|\sqrt{n}\,X_n\Big|\;\epsilon(X_n)$$

By the continuous mapping theorem $A\,\sqrt{n}\,X_n \xrightarrow{\mathcal{D}} AY$, so Slutsky's lemma implies the desired result if we can show that

$$\Big|\sqrt{n}\,X_n\Big|\;\varepsilon(X_n) \xrightarrow{P} 0\,. \tag{3.15}$$

As $\varepsilon(x) \to 0$ for $x \to 0$ we can for any given $\delta > 0$ find a $\rho > 0$ such that

$$|x| < \rho \quad \Rightarrow \quad |\varepsilon(x)| \leq \delta\,.$$

Hence

$$P\Big(|\varepsilon(X_n)| > \delta\Big) \leq P\Big(|X_n| \geq \rho\Big) \to 0 \quad \text{for } n \to \infty\,,$$

due to lemma 3.21. This implies that $\varepsilon(X_n) \xrightarrow{P} 0$. As $\Big|\sqrt{n}\,X_n\Big| \xrightarrow{\mathcal{D}} |Y|$ by the continuous mapping theorem, it follows from lemma 3.17 that

$$\Big(\Big|\sqrt{n}\,X_n\Big|,\, \varepsilon(X_n)\Big) \xrightarrow{\mathcal{D}} (|Y|,\, 0)\;.$$

The continuous mapping theorem applied to simple multiplication (similar to the idea in the proof of corollary 3.19) shows that

$$\Big|\sqrt{n}\,X_n\Big|\;\varepsilon(X_n) \xrightarrow{\mathcal{D}} 0\,.$$

And hence we conclude from lemma 3.16 that (3.15) is satisfied

□

Theorem 3.23 (Delta method) *Let $X_1, X_2, \ldots$ be stochastic variables with values in $\mathbb{R}^k$, and let $f : \mathbb{R}^k \to \mathbb{R}^m$ be measurable. If*

$$X_n \overset{as}{\sim} \mathcal{N}\left(\xi, \frac{1}{n}\Sigma\right) .$$

and if f is differentiable in ξ, then it holds that

$$f(X_n) \overset{as}{\sim} \mathcal{N}\left(f(\xi), \frac{1}{n}\, Df(\xi)\,\Sigma\, Df(\xi)^T\right) .$$

Proof: Essentially, this is just a rephrasing of theorem 3.22. We have assumed that $\sqrt{n}\,(X_n - \xi)$ converges in distribution to a stochastic variable Y which is $\mathcal{N}(0, \Sigma)$-distributed. The map

$$g(x) = f(x + \xi) - f(\xi)$$

satisfies that $g(0) = 0$ and is differentiable in 0 with $Dg(0) = Df(\xi)$. Hence it holds that

$$\sqrt{n}\left(f(X_n) - f(\xi)\right) = \sqrt{n}\, g(X_n - \xi) \overset{\mathcal{D}}{\to} Df(\xi)\, X .$$

The distribution for the limit variable follows from the usual transformation rules for normal distributions.

□

The reason for the term ’the delta method’ for theorem 3.23 comes from physics. Think of X_n as the asymptotic mean ξ plus a tiny stochastic perturbation. If the perturbation $X_n - \xi$ is so small that it can be considered infinitesimal, we can for practical purposes pretend that f is affine, and the theorem becomes obvious. It is typical physics jargon to argue in term of such infinitesimal perturbations, usually known as ’deltas’

In practical applications of the delta method, it may well happen that the map f is not a priori defined on all of $\mathbb{R}^k$. In a one-dimensional setting, f may be square root, and for any n there may be positive probability that X_n is negative. It may thus seem difficult to consider the stochastic variable $f(X_n)$ - it is not well-defined at all. But as long as f is well-defined in an open neighborhood of ξ, it is customary to disregard such formal obstacles. We can just extend f to all of $\mathbb{R}^k$ in an arbitrary manner (the spirit of [MT] is to let f be 0 in the parts of space where it is not properly defined already). As long as the extension is measurable, the limit distribution in the delta method will not depend on the details of the extension.

3.6 Triangular array of variables

So far we have been concerned with *sequences* of stochastic variables. To avoid artificial complications when dealing with the necessary normalizations in the central limit theorems, it is useful to introduce a more elaborate enumeration scheme known as a triangular array. In its most basic form this means stochastic variables

$$X_{n\,m}\,, \qquad n = 1, 2, \dots, \quad m = 1, \dots, n\,.$$

Such variables can naturally be written in the form

$$\begin{array}{llll} X_{1\,1} & & & \\ X_{2\,1} & X_{2\,2} & & \\ X_{3\,1} & X_{3\,2} & X_{3\,3} & \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

We denote this layout as a canonical triangular array. Seemingly more complicated triangular arrays can be obtained if there is not exactly n variables in the n'th row of the array, but say N_n. If $N_n \to \infty$ for $n \to \infty$, we can usually think of such general triangular arrays as a canonical triangular array with certain rows deleted, and most of the results we will show can be applied to non-canonical triangular arrays without problems. But for notational reasons we shall stick to canonical triangular arrays.

For a triangular array we can introduce the row-sums

$$S_n = \sum_{m=1}^{n} X_{n\,m} \quad n = 1, 2, \dots$$

Our goal will be to formulate conditions on the variables of the array that ensures that these row-sums converge in distribution to a normal distributed limit variable.

If every variable of the triangular array is real-valued and has second moment, it can be convenient to require that the array is **normalized**, meaning that

$$EX_{n\,m} = 0 \text{ for every } n, m, \quad V\left(\sum_{m=1}^{n} X_{n\,m}\right) = 1 \text{ for every } n\,.$$

If the array is normalized, the row-sums will all have mean 0 and variance 1, giving at least some plausibility to the idea that they may be converging in distribution.

Example 3.24 Let $Y_1, Y_2, \dots$ be independent, identically distributed real-valued stochastic variables with second moment. Let

$$EY_n = \xi\,, \quad VY_n = \sigma^2\,.$$

The corresponding normalized triangular array consists of the variables

$$X_{n\,m} = \frac{1}{\sqrt{n}}\,\frac{Y_m - \xi}{\sqrt{\sigma^2}} \qquad n = 1, 2, \dots, \quad m = 1, \dots, n\,,\,.$$

We see that $X_{n\,m}$ has mean 0 and variance $1/n$. Note that the variables in the n'th row of the array are computed from the same Y's as the previous row except for a single newcomer. What really changes from row to tow is the normalization of the occurring Y's. The row-sums of the normalized array is

$$S_n = \sum_{m=1}^{n} X_{n\,m} = \frac{\sqrt{n}}{\sqrt{\sigma^2}}\left(\frac{1}{n}\sum_{m=1}^{n} Y_i - \xi\right).$$

This is the type of normalization occurring in de Moivre's central limit theorem.

$\circ$

The main demand we will have to the triangular arrays under study here is *independence within rows*. We will assume that the variables$X_{n\,1}, X_{n\,2}, \dots, X_{n\,n}$ are independent for any fixed n. We do **not** assume that variables occurring in different rows are independent. On the contrary, as in example 3.24 the various rows will usually be computed from the same background variables, so a considerable amount of inter-row dependence will be present.

We will establish the fundamental convergence results in terms of an **invariance principle** for triangular arrays: Let $(X_{n\,m})$ and $(X^*_{n\,m})$ be two normalized triangular arrays of real-valued stochastic variables, both satisfying independence within rows and both with the same 2. order structure, meaning that

$$EX^2_{n\,m} = EX^{*\,2}_{n\,m} \quad \text{for every } n, m\,.$$

We will show that it is possible to replace one array with the other hardly without changing the distribution of the row-sums. More precisely we will prove that under mild extra assumptions it holds that

$$\left|\int f(S_n)\,dP - \int f(S^*_n)\,dP\right| \to 0$$

for all $C_b(\mathbb{R})$-functions - or at least for sufficiently many such functions. Here (S_n) and (S_n^*) are the two sequences of row-sums. From this it will follow that if one sequence of row-sums converge in distribution, so will the other, and to the same limit.

The normal distributions enter the picture in the following way: For any normalized triangular array (X_{nm}) we can construct a normalized array (X_{nm}^*) with the exact same second order structure and where all the variables are independent and normally distributed. For X^*-array, the row-sums are simply $\mathcal{N}(0,1)$-distributed for every n. In particular the row-sums converge in distribution to a $\mathcal{N}(0,1)$-distributed limit. So if we can get the invariance principle to hold for a replacement of the X-array with the X^*-array, we have shown that the row-sums of the original array converge in distribution to a $\mathcal{N}(0,1)$-distributed limit. The normal distribution of the limit will so to speak be forced upon us from two fundamental laws of probability: the invariance principle for normalized triangular arrays and the convolution properties of the normal distribution.

3.7 The fundamental replacement trick

In this section we will try to bound magnitudes of the form

$$\left| \int f\left(\sum_{i=1}^n X_i\right) - f\left(\sum_{i=1}^n X_i^*\right) dP \right|$$

where $X_1, \ldots, X_n, X_1^*, \ldots, X_n^*$ are independent real-valued variables, and where f is a suitably nice $C_b(\mathbb{R})$-function.

To make life easy, we will only consider functions $f : \mathbb{R} \to \mathbb{R}$ that are C^3, and where the functions f, f', f'' and f''' are all bounded. We call such functions **C^3-bounded**. Note that any $C_c^\infty(\mathbb{R})$-function is trivially C^3-bounded. So to prove weak convergence, it will suffice to check (3.2) for C^3-bounded functions, according to corollary 3.15.

For a given C^3-bounded function f we can define a map $R : \mathbb{R}^2 \to \mathbb{R}$ by the relation

$$f(y + x) = f(y) + f'(y)\,x + \frac{1}{2}\,f''(y)\,x^2 + R(y, x)\,. \tag{3.16}$$

We observed that R is the difference of two C^1-functions in two variables, so R is itself C^1. In particular R is $\mathbb{B}_2$-measurable.

A 3. order Taylor-expansion of f around y gives that

$$R(y,x) = \frac{1}{6}\, f'''(\xi)\, x^3\,,$$

for a suitable intermediate point ξ between y and $y+x$. This ξ is of course a function of x and y, but the existence of ξ is guaranteed in such an abstract way (it originates in the mean value theorem) that it is impossible to say if $(x,y) \mapsto \xi$ is measurable! Hence we should avoid working too directly with ξ, but it lends itself willingly to estimates such as

$$|R(y,x)| \leq \frac{\|f'''\|}{6}\, |x|^3\,. \tag{3.17}$$

Lemma 3.25 *Let X, X^* and Y be independent real-valued variables. We assume that*

$$EX = EX^*\,, \quad EX^2 = EX^{*\,2}\,,$$

and that X and X^* both have 3. moment. For any C^3-bounded function f it holds that

$$\left|\int f(Y+X) - f(Y+X^*)\, dP\right| \leq \frac{\|f'''\|}{6}\left(E|X|^3 + E|X^*|^3\right)\,.$$

Proof: Let f be a C^3-bounded function, and let R be the corresponding remainder term function from (3.16). For every $y, x, x^* \in \mathbb{R}$ we immediately see that

$$\begin{aligned} f(y,x) - f(y,x^*) &= \left(f(y) + f'(y)\,x + \frac{1}{2}f''(y)\,x^2 + R(y,x)\right) \\ &\qquad - \left(f(y) + f'(y)\,x^* + \frac{1}{2}f''(y)\,x^{*\,2} + R(y,x^*)\right) \\ &= f'(y)\,(x - x^*) + \frac{f''(y)}{2}(x^2 - x^{*\,2}) + R(y,x) - R(y,x^*)\,. \end{aligned}$$

Inserting the stochastic variables Y, X and X^*, it follows that

$$f(Y,X) - f(Y,X^*) = f'(Y)\,(X - X^*) + \frac{f''(Y)}{2}\,(X^2 - X^{*\,2}) + R(Y,X) - R(Y,X^*)\,.$$

As $f'(Y)$ and $f''(Y)$ are bounded, each term is integrable. Independence gives that

$$\int f'(Y)\,(X - X^*)\,dP = \int f'(Y)\,dP\left(\int X\,dP - \int X^*\,dP\right) = 0\,,$$

because the parenthesis is zero. Similarly, we obtain that

$$\int f''(Y)\,(X^2 - X^{*\,2})\,dP = \int f''(Y)\,dP\left(\int X^2\,dP - \int X^{*\,2}\,dP\right) = 0\,.$$

As a consequence of these observations it follows that

$$\int f(Y,X) - f(Y,X^*)\,dP = \int R(Y,X) - R(Y,X^*)\,dP\,.$$

If we combine this with (3.17), we obtain the desired inequality.

□

Theorem 3.26 *Let $X_1,\dots,X_n,X_1^*\,\dots,X_n^*$ be independent real-valued stochastic variables with 3. moment. Assume that*

$$EX_i = EX_i^*\,,\quad EX_i^2 = EX_i^{*\,2}\,,\quad \textit{for } i = 1,\dots,n\,.$$

For any C^3-bounded function f it holds that

$$\left|\int f\left(\sum_{i=1}^n X_i\right) - f\left(\sum_{i=1}^n X_i^*\right)\,dP\right| \le \frac{\|f'''\|}{6}\left(\sum_{i=1}^n E|X_i|^3 + \sum_{i=1}^n E|X_i^*|^3\right)\,.$$

Proof: We introduce new stochastic variables

$$Z_k = \sum_{i=1}^{n-k} X_i + \sum_{i=n-k+1}^{n} X_i^* \qquad \text{for } k = 0,1,\dots,n\,.$$

We see that $Z_0 = \sum_{i=1}^n X_i$, while $Z_n = \sum_{i=1}^n X_i^*$. The idea behind this mixed sums is that two successive variables Z_k and Z_{k+1} can fit into lemma 3.25, as

$$Z_k = Y_k + X_{n-k}\,,\quad Z_{k+1} = Y_k + X_{n-k}^*\,,$$

where

$$Y_k = \sum_{i=1}^{n-k-1} X_i + \sum_{i=n-k+1}^{n} X_i^*\,.$$

When all the X- and X^*-variables are independent of each other, then X_{n-k}, X_{n-k}^* and Y_k are independent. Observing the telescope phenomenon

$$f\left(\sum_{i=1}^n X_i\right) - f\left(\sum_{i=1}^n X_i^*\right) = f(Z_0) - f(Z_n) = \sum_{k=0}^{n-1} f(Z_k) - f(Z_{k+1})\,,$$

if thus follows that

$$\left|\int f\left(\sum_{i=1}^{n} X_i\right) - f\left(\sum_{i=1}^{n} X_i^*\right) dP\right| \leq \sum_{k=0}^{n-1}\left|\int f(Z_k) - f(Z_{k+1})\,dP\right|$$
$$\leq \frac{\|f'''\|}{6}\sum_{k=1}^{n}\left(E|X_i|^3 + E|X_i^*|^3\right),$$

which exactly is the desired inequality.

□

The replacement trick in theorem 3.26 is strong enough to give very useful versions of CLT. But as the essential bound in the theorem has to do with 3. moments of the occurring variables, the conditions are a bit tougher than necessary: the most flexible versions of CLT has conditions where only 2. moments enter, and where the existence of 3. moments is not necessary at all. To obtain such results we need a more sophisticated estimate of the remainder term $R(y, x)$ than (3.17). A 2. order Taylor-expansion of the C^3-function f shows that

$$R(y, x) = \frac{1}{2} f''(\eta)\,x^2 - \frac{1}{2} f''(y)\,x^2$$

for an intermediate point η between y and $y + x$. This gives rise to the inequality

$$|R(y, x)| \leq \|f''\|\,x^2 \quad \text{for } x, y \in \mathbb{R}. \tag{3.18}$$

At the face of it, one would imagine that a 2. order Taylor-expansion gives a poorer estimate of the remainder term than a 3. order expansion. And indeed this is true for **small** values of x. But for x far away from zero the estimate (3.18) is actually better than (3.17). Or more appropriately phrased: the estimate in (3.18) is most likely really bad, but not as bad as (3.17). We get the best of both estimates by observing that

$$|R(y, x)| \leq \begin{cases} \dfrac{\|f'''\|}{6}|x|^3 & \text{for } |x| < c \\ \|f''\|\,x^2 & \text{for } |x| \geq c \end{cases}$$

for a suitable $c > 0$. The inequality is true for any $c > 0$, but in applications the trick may very well be to use a carefully crafted c - or more likely, the whole range of inequalities obtained for varying the value of c. After inserting stochastic variables, we see that

$$\int |R(Y, X)|\,dP \leq \frac{\|f'''\|}{6}\int_{(|X|<c)} |X|^3\,dP + \|f''\|\int_{(|X|\geq c)} X^2\,dP$$
$$\leq \frac{c\,\|f'''\|}{6}\int X^2\,dP + \|f''\|\int_{(|X|\geq c)} X^2\,dP\,.$$

This inequality can directly be applied in the proof of lemma 3.25 and theorem 3.26, and this leads to the following theorem, which remarkably does not require the occurring variables to have 3. moment:

Theorem 3.27 *Let $X_1, \ldots, X_n, X_1^* \ldots, X_n^*$ be independent real-valued stochastic variables with 2. moment. Assume that*

$$EX_i = EX_i^*\,, \quad EX_i^2 = EX_i^{*\,2}\,, \quad \textit{for } i = 1, \ldots, n\,.$$

For any C^3-bounded function f and any $c > 0$ it holds that

$$\left| \int f\left(\sum_{i=1}^n X_i\right) - f\left(\sum_{i=1}^n X_i^*\right) dP \right|$$

is smaller than

$$\frac{c\,\|f'''\|}{6}\left(\sum_{i=1}^n EX_i^2 + EX_i^{*\,2}\right) + \|f''\|\left(\sum_{i=1}^n \int_{(|X_i|\geq c)} X_i^2\,dP + \int_{(|X_i^*|\geq c)} X_i^{*\,2}\,dP\right).$$

3.8 Lyapounov's CLT

Theorem 3.28 (Lyapounov's CLT) *Let $(X_{n\,m})$ be a triangular array of real-valued stochastic variables with 3. moment. Assume that the array satisfies independence within rows, assume that*

$$E\,X_{n\,m} = 0 \;\; \textit{for} \;\; n, m\,, \qquad \lim_{n\to\infty} \sum_{m=1}^n V\,X_{n\,m} = \sigma^2\,, \tag{3.19}$$

and assume that the array satisfies **Lyapounov's condition**,

$$\sum_{m=1}^n E\,|X_{n\,m}|^3 \to 0 \quad \textit{for} \;\; n \to \infty\,. \tag{3.20}$$

Under these assumptions the row-sums $S_n = \sum_{m=1}^n X_{n\,m}$ satisfy that

$$S_n \xrightarrow{\mathcal{D}} X \quad \textit{for} \;\; n \to \infty\,,$$

where the limit variable X is $\mathcal{N}(0, \sigma^2)$-distributed.

Proof: It is perfectly legitimate that the limit value σ^2 of the variance of the row-sums is zero. In that case, the statement is that the row-sums converge in probability to 0. This actually follows from the standard proof of the law of large numbers using Chebyshev's inequality, and does not need Lyapounov's condition at all. So we can safely ignore this case, and assume that $\sigma^2 > 0$ in the remainder of the proof.

Let us use the notation σ^2_{nm} for $E\,X^2_{nm} = V\,X_{nm}$. Construct a triangular array (X^*_{nm}) where all variables are independent, and where

$$X^*_{nm} \sim \mathcal{N}\left(0, \sigma^2_{nm}\right) \,.$$

We shall see in a moment that this array of normal distributed variables automatically satisfies Lyapounov's condition. If (S_n) and (S^*_n) are the row-sums of the X-array and of the X^*-array respectively, it follows from theorem 3.26 that

$$\left|\int f(S_n) - f(S^*_n)\; dP\right| \leq \frac{\|f'''\|}{6}\left(\sum_{m=1}^{n} E|X_{nm}|^3 + \sum_{m=1}^{n} E|X^*_{nm}|^3\right) \to 0 \quad \text{for } n \to \infty$$

for any C^3-bounded function f. The convolution properties of the normal distribution implies that

$$S^*_n \sim \mathcal{N}\left(0, \sum_{m=1}^{n} \sigma^2_{nm}\right) \,.$$

But it follows from example 3.5 that this sequence of normal distributions converge weakly to $\mathcal{N}(0,\sigma^2)$. And hence

$$\int f(S^*_n)\,dP \to \int f(X)\,dP \quad \text{for } \; n \to \infty \,,$$

where the variable X is $\mathcal{N}(0,\sigma^2)$-distributed. Combining the information, we obtain that

$$\left|\int f(S_n)\,dP - \int f(X)\,dP\right| \to 0 \quad \text{for } \; n \to \infty \,,$$

for any C^3-bounded function f. And thus $S_n \xrightarrow{\mathcal{D}} X$ as desired.

All that is left to prove is that the auxiliary X^*-array constructed in the course of the argument indeed satisfies Lyapounov's condition. If U has a standard normal distribution, then

$$E\,|X^*_{nm}|^3 = \sigma^3_{nm}\,E\,|U|^3 \,. \tag{3.21}$$

It is possible to express $E|U|^3$ as a Γ-integral, and actually it can be shown that

$$E\,|U|^3 = \sqrt{8/\pi}\,.$$

But the precise value of this absolute third moment is of no significance for the subsequent development - all that matters is that the value is finite.

Jensen's inequality, used on the convex function $x \mapsto |x|^{3/2}$, gives that

$$\sigma_{nm}^3 = \left(E\,X_{nm}^2\right)^{3/2} \leq E\,\left|X_{nm}^2\right|^{3/2} = E\,|X_{nm}|^3\ .$$

And hence we see that

$$\sum_{m=1}^{n} E\,|X_{nm}^*|^3 = E\,|U|^3 \sum_{m=1}^{n} \sigma_{nm}^3 \leq E\,|U|^3 \sum_{m=1}^{n} E\,|X_{nm}|^3\ .$$

As the original X-array satisfies Lyapounov's condition, so will the X^*-array.

□

Example 3.29 Let $Y_1, Y_2, \ldots$ be independent, identically distributed real-valued stochastic variables, and assume that these variables have 3. moment. Let us use the notation

$$\xi = EY_i\,, \qquad \sigma^2 = VY_i\,.$$

Introduce the triangular array

$$X_{nm} = \frac{Y_m - \xi}{\sqrt{n}} \qquad n = 1, 2, \ldots,\ m = 1, \ldots, n\,.$$

All the variables of this array have mean 0, and we see that $V\,X_{nm} = \sigma^2/n$ implying that (3.19) is satisfied. Furthermore,

$$E\,|X_{nm}|^3 = \frac{E\,|Y_m - \xi|^3}{n^{3/2}}\,,$$

and so

$$\sum_{m=1}^{n} E\,|X_{nm}|^3 = \sum_{m=1}^{n} \frac{E\,|Y_m - \xi|^3}{n^{3/2}} = \frac{n\,E\,|Y_1 - \xi|^3}{n^{3/2}} \to 0 \quad \text{for } n \to \infty\,.$$

Hence it follows from Lyapounov's CLT that

$$\frac{1}{\sqrt{n}} \sum_{m=1}^{n} (Y_m - \xi) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right)\,.$$

Or if you will, that

$$\frac{1}{n}\sum_{m=1}^{n} Y_i \overset{\text{as}}{\sim} \mathcal{N}\left(\xi, \frac{1}{n}\sigma^2\right) . \tag{3.22}$$

This statement is usually referred to as Laplace's CLT, and it is the centerpiece of weak convergence theory.

$\circ$

This particular proof of Laplace' CLT requires that the variables have third moment. This is an artefact of the proof - it will appear later that Laplace's CLT holds true under the assumption of second moments only. So Lyapounov's CLT may not give the strongest possible results. But as Lyapounov's condition is so easy to check, theorem 3.28 is arguably the most popular of the general versions of CLT.

Example 3.30 In figure 3.1 and figure 3.2 we have examined the practical content of (3.22) for exponentially distributed variables. The exponential distribution itself is not at all like a normal distribution, so if the averages tend behave like normals, it is a powerful illustration of the strength of the force behind the CLT. For four different values of n we have generated 1 000 variables of the form

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \, , \tag{3.23}$$

where the Y's are independent and exponentially distributed with mean 1. Figure 3.1 illustrates the results in a histogram. It shows that for large values of n, the averages (3.23) will concentrate quite visibly around the value 1. This concentration is no big surprise - it is the law of large numbers in action.

In figure 3.2 we have shown a QQ-plot of the empirical distribution of the averages against the approximating normal distribution from (3.22), that is against the $\mathcal{N}\left(1, \frac{1}{n}\right)$-distribution. The agreement is horribly bad for small values of n - as expected, due to the non-normal behavior of the exponential distribution. But we also see a quite convincing agreement for larger values of n, in particular for $n = 100$.

The conclusion is that the averages (3.23) will not only concentrate around the expected value ξ when n is large: we can with impressive accuracy describe the variability around ξ with a normal distribution with small variance. And perhaps most important: the value of this small variance in the approximating normal distribution is actually known in advance!

$\circ$

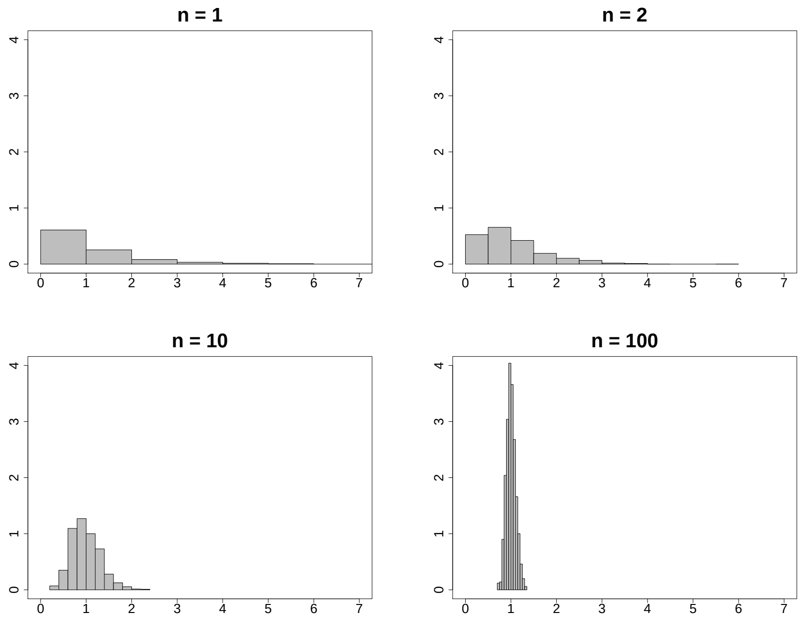


Figure 3.1: *The empirical distribution of 1 000 averages of the form (3.23), when the variables involved are independent and exponentially distributed with mean 1.*

Example 3.31 Let $Y_1, Y_2, \ldots$ be independent, identically distributed real-valued stochastic variables with common distribution function F. Let $\hat{F}_n$ be the empirical distribution function, calculated from the first n observations, that is

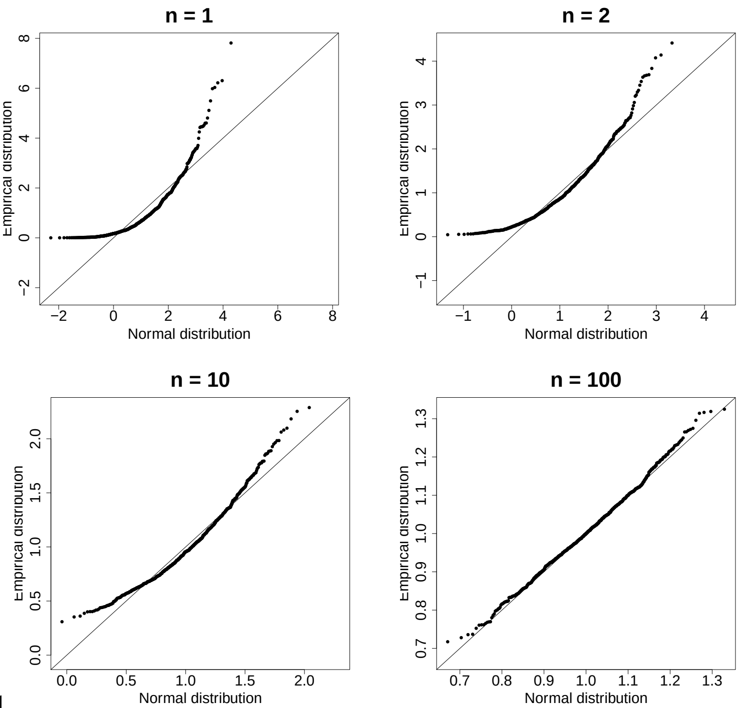
$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x]}(Y_i)\,.$$

The variables $1_{(-\infty,x]}(Y_m)$ are independent Bernoulli variables with success probability $F(x)$. Hence they have mean $F(x)$ and variance $F(x)(1-F(x))$, and they certainly have 3. moment. It follows from example 3.29 that for fixed x it holds that

$$\sqrt{n}\left(\hat{F}_n(x) - F(x)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, F(x)(1-F(x))\Big) \quad \text{for} \ \ n \to \infty\,.$$

For fixed n this gives rise to the construction of a pointwise confidence-band around the theoretical distribution function F. If we draw

$$x \mapsto F(x) - \frac{1.96}{\sqrt{n}} \sqrt{F(x)(1-F(x))}$$



1

Figure 3.2: *A QQ-plot of the empirical distribution of 1 000 averages of the form (3.23), when the variables involved are independent and exponentially distributed with mean 1. The QQ-plot is against the postulated approximating normal distribution from (3.22).*

as a lower bound and

$$x \mapsto F(x) + \frac{1.96}{\sqrt{n}} \sqrt{F(x)(1 - F(x))}$$

as an upper bound, we will expect that the empirical distribution function 'by and large' is contained in this band. The precise statement is that for fixed x there is a probability converging to 95% as $n \to \infty$, that the empirical distribution function is between these two limits.

In figure 3.3 these limits are drawn in an example where the Y's are uniformly distributed on $(0, 1)$ and where $n = 100$. There is added a single empirical distribution function, based on 100 uniformly distributed variables. This particular empirical distribution function is well within the confidence band, and this is not at all uncommon.

It does however occur for less than 95% of a large numbers of empirical distribution functions.

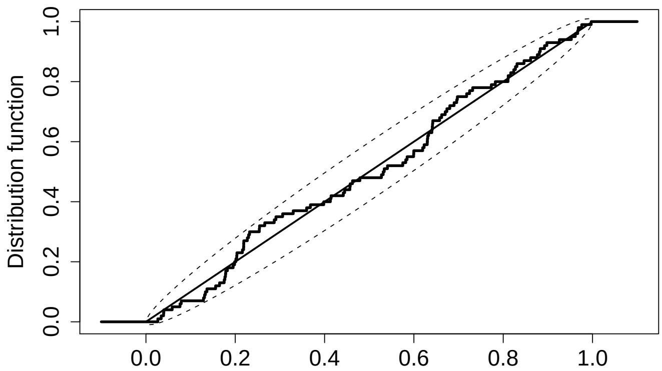


Figure 3.3: *Pointwise asymptotic confidence-bands for the empirical distribution function, based on a sample of 100 independent uniformly distributed random variables. A single example of such a distribution function has been added.*

The probability statement is pointwise for fixed x, it is not uniform - the empirical distribution function may well escape the band here and there, but the domain of this wild behavior will change from replication to replication. It is possible to construct uniform confidence bands, but this requires considerably more advanced weak convergence theory.

$\circ$

Example 3.32 Let us now consider an example where the triangular array formalism proves to be worthwhile. Let $Y_1, Y_2, \ldots$ be independent, identically distributed real-valued stochastic variables with a common distribution function F. Assume that F has a unique p-quantile x_p, and that F is differentiable in x_p with $F'(x_p) = a$ for a suitable constant a. We assume that $a \neq 0$. Note that F is continuous in x_p. In particular it holds that $F(x_p) = p$.

The empirical p-quantile, based on the observations $Y_1, \ldots, Y_n$, may not be uniquely determined, but we choose to use

$$Z_n = Y_{(n:\lceil np \rceil)}\,,$$

where $\lceil y \rceil$ denotes the smallest integer larger than or equal to y, and where $Y_{(n:1)}, Y_{(n:2)}, \dots, Y_{(n:n)}$ are the ordered values of $Y_1, \dots, Y_n$. Though this choice probably makes hard reading, it is quite unproblematic if np is not an integer, for in that case the p-quantile of the empirical distribution function **is** unique, and it is exactly what the formula produces. If np by an unfortunate twist of faith is an integer, things are more complicated: there is a whole interval of p-quantiles for the empirical distribution function - and our definition happens to these the smallest of these.

If the Y's are uniformly distributed on the unit interval, then according to example 20.20 of [MT], Z_n is B-distributed with shape parameters $\lceil np \rceil$ and $n + 1 - \lceil np \rceil$. In that case

$$E\,Z_n = \frac{\lceil np \rceil}{n+1} \approx p\,, \quad V\,Z_n = \frac{\lceil np \rceil\,(n+1-\lceil np \rceil)}{(n+1)^2\,(n+2)} \approx \frac{p(1-p)}{n}\,.$$

As x_p in case equals p, this result supports the idea that is should be possible to say something useful about the asymptotic distribution of

$$\sqrt{n}\,(Z_n - x_p)\,. \tag{3.24}$$

We return to the general case and consider the distribution function for (3.24):

$$\begin{aligned}
P\Big(\sqrt{n}\,(Z_n - x_p) \leq x\Big) &= P(Z_n \leq x_p + x/\sqrt{n}) \\
&= P(\text{at least } \lceil np \rceil \text{ of } Y_1, \dots, Y_n \text{ are smaller that or equal to } \ x_p + x/\sqrt{n}) \\
&= P\left(\sum_{m=1}^{n} 1_{(-\infty, x_p + x/\sqrt{n}]}(Y_m) \geq \lceil np \rceil\right) \\
&= P\left(\frac{1}{\sqrt{n}} \sum_{m=1}^{n} 1_{(-\infty, x_p + x/\sqrt{n}]}(Y_m) - F(x_p + x/\sqrt{n}) \geq \frac{\lceil np \rceil - nF(x_p + x/\sqrt{n})}{\sqrt{n}}\right) \\
&= P\left(\sum_{m=1}^{n} X_{n\,m} \geq x_n\right),
\end{aligned}$$

where we have introduced the triangular array $(X_{n\,m})$ with variables

$$X_{n\,m} = \frac{1_{(-\infty, x_p + x/\sqrt{n}]}(Y_m) - F(x_p + x/\sqrt{n})}{\sqrt{n}}\,,$$

and where we have introduced the sequence of real numbers

$$x_n = \frac{\lceil np \rceil - nF(x_p + x/\sqrt{n})}{\sqrt{n}}\,.$$

The triangular array clearly has independence within rows, and we see that

$$EX_{n\,m} = 0\,, \quad VX_{n\,m} = \frac{F(x_p + x/\sqrt{n})(1 - F(x_p + x/\sqrt{n}))}{n}\,, \quad E|X_{n\,m}|^3 \leq n^{-3/2}\,.$$

Obviously the 3. moments of the array satisfy Lyapounov's condition, and as

$$\sum_{m=1}^{n} VX_{n\,m} \to p(1-p) \quad \text{for} \;\; n \to \infty\,,$$

we find from Lyapounov's CLT that

$$\sum_{m=1}^{n} X_{n\,m} \xrightarrow{\mathcal{D}} X\,,$$

where the limit variable X is $\mathcal{N}(0, p(1-p))$-distributed. Considering the sequence x_n we observe - by adding and subtracting $p = F(x_p)$ - that

$$x_n = \frac{\lceil np \rceil - np}{\sqrt{n}} - x\,\frac{F(x_p + x/\sqrt{n}) - F(x_p)}{x/\sqrt{n}} \to 0 - x\,a \quad \text{for} \;\; n \to \infty\,.$$

Using that the distribution function for a normal distribution is continuous, we see that

$$P\left(\sum_{m=1}^{n} X_{n\,m} \geq x_n\right) \to P(X \geq -xa)\,.$$

But the limit variable X has a distribution which is symmetric around 0, so

$$P(X \geq -xa) = P(X \leq xa)\,.$$

All in all we can conclude that

$$P\left(\sqrt{n}(Z_n - x_p) \leq x\right) \to P(X \leq xa) = P(X/a \leq x)\,.$$

So the distribution function for the normalized empirical p-quantiles in (3.24) converges pointwise to the distribution function for a normal distribution with mean 0 and variance $\frac{p(1-p)}{a^2}$. This can be rephrased as:

$$Z_n \overset{as}{\sim} \mathcal{N}\left(p, \frac{p(1-p)}{n\,a^2}\right)\,. \tag{3.25}$$

We have been working under the assumption that F is differentiable in x_p with a non-zero derivative. This assumption cannot be relaxed. If it does not hold, the $\sqrt{n}$-normalization is typically not the right choice, and even if this is corrected, the limit distribution will in most cases not be normal.

∘

Theorem 3.33 (Multidimensional Lyapounov CLT) *Let $(X_{n\,m})$ be a triangular array of stochastic variables with values in $\mathbb{R}^k$ and 3. moment. . Assume that the array satisfies independence within rows, assume that*

$$E\,X_{n\,m} = 0 \;\; \textit{for } n, m\,, \qquad \lim_{n\to\infty} \sum_{m=1}^{n} V\,X_{n\,m} = \Sigma$$

for some $k \times k$ matrix Σ, and assume that the array satisfies Lyapounov's condition

$$\sum_{m=1}^{n} E\,|X_{n\,m}|^3 \to 0 \quad \textit{for} \;\; n \to \infty\,. \tag{3.26}$$

Under these assumptions the row-sums $S_n = \sum_{m=1}^{n} X_{n\,m}$ satisfy that

$$S_n \xrightarrow{\mathcal{D}} X \quad \textit{for} \;\; n \to \infty\,,$$

where the limit variable X is $\mathcal{N}(0, \Sigma)$-distributed.

Proof: The result is an easy combination of the one-dimensional CLT with Lyapounov's condition and the Cramér-Wold theorem. Take a fixed vector $v \in \mathbb{R}^k$, and consider the triangular array $(v^T X_{n\,m})$ of real-valued variables.

This new array has independence within rows, is has

$$E\,v^T X_{n\,m} = v^T E\,X_{n\,m} = 0\,,$$

it has

$$V\,v^T X_{n\,m} = v^T\;V\,X_{n\,m}\,v\,,$$

and according to the Cauchy-Schwarz inequality it has

$$E\,|v^T X_{n\,m}|^3 \leq |v|^3\,E\,|X_{n\,m}|^3\,.$$

The latter observation clearly shows that this new array satisfies Lyapounov's condition, and as

$$\sum_{m=1}^{n} V\,v^T X_{n\,m} = v^T \left(\sum_{m=1}^{n} V\,X_{n\,m} \right) v \to v^T\,\Sigma\,v \quad \text{for} \;\; n \to \infty$$

we see that

$$v^T \sum_{m=1}^{n} X_{n\,m} = \sum_{m=1}^{n} v^T X_{n\,m} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, v^T\,\Sigma\,v\right)\,.$$

But this limit distribution is exactly the distribution of $v^T X$, if X is $\mathcal{N}(0, \Sigma)$-distributed. So the desired multidimensional result follows from the Cramér-Wold theorem.

□

As an easy application of theorem 3.33 we can repeat the calculations from example 3.29 to see that if $Y_1, Y_2, \dots$ are independent, identically distributed stochastic variables with values in $\mathbb{R}^n$ and with

$$E\,Y_i = \xi\,, \qquad V\,Y_i = \Sigma\,,$$

then it holds that

$$\frac{1}{n}\sum_{m=1}^{n} Y_i \overset{\text{as}}{\sim} \mathcal{N}\left(\xi, \frac{1}{n}\Sigma\right). \tag{3.27}$$

under the extra condition that $E\,|Y_i|^3 < \infty$. As in one dimension, the assumption on existence of 3. moments is an artefact, and as we will see in the next section, it can be replaced by the more natural condition of existence of 2. moments.

3.9 Lindeberg's CLT

Lyapounov's condition is very easy to use, and it is the natural tool for proving CLTs in most specific examples. But as we have mentioned several times, the condition is a tad too strong, and this gives rise to superfluous or artificial limitations. More sophisticated conditions for CLTs can be a blessing. A triangular array $(X_{n\,m})$ of real-valued stochastic variables satisfies **Lindeberg's condition** if

$$\lim_{n\to\infty}\sum_{m=1}^{n}\int_{(|X_{n\,m}|>c)} X_{n\,m}^2\,dP = 0 \qquad \text{for every } c > 0\,. \tag{3.28}$$

This condition does not appear as natural as Lyapounov's condition. But it works in the same way: it protects us form working with triangular arrays where the variables are progressively more heavy-tailed as we go from row to row.

Note the trivial integration inequality

$$\int_{(|X_{n\,m}|>c)} X_{n\,m}^2\,dP \le \int_{(|X_{n\,m}|>c)} \frac{|X_{n\,m}|^3}{c}\,dP \le \frac{E\,|X_{n\,m}|^3}{c}\,. \tag{3.29}$$

This inequality implies that a triangular array satisfying Lyapounov's condition will also satisfy Lindeberg's condition.

Lemma 3.34 *Let $(X_{n\,m})$ be a triangular array of real-valued stochastic variables with 2. moment. If $EX_{n\,m} = 0$ for every n and m, and if the array satisfies Lindeberg's condition (3.28), then it holds that*

$$\max_{m=1,\dots,n} V X_{n\,m} \to 0 \quad \textit{for } n \to \infty\,. \tag{3.30}$$

Proof: For every $c > 0$ and every n, m we have that

$$\int X_{n\,m}^2\, dP = \int_{(|X_{n\,m}| \le c)} X_{n\,m}^2\, dP + \int_{(|X_{n\,m}| > c)} X_{n\,m}^2\, dP \le c^2 + \sum_{k=1}^{n} \int_{(|X_{n\,k}| > c)} X_{n\,k}^2\, dP\,.$$

This implies that

$$\max_{m=1,\dots,n} V X_{n\,m} \le c^2 + \sum_{k=1}^{n} \int_{(|X_{n\,k}| > c)} X_{n\,k}^2\, dP\,,$$

and as the last term converges to zero, we can find an N such that

$$\max_{m=1,\dots,n} V X_{n\,m} \le 2\,c^2 \qquad \text{for } n \ge N\,.$$

But c can be chosen arbitrarily small, and hence (3.30) follows.

□

Theorem 3.35 (Lindeberg's CLT) *Let $(X_{n\,m})$ be a triangular array of real-valued stochastic variables. Assume that the array satisfies independence within rows, that*

$$E\, X_{n\,m} = 0 \;\; \textit{for every } n, m\,, \qquad \lim_{n\to\infty} \sum_{m=1}^{n} V X_{n\,m} = \sigma^2\,,$$

and that the array satisfies Lindeberg's condition,

$$\lim_{n\to\infty} \sum_{m=1}^{n} \int_{(|X_{n\,m}| > c)} X_{n\,m}^2\, dP = 0 \qquad \textit{for every } c > 0\,. \tag{3.31}$$

Under these assumptions, the row-sums $S_n = \sum_{m=1}^{n} X_{n\,m}$ satisfy that

$$S_n \xrightarrow{\mathcal{D}} X \quad \textit{for} \;\; n \to \infty\,,$$

where the limit variable X is $\mathcal{N}(0, \sigma^2)$-distributed.

Proof: By and large we copy the proof of Lyapounov's CLT. We use the notation σ^2_{nm} for the 2. moment $E\,X^2_{nm} = V\,X_{nm}$. We construct a new triangular array (X^*_{nm}) where all the variables are independent, and where

$$X^*_{nm} \sim \mathcal{N}\left(0, \sigma^2_{nm}\right) \,.$$

We shall see in a moment that this new array of normal distributed variables automatically satisfies Lindeberg's condition. If (S_n) and (S^*_n) are the row-sums of the X-array and of the X^*-array respectively, it follows from theorem 3.27 that

$$\left|\int f(S_n) - f(S^*_n)\; dP\right| \leq \frac{2\,c\,\|f'''\|}{6}\left(\sum_{m=1}^{n}\sigma^2_{nm}\right)$$
$$+\,\|f''\|\left(\sum_{m=1}^{n}\int_{(|X_{nm}|>c)} X_i^2\,dP + \int_{(|X^*_{nm}|>c)} X^{*\,2}_{nm}\,dP\right),$$

for any C^3-bounded function f. Lindeberg's condition ensures that the last term of this upper bound converges to zero. However, the first term will not disappear. We see that

$$\limsup_{n\to\infty}\left|\int f(S_n) - f(S^*_n)\; dP\right| \leq \frac{c\,\|f'''\|\,\sigma^2}{3} \,.$$

But c can be chosen arbitrarily small, so in fact we can conclude that

$$\limsup_{n\to\infty}\left|\int f(S_n) - f(S^*_n)\; dP\right| = 0 \,.$$

The convolution properties of the normal distribution implies that

$$S^*_n \sim \mathcal{N}\left(0, \sum_{m=1}^{n}\sigma^2_{nm}\right) \,.$$

But it follows from example 3.5 that this sequence of normal distributions converge weakly to $\mathcal{N}(0,\sigma^2)$. And hence

$$\int f(S^*_n)\,dP \to \int f(X)\,dP \quad \text{for } n\to\infty \,,$$

where the variable X is $\mathcal{N}(0,\sigma^2)$-distributed. Combining the information, we obtain that

$$\left|\int f(S_n)\,dP - \int f(X)\,dP\right| \to 0 \quad \text{for } n\to\infty \,,$$

for any C^3-bounded function f. And thus $S_n \xrightarrow{\mathcal{D}} X$ as desired.

All that is left to prove is that the auxiliary X^*-array constructed in the course of the argument indeed satisfies Lindeberg's condition. If U has a standard normal distribution, then we get from (3.29) that

$$\sum_{m=1}^{n} \int_{(|X^*_{nm}|>c)} X^{*\,2}_{nm}\, dP \leq \frac{E\,|U|^3}{c} \sum_{m=1}^{n} \sigma^3_{nm} \leq \frac{E\,|U|^3}{c} \left(\max_{k=1,\dots,n} \sigma_{nk} \right) \sum_{m=1}^{n} \sigma^2_{nm} \, .$$

According to lemma 3.34 the row-wise maximal standard deviation will converge to zero, while the row-sum of the variances will stay bounded (they will in fact converge). Hence this upper bound converges to zero, for any choice of c.

□

Example 3.36 Let $Y_1, Y_2, \dots$ be independent, identically distributed real-valued stochastic variables, and assume that these variables have 2. moment. Let us use the notation

$$\xi = EY_i \, , \qquad \sigma^2 = VY_i \, .$$

Introduce the triangular array

$$X_{nm} = \frac{Y_m - \xi}{\sqrt{n}} \qquad n = 1, 2, \dots, \; m = 1, \dots, n \, .$$

Let us prove that this array satisfies Lindeberg's condition. Then we can follow the remaining computations from example 3.30 to obtain that

$$\frac{1}{n} \sum_{m=1}^{n} Y_i \overset{\text{as}}{\sim} \mathcal{N}\left(\xi, \frac{1}{n} \sigma^2 \right) \, ,$$

this time without any assumption on 3. moments.

To prove that Lindeberg's condition is satisfied, we see that

$$\sum_{m=1}^{n} \int_{(|X_{nm}|>c)} X^2_{nm}\, dP = \sum_{m=1}^{n} \int_{(|Y_m|>\sqrt{n}\,c)} \frac{(Y_m - \xi)^2}{n}\, dP = \int_{(|Y_1|>\sqrt{n}\,c)} (Y_1 - \xi)^2 \; dP \, .$$

Observing that

$$1_{(|Y_1|>\sqrt{n}\,c)} \; (Y_1 - \xi)^2 \to 0 \qquad \text{for } n \to \infty \, ,$$

dominated by $(Y_1 - \xi)^2$, it follows from Lebesgue's theorem on dominated convergence that Lindeberg's condition is satisfied.

∘

Theorem 3.37 (Multidimensional Lindeberg CLT) *Let (X_{nm}) be a triangular array of stochastic variables with values in $\mathbb{R}^k$ and 2. moment. . Assume that the array satisfies independence within rows, assume that*

$$E\,X_{nm} = 0 \;\; \textit{for } n, m\,, \qquad \lim_{n\to\infty} \sum_{m=1}^{n} V\,X_{nm} = \Sigma$$

for some $k \times k$ matrix Σ, and assume that the array satisfies Lindeberg's condition

$$\lim_{n\to\infty} \sum_{m=1}^{n} \int_{(|X_{nm}|>c)} |X_{nm}|^2 \, dP = 0 \qquad \textit{for every } c > 0\,. \tag{3.32}$$

Under these assumptions the row-sums $S_n = \sum_{m=1}^{n} X_{nm}$ satisfy that

$$S_n \xrightarrow{\mathcal{D}} X \quad \textit{for } \; n \to \infty\,,$$

where the limit variable X is $\mathcal{N}(0, \Sigma)$-distributed.

Proof: The argument is completely analogous to the proof of the multivariate version of Lyapounov's CLT. For any fixed vector $v \in \mathbb{R}^k$ we establish that the triangular array with variables $v^T X_{nm}$ satisfy a CLT with sufficient consistency in the limit variance to enable an application of the Cramér-Wold theorem.

As the computational details are the same as before, we will skip them and focus on the only new element: We have to prove that when the original X-array satisfies (3.32), then the $v^T X$-array satisfies the one-dimensional Lindeberg condition. So pick $v \in \mathbb{R}^k$, and let us assume that $v \neq 0$. Cauchy-Schwarz' inequality ensures that

$$\left|v^T X_{nm}\right| \leq |v| \;\; |X_{nm}| \;.$$

For any given $c > 0$ we have that

$$\left(|v^T X_{nm}| > c\right) \subset \left(|X_{nm}| > c/|v|\right).$$

It follows that

$$\int_{(|v^T X_{nm}|>c)} \left(v^T X_{nm}\right)^2 \, dP \leq \int_{(|X_{nm}|>c/|v|)} |X_{nm}|^2 \, |v|^2 \, dP\,,$$

since both the integrand and the integration domain becomes larger. Now it is quite evident that the $v^T X$-array satisfies Lindeberg's condition.

□

3.10 Problems

3.1. Let f be a fixed probability density on $\mathbb{R}$. Consider $f_n(x) = n\, f(x/n)$. Explain that each f_n is a probability density, and show that $f_n \cdot m \overset{\text{wk}}{\to} \epsilon_0$.
(Note that this is an example of continuous measures converging weakly to a discrete limit.)

3.2. Let $\nu, \nu_1, \nu_2, \dots$ be probability measures on $(\mathbb{R}, \mathbb{B})$, and let $F, F_1, F_2, \dots$ be the corresponding distribution functions.

3.2(a). Show that if

$$F_n(x) \to F(x) \qquad \text{for } \; n \to \infty$$

for every $x \in \mathbb{R}$, then the convergence will in fact be uniform.

Hint: distribution functions have certain function theoretic properties (monotonicity, right continuity, etc.) that can be applied. Start with the easier case where F is continuous.

3.2(b). Assume that the limit function F is continuous. Show that if there is a dense subset $A \subset \mathbb{R}$ such that

$$F_n(x) \to F(x) \qquad \text{for } \; n \to \infty$$

for every $x \in A$, then F_n will in fact converge to F in every point - indeed it will converge uniformly.

3.3. Let (X_n, Y_n) follow a 2-dimensional normal distribution for each n. Assume that $X_n \sim \mathcal{N}(0, 1)$ and that $Y_n \sim \mathcal{N}(0, 1)$ for all n. Does it follow that (X_n, Y_n) converges in distribution?

3.4. Let (X_{nm}) be a triangular array of real-valued stochastic variables with independence within rows and satisfying (3.19). Let $\delta > 0$, and assume that the array satisfies the $2 + \delta$ version of Lyapounov' condition,

$$\sum_{m=1}^{n} E\,|X_{n\,m}|^{2+\delta} \to 0 \quad \text{for } \; n \to \infty\,.$$

Show that

$$\sum_{m=1}^{n} X_{n\,m} \overset{\mathcal{D}}{\to} X \quad \text{for } \; n \to \infty\,,$$

where the limit variable X is $\mathcal{N}(0, \sigma^2)$-distributed.

Hint: Show that the $2 + \delta$ version of Lyapounov' condition implies Lindeberg's condition.