Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Matematica

History of the Central Limit Theorem

Tesi di Laurea in Probabilità

Relatore:
Chiar.mo Prof.
Andrea Pascucci

Presentata da:
Lisa Bettini

Anno Accademico 2023/2024

Al mio fratellino Giulio,
forse un altro futuro matematico.

Introduction

The term "Central Limit Theorem" (in short CLT), indicates a collection of theorems, formulated between 1810 and 1935, regarding the convergence of distributions, densities and discrete probabilities. In this thesis we gathered the most notable approximations and theorems developed through the years, which all led to the current version of the theorem. Probabilities for sums of independent random variables played already an important role in probability theory of the 18th century, as in problems involving games of chance (e.g. with regard to sums of dice rolls), and in the field of the theory of errors, which began to emerge around 1750. Through this analysis, we can observe the fundamental changes of Probability Theory, which started as a tool to solve physical, social and moral problems, then gradually became essential for studying more abstract problems, stochastic models and analytic methods.

In the first chapter, we examine the beginning of the history of the CLT during the 19th century. The protagonists of the first half of this period are Pierre-Simon de Laplace, Siméon-Denis Poisson, Johann Peter Gustav Lejeune Dirichlet and Augustin-Louis Cauchy; then, in the second half, from the St. Petersburg school, Pafnuty Lvovich Chebychev, Andrey Andreyevich Markov and finally Aleksandr Mikhailovich Lyapunov. These authors tried to approximate the distribution of a sum of "errors of observation" (the concept of random variables was coined by Kolmogorov only in 1933), but they all obtained formulae too complicated for a direct numerical evaluation if the number of errors exceeded a relatively small value. A notable result for "choses", an early concept for random variable, was studied by Poisson, who also coined the term "Law of Large Numbers", referring to the fact that, observing a large number of phenomena of the same kind, he found that ratios of these numbers are "almost" constant. In the second half of the century, the Russians of St. Petersburg introduced the CLT and started writing proofs, by adjusting the conditions on the initial variables, obtaining further approximations of sums of "independent quantités" to a normal distribution.

In the second chapter we continue the analysis on newer version of the CLT during the 20th Century, when probability theory became an object of study within mathematics

itself. In 1920 George Pólya coined the name "Central Limit Theorem", to underline its central role in probability theory. Also in 1920, Jarl Waldemar Lindeberg succeeded in finally proving the theorem. We have also reported the work of Felix Hausdorff, Sergei Natanovich Bernstein and Paul Lévy, who mostly worked with characteristic functions. In the second section of the chapter we write about Lévy's and Feller discoveries on necessary and sufficient conditions for the theorem.

The third and final chapter presents the Weak and Strong Law of Large Numbers and their proofs. In conclusion, we examine the modern version of the CLT and provide its proof using Lévy's continuity theorem.

Contents

Chapter 1

The beginning of the history of CLT

The term "Central Limit Theorem", abbreviated with CLT, indicates a collection of theorems, formulated between 1810 and 1935, regarding the convergence of distributions, densities or discrete probabilities. The term itself was the title of a paper published in 1920 by George Pólya, in order to underline its central role in probability theory. Therefore, strictly speaking, one should not really refer to *the* central limit theorem in connection with sums of independent random variables, but rather to *a* central limit theorem on a case-by-case basis. We will now discuss the development of this theorem from a basic idea in the natural and social sciences into an autonomous mathematical theorem, or more correctly into an entire group of such theorems. The first approaches to the theorem were influenced by Abraham De Moivre's approximations to binomial distributions. Before starting to illustrate the history of the theorem, we will give its statement, as a reference for the approximations that follow.

Theorem 1.1 (**Central Limit Theorem**). *Let $(X_n)_n$ be a sequence of real-valued random variables i.i.d. in $\mathcal{L}^2$, with $\mathbb{E}[X_n] = \mu$ and $Var(X_n) = \sigma^2$ $\forall n$. If $\sigma > 0$, we have*

$$\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1). \tag{1.1}$$

The "prehistory": De Moivre's approximations.

In 1733, De Moivre set himself the task of refining the main theorem of *ars conjectandi* [1713] by Jakob Bernoulli, also known today as "Law of Large Numbers". Bernoulli had shown that for n identical and independent trials, if h_n is the relative frequency of a particular event occurring with probability p, then

$$\lim_{n\to\infty} P(|h_n - p| \leq \epsilon) = 1 \qquad \forall \epsilon > 0.$$

De Moivre was interested in a more precise approximation to the binomial distribution, and described his method for the special case of $p = \frac{1}{2}$: he started his work with

$$P\left(Z = \left[\frac{n}{2}\right] + i\right) = 2^{-n}\binom{n}{\left[\frac{n}{2}\right] + i}. \tag{1.2}$$

To approximate the probability (1.2) that exactly $\left[\frac{n}{2}\right] + i$ "successes" Z will be achieved for a large number n of trials, De Moivre first provided the approximations

$$\frac{\binom{n}{\left[\frac{n}{2}\right]}}{2^n} \approx \frac{2}{\sqrt{2\pi n}} \quad \text{and} \quad \log\left(\frac{\binom{n}{\left[\frac{n}{2}\right]+i}}{\binom{n}{\left[\frac{n}{2}\right]}}\right) \approx -2\frac{i^2}{n}. \tag{1.3}$$

From (1.3), it follows

$$P\left(Z = \left[\frac{n}{2}\right] + i\right) \approx \frac{2}{\sqrt{2\pi n}}\mathrm{e}^{-2\frac{i^2}{n}}. \tag{1.4}$$

This statement can be "translated" into a more modern form, a "local" limit theorem, but this was not De Moivre's main goal, which was instead to approximate $\mathbb{P}(|Z - \left[\frac{n}{2}\right]| \leq t)$. Furthermore, he still did not have a concept at his disposal to adequately match the idea of the exponential function. He approximated the probability above according to

$$\begin{aligned}\mathbb{P}\left(\left|Z - \left[\frac{n}{2}\right]\right| \leq t\right) &\approx 2\frac{2}{\sqrt{2\pi n}}\sum_{i=0}^{t}\mathrm{e}^{-2\frac{i^2}{n}} \\ &\approx \frac{4}{\sqrt{2\pi}}\int_0^{t/\sqrt{n}} e^{-2y^2}\mathrm{d}y = \frac{4}{\sqrt{2\pi n}}\int_0^t e^{-2\frac{x^2}{n}}\mathrm{d}x.\end{aligned}$$

Probabilities for sums of independent random variables played an important role in probability theory of the 18th century, as in problems involving games of chance (e.g. with regard to sums of dice rolls) and in the field of the theory of errors, which began to emerge around 1750. Thanks to De Moivre, around 1730 it was possible to establish formulae for probabilities and density functions of sums of i.i.d. random variables if the distribution of the individual summands could be expressed by simple algebraic terms. However, even with a number of random variables that was still fairly small it became impossible to numerically and analytically evaluate the results obtained in this way. Although Daniel Bernoulli succeeded in 1780 in introducing an approximation method that was completely different from the de Moivrian approach, its scope of application remained limited to binomial distributions and thus to distributions of sums of two-valued random variables. In the 18th century, it was impossible to get significantly beyond de Moivre's "limit" theorems.

1.1 First approximations

Laplace

The history of the CLT actually starts in 1812, when Pierre-Simon de Laplace published the first edition of his *Théorie analytique des probabilités*. This book considerably influenced probability theory and mathematical statistics of the 19th century, by analyzing typical problems, stochastic models and analytic methods. Laplace was one of the first mathematicians that pointed out the importance of probability theory in mathematics, and not only for applications to physical, social and moral problems. In many problems referring to stochastic models depending on a large number of trials, probabilities could only be expressed by formulae too complicated for direct numerical evaluation. For example, he tried to compute the probability that the sum of the angles of inclination of comet orbits (or the arithmetic mean of these angles, respectively) lay within certain limits. Laplace assumed that the angles, measured against the ecliptic, were uniformly distributed between 0° and 90°, and also implicitly assumed that all angles were stochastically independent. Using induction, Laplace successfully computed these probabilities for an arbitrary number of comets. In the most simple case, each of the n variables had the same uniform distribution between 0 and h. For the probability P that the sum of those variables was between a and b with $0 \leq a \leq b \leq nh$, Laplace obtained

$$P = \frac{1}{h^n n!} \left(\sum_{i=0}^{N} \binom{n}{i} (-1)^i (b - ih)^n - \sum_{i=0}^{M} \binom{n}{i} (-1)^i (a - ih)^n \right), \tag{1.5}$$

where $N = \min\left(n, \left[\frac{b}{h}\right]\right)$ and $M = \min\left(n, \left[\frac{a}{h}\right]\right)$. Formulae of this kind were too complicated for a direct numerical evaluation if the number of random variables exceeded a relatively small value. Through the use of (1.5) alone, Laplace was unable to address the hypothesis that the comets' planes of motion resulted at "random", so this work could not develop usable approximations.
Thus, for a reasonable application of many of the results of probability calculus, particular considerations were needed to obtain useful approximations of the "formulae of large numbers".

Laplace had his first approach to the CLT in 1810, after modifying generating functions. He considered $X_1, \ldots, X_n$ i.i.d. random variables, with zero mean and which take the values $\frac{k}{m}$, with m a given natural number and $k = -m, -m+1, \ldots, m-1, m$ and respective probabilities p_k. For the calculation of the probability

$$P_j = \mathbb{P}\left(\sum_{l=1}^{n} X_l = \frac{j}{m} \right) \quad \text{for } j \in [-nm, nm],$$

Laplace made use of the generating function $T(t) = \sum_{k=-m}^{m} p_k t^k$. Due to the mutual independence of the X_l's, P_j is equal to the coefficient of t^j in $[T(t)]^n$, after carrying out the multiplication. He then changed variable from t to e^{ix} and introduced the now so-called *characteristic functions* in a special case. From

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{-itx}\mathrm{e}^{isx}\mathrm{d}x = \delta_{ts}, \quad t, s \in \mathbb{Z},$$

it follows that

$$P_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{-ijx}\left[\sum_{k=-m}^{m} p_k \mathrm{e}^{ikx}\right]^n \mathrm{d}x. \tag{1.6}$$

The last integral above was at least formally accessible to Laplace's method of approximation. There was, however, a certain modification necessary, as Laplace did not consider an expansion of the whole integrand around its maximum at $x = 0$, but only of the power with exponent n (equal to the characteristic function). By expanding e^{ikx} in (1.6) into powers of x one gets

$$P_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ijx}\left[\sum_{k=-m}^{m} p_k \left(1 + ikx - \frac{k^2x^2}{2} - \frac{ik^3x^3}{6} + \cdots\right)\right]^n dx.$$

Considering $\sum_{k=-m}^{m} p_k k = 0$, and by replacing $m^2\sigma^2 = \sum_{k=-m}^{m} p_k k^2$, we obtain

$$P_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ijx}\left[1 - \frac{m^2\sigma^2x^2}{2} - iAx^3 + \cdots\right]^n \mathrm{d}x,$$

where A is a constant depending on $\sum_{k=-m}^{m} p_k k^3$. By expanding

$$\log\left[1 - \frac{m^2\sigma^2x^2}{2} - iAx^3 + \cdots\right]^n =: \log z$$

into a series of powers of x, we obtain

$$\log z = -\frac{m^2\sigma^2 n x^2}{2} - iAnx^3 + \cdots,$$

so

$$z = \mathrm{e}^{-\frac{m^2\sigma^2nx^2}{2} - iAnx^3 + \cdots} = \mathrm{e}^{-\frac{m^2\sigma^2nx^2}{2}}(1 - iAnx^3 + \cdots)).$$

After transforming the variable of integration according to $x = \frac{y}{\sqrt{n}}$, the result is

$$P_j = \frac{1}{2\pi\sqrt{n}}\int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} e^{-ij\frac{y}{\sqrt{n}}}\mathrm{e}^{-\frac{m^2\sigma^2y^2}{2}}\left(1 - \frac{iAy^3}{\sqrt{n}} + \cdots\right)\mathrm{d}y,$$

For an approximation with a very large n he ignored all series terms with a power of $\sqrt{n}$ in the denominator, and set the limits of integration to $\pm\infty$. This way we get

$$P_j \approx \frac{1}{2\pi\sqrt{n}}\int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} e^{-ij\frac{y}{\sqrt{n}}}\,\mathrm{e}^{-\frac{m^2\sigma^2y^2}{2}}\mathrm{d}y,$$

where the last integral is equal to

$$\frac{1}{m\sigma\sqrt{2\pi n}}\mathrm{e}^{-\frac{j^2}{2m^2\sigma^2 n}}. \tag{1.7}$$

Summing up (1.7) for $\frac{j}{m} \in [r_1\sqrt{n}, r_2\sqrt{n}]$, which can be approximated by integration ($\mathrm{d}x \approx \frac{1}{\sqrt{n}}$), we obtain

$$\begin{aligned} P(r_1\sqrt{n} \leq \sum X_l \leq r_2\sqrt{n}) &\approx \sum_{j\in[mr_1\sqrt{n};mr_2\sqrt{n}]} \frac{1}{m\sigma\sqrt{2\pi n}} e^{-\frac{j^2}{2m^2\sigma^2 n}} \\ &\approx \int_{mr_1}^{mr_2} \frac{1}{m\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2m^2\sigma^2}}\,dx = \int_{r_1}^{r_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\,dx, \end{aligned}$$

which corresponds to the integral form of the CLT. However, Laplace only treated particular problems concerning the approximation of probabilities of sums or linear combinations of a great number of *random variables*[1] (in many cases errors of observation), and did not justify neglecting series terms of "higher order". His most general version of the CLT was as follows. Let $\epsilon_1, \ldots, \epsilon_n$ be a large number of independent errors of observation, each having the same density with mean μ and variance σ^2. If $\lambda_1, \ldots, \lambda_n$ are constant multipliers and $a > 0$, then

$$P\left(\left|\sum_{j=1}^{n} \lambda_j(\epsilon_j - \mu)\right| \leq a\sqrt{\sum_{j=1}^{n} \lambda_j^2}\right) \approx \frac{2}{\sigma\sqrt{2\pi}} \int_0^a \mathrm{e}^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x.$$

Poisson

From 1824, Siméon-Denis Poisson formulated and proved the CLT generally for "choses", an early concept of random variables, and tried to discuss the validity of this theorem mainly through counterexamples, without really proving it. He investigated the asymptotic behaviour of the distribution of a sum of functions of the values of a "chose", where in several independent experiments these values were obtained with possibly different probabilities. He complicated his approach by considering a function essentially to cover both sums of random values and of powers of these values in the same theorem: all of these quantities are now described as random variables (a term coined by Kolmogorov in 1933). We will now use modern terminology to describe his work, e.g. we will use "random variable" instead of "chose" and so on. Poisson considered $X_1, \ldots, X_s$ to be a great number of random variables with density functions $f_n(x) = F_n'(x)$, where

[1] The term "random variable" was coined by Kolmogorov only in 1933. We abuse of the term to indicate errors of observation, values, or general quantities that these authors studied, which can now be interpreted as random variables.

$F_n(x) = \mathbb{P}(X_n \leq x)$, which decrease sufficiently fast (he did not specify exactly how fast) as their arguments tend to $\pm\infty$. He defined

$$\rho_n(\alpha)\cos(\phi_n) := \int_a^b f_n(x)\cos(\alpha x)\mathrm{d}x \quad \text{and} \quad \rho_n(\alpha)\sin(\phi_n) := \int_a^b f_n(x)\sin(\alpha x)\mathrm{d}x \tag{1.8}$$

It is supposed that, for the absolute values $\rho_n(\alpha)$ of the characteristic functions of X_n (1.8), there exists a function $r(\alpha)$ independent of n with $0 \leq r(\alpha) < 1$, for all $\alpha \neq 0$ such that

$$\rho_n(\alpha) \leq r(\alpha).$$

Then, for arbitrary γ, γ',

$$\mathbb{P}\left(\gamma \leq \frac{\sum_{n=1}^{s}(X_n - \mathbb{E}[X_n])}{\sqrt{2\sum_{n=1}^{s}\mathrm{Var}(X_n)}} \leq \gamma'\right) \approx \frac{1}{\sqrt{\pi}}\int_{\gamma}^{\gamma'} \mathrm{e}^{-u^2}\mathrm{d}u, \tag{1.9}$$

where the approximation improves the larger s is, and the right side is the distribution function of a normal distribution with expectation 0 and variance $\frac{1}{2}$. Strictly speaking, Poisson's analysis could be used for arbitrary γ, γ', though he explicitly expressed end results only for the special case $\gamma = -\gamma' < 0$. We can write (1.9) using the standard normal distribution: by replacing $u = \frac{v}{\sqrt{2}}$, we obtain

$$\mathbb{P}\left(\gamma\sqrt{2} \leq \frac{\sum_{n=1}^{s}(X_n - \mathbb{E}[X_n])}{\sqrt{2\sum_{n=1}^{s}\mathrm{Var}(X_n)}} \leq \gamma'\sqrt{2}\right) \approx \frac{1}{\sqrt{2\pi}}\int_{\sqrt{2}\gamma}^{\sqrt{2}\gamma'} \mathrm{e}^{-\frac{v^2}{2}}\mathrm{d}v.$$

Poisson was convinced that this CLT was also valid for discrete random variables. In this case one could assume that the values $c_1, \ldots, c_v$ of a random variable of this kind were subject to the respect probabilities $\gamma_1, \ldots, \gamma_v$, which were represented by $\gamma_i = \int_{c_i-\delta}^{c_i+\delta} f(z)\mathrm{d}z$, with an "infinitely small" quantity δ and a "discontinuous" density function f.

The approximate stability of arithmetic means or relative frequencies, quite often observed within different sequences of random experiments of the same kind, was so important for Poisson's probabilistic approach that he coined the term "Law of large numbers" for this fact. In the introduction of his *Recherches*[2], he characterized this law as follows: *The phenomena of any kind are subject to a general law, which one can call the "Law of Large Numbers". It consists in the fact, that, if one observes very large numbers of phenomena of the same kind depending on constant or irregularly changeable causes, however not progressively changeable, but one moment in the one sense, the other moment in the other sense; one finds ratios of these numbers which are almost constant.*

[2]Poisson, Siméon Denis 1837. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédés des règles générales du calcul des probabilités.* Paris: Bachelier.

It must be emphasized that Poisson's interpretation of "law of large numbers" is different from the modern definition of this term, which we will illustrate in Chapter 3. After Poisson, probability theory lost its application to moral sciences and moved toward a more mathematical point of view.

Dirichlet

In 1846, Johann Peter Gustav Lejeune Dirichlet discussed linear combinations $\alpha_1 x_1 + \cdots + \alpha_n x_n$ of random errors. The densities of these errors were not only considered to be symmetric and concentrated on a fixed interval, but also to be smooth, which implies the existence of continuous derivatives. He presupposed that the sequence of the α_v had a positive lower bound α and a positive upper bound A, and that all variances of the random errors should be uniformly bounded away from 0. For non-identically distributed observation errors, it has to be assumed also a certain uniformity in the shape of all the density functions f_v, e.g. the existence of an upper bound C such that $|f_v'(x)| < C$ for all $x \in [-a, a]$ and all v. His main result was (expressed in "modern" limit assertion):

$$\left| \mathbb{P}\left(-\lambda\sqrt{n} \leq \sum_{v=1}^{n} \alpha_v x_v \leq \lambda\sqrt{n} \right) - \frac{2}{\sqrt{\pi}} \int_0^{\frac{\lambda}{r}} \mathrm{e}^{-s^2} \mathrm{d}s \right| \xrightarrow{n\to\infty} 0,$$

where

$$r = 2\sqrt{\frac{1}{n}\sum_{v=1}^{n} k_v \alpha_v^2}.$$

In his formula, Dirichlet used the integral form of the CLT with a few differences with the modern one, where we usually use the standard normal distribution. In fact, by doing a few changes, e.g. dividing both members by $\frac{1}{2} r\sqrt{n}$ and changing variables with $x = s\sqrt{2}$, we obtain a more familiar version of the CLT:

$$\left| \mathbb{P}\left(-\frac{2\lambda}{r} \leq \frac{\sum_{v=1}^{n} \alpha_v x_v}{\sum_{v=1}^{n} k_v \alpha_v^2} \leq \frac{2\lambda}{r} \right) - \frac{1}{\sqrt{2\pi}} \int_{-\frac{2\lambda}{r}}^{\frac{2\lambda}{r}} \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x \right| \xrightarrow{n\to\infty} 0,$$

where we can consider $\sqrt{\sum_{v=1}^{n} k_v \alpha_v^2}$ as the variance of the linear combination of errors.

Cauchy

In 1853, Augustin-Louis Cauchy established upper bounds for the error of a normal approximation to the distribution of a linear combination $\sum_{j=1}^{n} \lambda_j \epsilon_j$ of i.i.d. errors ϵ_j with a symmetric density f vanishing for arguments beyond the compact interval $[-k, k]$. He

additionally required that the λ_j should have "the order of magnitude" of $\frac{1}{n}$ or less and $\sum_{j=1}^{n} \lambda_j =: \Lambda$ should be of the order $\frac{1}{n}$. More precisely, the first requirement means there exists positive constants α and β independent of n such that, for all $j = 1, \ldots, n$ there is a $\gamma(j) \geq 1$ with

$$\alpha \leq n^{\gamma(j)} |\lambda_j| \leq \beta.$$

Cauchy used the notation $c := \int_0^k x^2 f(x) \mathrm{d}x$, and for $v > 0$ he obtained

$$\left| \mathbb{P}\left(-v \leq \sum_{j=1}^{n} \lambda_j \epsilon_j \leq v \right) - \frac{2}{\sqrt{\pi}} \int_0^{\frac{v}{2\sqrt{c\Lambda}}} \mathrm{e}^{-\theta^2} \mathrm{d}\theta \right| \leq C_1(n) + C_2(n) + C_3(n), \tag{1.10}$$

where the functions C_1, C_2 and C_3 tend to 0 as n increases, independently of v. These results can be interpreted as a quite rigorous proof of the finite version of a CLT for linear combinations of i.i.d. random variables. In fact, a "modern" CLT can be inferred from Cauchy's version by considering a sequence of independent random variables X_j, distributed like Cauchy's observational errors, and by setting in (1.10) $\lambda_j = \frac{1}{n}$, $v = \frac{a}{\sqrt{n}} (a > 0)$ and $c = \frac{1}{2}\mathrm{Var}(X_1)$, we obtain

$$\begin{aligned} &\left| \mathbb{P}\left(-a\sqrt{n} \leq \sum_{j=1}^{n} X_j \leq a\sqrt{n} \right) - \frac{2}{\sqrt{\pi}} \int_0^{\frac{a}{2\sqrt{c}}} \mathrm{e}^{-x^2} \mathrm{d}x \right| \\ &\qquad \leq C_1(n) + C_2(n, \frac{a}{\sqrt{n}}) + C_3(n) \xrightarrow{n\to\infty} 0. \end{aligned} \tag{1.11}$$

Like the previous case, we can write a more familiar formula containing the standard normal distribution. By replacing $x = \frac{y}{\sqrt{2}}$, we obtain

$$\left| \mathbb{P}\left(-\frac{a}{\sqrt{2c}} \leq \frac{\sum_{i=1}^{n} X_i}{\sqrt{n\mathrm{Var}(X_1)}} \leq \frac{a}{\sqrt{2c}} \right) - \frac{1}{\sqrt{2\pi}} \int_{-\frac{a}{\sqrt{2c}}}^{\frac{a}{\sqrt{2c}}} \mathrm{e}^{-\frac{-y^2}{2}} \mathrm{d}y \right| \xrightarrow{n\to\infty} 0,$$

and since we consider the errors to be i.i.d., we can consider $\sqrt{n\mathrm{Var}(X_1)}$ as the variance of $\sum_{i=1}^{n} X_i$.

1.2 The founders of "St. Petersburg school"

The founders of "St. Petersburg school", in particular Chebychev, Markov and Lyapunov, all had an influence on the history of the CLT. The first two worked with moments, and this Theorem appears in their work to illustrate their methods in moment theory, while the latter worked with it as a mathematical object of its own and he was the first who rigorously proved the CLT.

Chebychev

In 1887, Pafnuty Lvovich Chebychev introduced the CLT in the following version, without a complete proof. Let u_i be a sequence of independent random variables, each with zero expectation. For these random variables he presupposed non negative densities ϕ_i with moments of arbitrarily high order. He assumed that, for each order, an upper and lower bound of the moments existed, uniformly for all random variables. Under these assumptions, Chebychev stated that for any $t < t' \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(t \leq \frac{\sum_{i=1}^n u_i}{\sqrt{2\sum_{i=0}^n \mathbb{E}[u_i^2]}} \leq t' \right) = \frac{1}{\sqrt{\pi}} \int_t^{t'} \mathrm{e}^{-x^2} \mathrm{d}x.$$

As for the results in the previous section, by replacing $y = \sqrt{x}$ and defining $r_1 := \sqrt{2}\, t_1$ and $r_2 := \sqrt{2}\, t_2$, we obtain the usual formula

$$\lim_{n\to\infty} \mathbb{P}\left(r_1 \leq \frac{\sum_{i=1}^n u_i}{\sqrt{2\sum_{i=0}^n \mathbb{E}[u_i^2]}} \leq r_2 \right) = \frac{1}{\sqrt{2\pi}} \int_{r_1}^{r_2} \mathrm{e}^{-y^2} \mathrm{d}y.$$

We can consider $\sqrt{\sum_{i=0}^n \mathbb{E}[u_i^2]}$ as the standard deviation of $\sum_{i=0}^n u_i$, since they are independent and for all j is valid that $\mathbb{E}[u_j] = 0$, so $\mathrm{Var}(u_j) = \mathbb{E}[u_j^2]$. Chebychev did not prove the CLT rigorously, but his theorem was still important for two main reasons: first, he stated his theorem for "quantités" and not for errors, as the other authors before him; second, he explicitly stated conditions for the validity of the assertion, so he was the first to express the CLT properly as a limit theorem.

Markov

After Chebychev had retired from teaching probability theory in 1882, his successor became Andrey Andreyevich Markov, but only in 1898 he started his work on finding a moment theoretic proof of the CLT. His proof of the CLT was actually a corollary of more general moment theoretic results. His version of the CLT was as follows. Let $u_1, u_2, \ldots$ be "independent quantités", obeying the following conditions: $\mathbb{E}[u_k] = 0$ for all k, for all natural $m \geq 2$ there exists a constant C_m such that $|\mathbb{E}[u_k^m]| < C_m$ for all $k \in \mathbb{N}$ and $\mathbb{E}[u_k^2]$ has a positive lower bound. Then

$$\mathbb{P}\left(\alpha \sqrt{2\sum_{k=1}^n \mathbb{E}[u_k^2]} \leq \sum_{k=1}^n u_k \leq \beta \sqrt{2\sum_{k=1}^n \mathbb{E}[u_k^2]} \right) \xrightarrow[n\to\infty]{} \frac{1}{\sqrt{\pi}} \int_\alpha^\beta \mathrm{e}^{-x^2} \mathrm{d}x$$

for any $\alpha < \beta$. As we can see, this formula is very similar to Chebychev's results, with just some differences in the conditions. Indeed, in 1898 Markov did not prove that the

moments of each order of the suitably normed sum of random variables converge to those of the normal distribution respectively, as it would have been essential for the application of his main theorem to the case of the CLT. He gave that proof the year later and his main result was: let $X_1, X_2, \dots$ be a sequence of independent random variables, each with expectation $\mathbb{E}[X_k]$ and variance $\sigma_k^2 > 0$, respectively, where

$$\frac{\sum_{k=1}^n \mathbb{E}[|X_k - \mathbb{E}[X_k]|^r]}{(\sum_{k=1}^n \sigma_k^2)^{\frac{r}{2}}} \xrightarrow[n\to\infty]{} 0 \tag{1.12}$$

and

$$\frac{\sum_{k=1}^n (\sigma_k^2)^{r-1}}{(\sum_{k=1}^n \sigma_k^2)^{r-1}} \xrightarrow[n\to\infty]{} 0 \tag{1.13}$$

for all natural $r \geq 3$, then

$$\mathbb{E}\left[\left(\frac{\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}{\sqrt{2\sum_{i=1}^n \sigma_i^2}}\right)^m\right] \xrightarrow[n\to\infty]{} \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^m \mathrm{e}^{-t^2} \mathrm{d}t.$$

for each natural m. Afterwards, Markov noticed that condition (1.13) was superfluous, because it could be deduced from the first condition (1.12) by means of the inequality

$$(\sigma_k^2)^{r-1} \leq \mathbb{E}[(X_k - \mathbb{E}[X_k])]^{2r-2} \quad (r \geq 3),$$

whose proof is quiet easy.

Markov's and Chebychev's inequalities.

We recall now two important results by these last two authors, which we still very much use in modern probability theory. Markov's result was as follows.

Theorem 1.2. *For all random variables X which take values in $\mathbb{R}^d$, $\lambda > 0$, and $p \in [0, +\infty[$, it holds*

$$\mathbb{P}(|X| \geq \lambda) \leq \frac{\mathbb{E}[|X|^p]}{\lambda^p}. \tag{1.14}$$

In particular, if $Y \in \mathcal{L}^2$ is a real random variable, the Chebychev's inequality holds as

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq \lambda) \leq \frac{\mathrm{Var}(Y)}{\lambda^2}. \tag{1.15}$$

Proof. As for (1.14), if $\mathbb{E}[|X|^p] = +\infty$ there is nothing to prove, otherwise, by the monotonicity property, we have

$$\mathbb{E}[|X|^p] \geq \mathbb{E}\left[|X|^p \mathbb{1}_{\{|X|\geq\lambda\}}\right] \geq \lambda^p \,\mathbb{E}\left[\mathbb{1}_{\{|X|\geq\lambda\}}\right] = \lambda^p \,\mathbb{P}(|X| \geq \lambda).$$

(1.15) follows from (1.14) by setting $p = 2$ and $X = Y - \mathbb{E}[Y]$, indeed

$$P(|Y - \mathbb{E}[Y]| \geq \lambda) \leq \frac{\mathbb{E}\left[|Y - \mathbb{E}[Y]|^2\right]}{\lambda^2} = \frac{\mathrm{Var}(Y)}{\lambda^2}.$$

□

Lyapunov

In 1900, Aleksandr Mikhailovich Lyapunov considered Chebychev's and Markov's work on the CLT, he barely used moment theory whereas he tried to simplify their results in order to find more general conditions for the CLT. In his proof, he used the so-called "Lyapunov inequality".

Lemma 1.3 (Lyapunov inequality)**.** *Let $x', x'', x''', \dots$ be a sequence of positive numbers, and let $f(x)$ be any function whose values $f(x'), f(x''), f(x'''), \dots$ are all positive. If one generally sets*

$$f(x') + f(x'') + f(x''') + \cdots = \sum f(x),$$

and by l, m, n one understands any numbers which are according to the inequalities

$$l > m > n \geq 0,$$

then one has

$$\left(\sum f(x)x^m\right)^{l-n} < \left(\sum f(x)x^n\right)^{l-m} \left(\sum f(x)x^l\right)^{m-n}.$$

Lyapunov proved the following theorem. Let $x_1, x_2, x_3, \dots$ be an infinite sequence of independent random variables, for which the expectations $\mathbb{E}[x_i] =: \alpha_i$, $\mathbb{E}[(x_i - \alpha_i)^2] = a_i$, and $\mathbb{E}[|x_i^3|] =: l_i$ exist, respectively. Furthermore, let

$$A_n := \frac{1}{n}\sum_{i=1}^{n} a_i \quad \text{and} \quad L_n^3 := \max_{1\leq i\leq n} l_i.$$

Under the condition

$$\frac{L_n^2}{A_n} n^{-\frac{1}{3}} \xrightarrow[n\to\infty]{} 0, \tag{1.16}$$

for all $z_1 < z_2 \in \mathbb{R}$, the modulus of

$$\mathbb{P}\left(z_1\sqrt{2nA_n} < \sum_{i=1}^{n}(x_i - \alpha_i) < z_2\sqrt{2nA_n}\right) - \frac{1}{\sqrt{\pi}}\int_{z_1}^{z_2} \mathrm{e}^{-z^2}\mathrm{d}z$$

has an upper bound Ω_n independent of z_1, z_2, such that

$$\Omega_n \xrightarrow[n\to\infty]{} 0.$$

The condition (1.16) is met, for example, if the absolute moments of third order l_i of each single random variable have a uniform lower bound c. Then we have

$$\frac{L_n^2}{A_n} n^{-\frac{1}{3}} \leq \frac{C^2}{c} n^{-\frac{1}{3}} \xrightarrow[n\to\infty]{} 0.$$

Later on he weakened the conditions for his theorem, by presupposing the existence of the respective expectations α_i, a_i and $d_i := \mathbb{E}[|x_i - \alpha_i|^{2+\delta}]$, with $\delta > 0$ arbitrarily small. From these it follows

$$\frac{(d_1 + \cdots + d_n)^2}{(a_1 + \cdots + a_n)^{2+\delta}} \xrightarrow[n\to\infty]{} 0. \tag{1.17}$$

Chapter 2

The CLT at the beginning of the 20th Century

2.1 The CLT in the Twenties

After the First World War, probability theory began to be discovered as a field for ambitious analysts, even outside of Russia. The CLT consequently ceased to be an issue merely for "users", such as astronomers, geodetics specialists, insurance specialists, or economists, who had actually produced quite impressive results in the second half of the 19th century, particularly in the field of error theory, and became an object of study within mathematics itself. In the Twenties a lot of authors, such as von Mises, Pólya, Lindeberg, Lévy, Bernstein and Hausdorff worked on the CLT, and found necessary and sufficient conditions for the theorem.

Von Mises and Pólya

In 1919, Richard Von Mises conceived the notion of "distribution" as a monotonically increasing function, being right continuous and having limit 0 as $x \to -\infty$ and limit 1 as $x \to \infty$, which was important for generality and also precision of analytic exposition. Apparently one of the first to do so, he represented probabilities, as well as higher moments, by Stieltjes integrals[1] referring to those distributions. The use of Stieltjes

[1]The Stieltjes integral is a generalization of Riemann's integral. Given two real-valued functions $f, g : \mathbb{R} \to \mathbb{R}$, let $x_0 = a < x_1 < x_2 < \cdots < x_i < \cdots < x_n = b$ be a partition of the interval $[a, b] \subseteq \mathbb{R}$. From each subinterval defined by the partition, consider a point $c_i \in [x_i, x_{i+1}]$. The mesh $\delta(P)$ of the partition P is $\delta(P) := \max_{x_i \in P} |x_{i+1} - x_i|$. The Riemann-Stieltjes integral of f with respect to g, denoted by $\int_a^b f(x)\,\mathrm{d}g(x)$ is defined as follows: $\lim_{\delta(P)\to 0} \sum_{x_i \in P} f(c_i)(g(x_{i+1}) - g(x_i))$, if it exists independently of the choice of points c_i.

integrals, in addition to the analytic skill employed in dealing with moments, proves that von Mises was informed about the current development of moment theory, at least in its main features. Based on Stieltjes integrals, von Mises formulated and proved his local and integral CLTs for real- and vector-valued random variables as statements about convolutions of discrete probability functions, densities, and distribution functions, respectively. So, his exposition was purely analytic and did not resort to probabilistic interpretations and concepts.

Let us briefly recall Pólya, the mathematician who coined the name "Central limit theorem" in a 1920 article, to underline its *central* role in probability theory.

Lindeberg

The complete mathematical work of Jarl Waldemar Lindeberg contains only one truly outstanding performance: the proof of the CLT under a very weak condition, which under certain "natural" assumptions even proved to be necessary. Lindeberg's arguments were based on an entirely new analytic method, which would later be applied to far more general problems. In 1920 Lindeberg, still without any knowledge of Lyapunov's works, had already proven the CLT for normed sums $\sum_{k=1}^{n} \frac{X_k}{r_n}$ of mutually independent random variables X_k , each with distribution U_k , with zero expectation, variance σ_k^2, and finite absolute moment of third order, presupposing that

$$\frac{1}{r_n^3} \sum_{k=1}^{n} \int_{-\infty}^{\infty} |x|^3 \mathrm{d}U_k(x) \xrightarrow[n\to\infty]{} 0, \qquad r_n = \sqrt{\sum_{k=1}^{n} \sigma_k^2}.$$

After certain modifications of his arguments, in 1922 he was able to publish his famous proof of the CLT under even weaker conditions. He expressed this theorem in several versions. The version which comes closest to Lindeberg's concepts is probably his "Theorem III": let $U_1, \ldots, U_n$ be the distribution functions of n mutually independent "probability quantities" $u_1, \ldots, u_n$ each with expectation 0 and variance σ_k^2, where $\sum_{k=1}^{n} \sigma_k^2 = 1$. Let

$$U(x) := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} U_n(x - t_1 - t_2 - \cdots - t_n)\, \mathrm{d}U_{n-1}(t_{n-1}) \cdots \mathrm{d}U_1(t_1).$$

Then U is the distribution of the sum of all random variables. Let

$$s(x) := \begin{cases} |x|^3 & \text{if } |x| < 1 \\ x^2 & \text{otherwise.} \end{cases} \tag{2.1}$$

Even if the positive number ϵ is taken arbitrarily small, a positive number η can be chosen such that

$$\left| U(x) - \int_{-\infty}^{x} \frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\, \mathrm{d}t \right| < \epsilon \tag{2.2}$$

if

$$\sum_{k=1}^{n} \int_{-\infty}^{\infty} s(x)\, \mathrm{d}U_k(x) < \eta. \tag{2.3}$$

Since U is the distribution of the sum of all random variables, it is equal to $\mathbb{P}(\sum_{k=1}^{n} U_k < x)$. By setting $a_k = \mathbb{E}[U_k] = 0$ and $b = \sqrt{\sum_{k=1}^{n} \sigma_k^2} = 1$, we can write

$$U(x) = \mathbb{P}\left(\frac{\sum_{k=1}^{n} (U_k - a_k)}{b} < x \right)$$

So, Lindeberg proved a theorem which can be applied both to normed partial sums related to simple sequences of random variables and to sums of elements within different rows of a triangular array of random variables. At the end of the proof he obtained

$$\left| U(x) - \int_{-\infty}^{x} \varphi(t,1)\mathrm{d}t \right| < 3\sqrt[4]{\frac{3}{2}} \left(\sum_{i=1}^{n} \int_{-\infty}^{\infty} s(x)\mathrm{d}U_i(x) \right)^{\frac{1}{4}}, \tag{2.4}$$

with the abbreviation $\varphi(x,\sigma) := \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2\sigma^2}}$. In 1922, from (2.4) he obtained the CLT in its usual form: let (X_i) be a sequence of independent random variables with distributions V_i. For simplicity it is assumed that $\mathbb{E}[X_i] = 0$ for all i. Then, for all natural n and for all $i \leq n$, the random variables $u_i := \frac{X_i}{r_n}$, with $r_n = \sqrt{\sum \mathrm{Var}(X_i)}$ and distribution functions $U_i(x) = V_i(r_n x)$, are of the type required for Lindeberg's Theorem (equations (2.2) and (2.3)). From (2.4), under the condition

$$\sum_{i=1}^{n} \int_{-\infty}^{\infty} s\left(\frac{x}{r_n}\right) \mathrm{d}V_i(x) \xrightarrow[n\to\infty]{} 0,$$

with s as defined in (2.1), it follows that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i}{r_n} \leq x \right) = \int_{-\infty}^{x} \frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \mathrm{d}t.$$

Lévy in 1924 used the following condition, that was later called the "*Lindeberg condition*", despite Lindeberg himself never wrote about it:

$$\frac{1}{r_n^2} \sum_{i=1}^{n} \int_{|x|>tr_n} x^2 \mathrm{d}V_i(x) \to 0 \qquad \forall t > 0. \tag{2.5}$$

Hausdorff

Felix Hausdorff was mainly interested in the integral version of the CLT and used Lindeberg's method of proving it in his work. He presupposed "variables" (without explaining this notion) $X_1, \ldots, X_n$ with zero means, second-order moments $a_1, \ldots, a_n$, and absolute third-order moments $c_1, \ldots, c_n$. He deduced a finitary version of the CLT, which he named "Lyapunov's limit theorem", and which can be summarized as follows. If Φ_n denotes the distribution function[2] of $\sum_{k=1}^{n} \frac{X_k}{\sqrt{2}b_n}$ where $b_n^2 = a_1^2 + \cdots + a_n^2$, and

[2]Hausdorff defined the distribution function $F(x)$ of a random variable X by $F(x) := \mathbb{P}(X < x)$.

$d_n = (c_1^3 + \cdots + c_n^3)^{\frac{1}{3}}$. Then, with the denotation $\Phi(x) := \frac{1}{\sqrt{\pi}} \int_{-\infty}^{x} \mathrm{e}^{-t^2} \mathrm{d}t$,

$$|\Phi_n(x) - \Phi(x)| \leq \mu \left(\frac{d_n}{b_n} \right)^{\frac{3}{4}},$$

where μ is a numerical constant. Hausdorff additionally noticed that the condition

$$\frac{d_n}{b_n} \xrightarrow[n\to\infty]{} 0$$

was sufficient for the uniform convergence of $\Phi_n(x)$ to $\Phi(x)$. If we look closer to this condition, we notice that the "Lyapunov condition" (1.17) from 1901 implies it. To prove that, we show that, since $\mathbb{E}[X_i] = 0$, $a_i^2 = \mathbb{E}[X_i^2] = \mathbb{E}[(X_i - \mathbb{E}[X_i])^2]$ and $c_i^3 = \mathbb{E}[|X_i|^3] = \mathbb{E}[|X_i - \mathbb{E}[X_i]|^3]$, and then we use (1.17). We obtain

$$\left(\frac{d_n}{b_n} \right)^{\frac{3}{4}} = \left(\frac{(d_1 + \cdots + d_n)^{\frac{1}{3}}}{(a_1 + \cdots + a_n)^{\frac{1}{2}}} \right)^{\frac{3}{4}} = \frac{(d_1 + \cdots + d_n)^{\frac{1}{4}}}{(a_1 + \cdots + a_n)^{\frac{3}{8}}} = \left(\frac{(d_1 + \cdots + d_n)^2}{(a_1 + \cdots + a_n)^3} \right)^{\frac{1}{4}} \xrightarrow[n\to\infty]{} 0.$$

Lévy

In a 1922 brief discussion of counterexamples to the CLT, Paul Lévy had considered functions

$$\varphi(t) = \mathrm{e}^{-a|t|^\alpha}, \quad (a > 0,\ 0 < \alpha \leq 2),$$

which he referred to as "characteristic functions" of *stable* laws. A probability law $\mathscr{L}$ is called "stable" if it does not correspond to a degenerate distribution (i.e. a distribution concentrated in one point), and if for independent random variables X_1, X_2, each with probability law $\mathscr{L}$, this condition is valid: for all $a_1, a_2 > 0$ such that $\frac{1}{a}(a_1 X_1 + a_2 X_2)$ likewise obeys the law $\mathscr{L}$, with a necessarily uniquely determined. He emphasized the advantages of characteristic functions over generating functions, especially because of their property to be "*always well defined without any restrictions on the probability law*". The notion of stable law directly implies the following property: if there exists a sequence of i.i.d. random variables $(X_i)_{i\in\mathbb{N}}$, and a sequence of positive numbers $(N_n)_{n\in\mathbb{N}}$ and a distribution function V such that

$$\mathbb{P}\left(\frac{\sum_{i=1}^n X_i}{N_n} \leq x \right) + \frac{1}{2}\,\mathbb{P}\left(\frac{\sum_{i=1}^n X_i}{N_n} = x \right) \xrightarrow[n\to\infty]{} V(x)$$

in all points of continuity x of V, then V is the distribution function of a stable law. Lévy considered stable distributions "natural" generalizations of the classic Gaussian law, therefore, with regard to this author, the history of the CLT is closely connected with his more general discussion of stable limit distributions.

The laws of type $L_{\alpha,\beta}$. We need $\psi(t) = -c|t|^\alpha$, where the complex coefficient x may depend on the sign of t. From the general properties of characteristic functions, which had to be valid also for $\varphi(x) = \mathrm{e}^{\psi(x)}$ (in particular $|\varphi(x)| \leq 1,\ \varphi(0) = 1,\ \varphi(-x) = \overline{\varphi(x)}$, continuity), it followed

$$\psi(t) = -(c_0 + \operatorname{sgn}(t)\, c_1 i)|t|^\alpha, \tag{2.6}$$

where $\alpha > 0,\ c_0 \geq 0,\ c_1 \in \mathbb{R}$. For a closer specification of the constants c_0 and c_1, Lévy designated certain probability laws, as "Laws of type $L_{\alpha,\beta}$" if their characteristic function had the form $\mathrm{e}^{\psi(t)}$, where $\psi(t)$ was a function according to (2.6), $c_0 > 0$, and

$$\frac{c_1}{c_0} = \begin{cases} \beta \tan(\frac{\pi}{2}\alpha) & \text{for } \alpha \in]0,1[\cup]1,2[\\ \beta & \text{for } \alpha \in \{1;2\}. \end{cases}$$

He showed that, for all values of β and $\alpha \neq 1, 2$ under consideration, there exists a probability density f with a characteristic function φ such that

$$\left(\varphi\left(\frac{t}{n^{\frac{1}{\alpha}}}\right)\right)^n \xrightarrow[n\to\infty]{} \mathrm{e}^{\psi(t)}. \tag{2.7}$$

Lévy succeeded in proving that the convergence in (2.7) was uniform in each finite interval of t-values, $\psi(t)$ as defined above. In 1922, he proved a first version of the CLT, as a special case of his theorem on the convergence to distributions of type $L_{\alpha,\beta}$. For a sequence of distribution functions $(F_k)_{k\in\mathbb{N}}$ of independent random variables X_k, each with zero expectation and variance 1, let

$$\forall \epsilon > 0\ \exists a > 0\ \forall k \in \mathbb{N} : \int_{|\xi|\leq a} \xi^2\, \mathrm{d}F_k(\xi) \geq 1 - \epsilon.$$

Let $(m_k)_{k\in\mathbb{N}}$ be a sequence of positive numbers with

$$\frac{\max_{k=1,\ldots,n} m_k^2}{\sum_{k=1}^n m_k^2} \xrightarrow[n\to\infty]{} 0. \tag{2.8}$$

Then

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{k=1}^n m_k X_k}{\sqrt{\sum_{k=1}^n m_k^2}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \mathrm{e}^{-\frac{t^2}{2}}\, \mathrm{d}t.$$

Lévy stressed the independence of his and Lindeberg's work. He proved the CLT in 1924 by use of the method of characteristic functions under the modified "Lindeberg condition" (2.5), which was especially appropriate to characteristic functions. This proof largely corresponds to the now common standard proof contained in many textbooks on probability theory. For a comparison of Lévy's conditions with the modified Lindeberg condition one has to refer to random variables X_k with zero expectations and variances σ_k^2. In this case, Lévy's conditions are

$$\forall \epsilon > 0\ \exists a > 0 :\ \forall k \in \mathbb{N}, \quad \frac{1}{\sigma_k^2} \int_{|\xi|\leq a\sigma_k} \xi^2 \mathrm{d}F_k(\xi) \geq 1 - \epsilon, \tag{2.9}$$

and after writing σ_k^2 instead of m_k^2 in (2.8), the Lindeberg condition, with the abbreviation $r_n^2 := \sum_{k=1}^n \sigma_k^2$ is

$$\forall t > 0\ \forall \eta > 0\ \exists n_0 :\ \forall n \geq n_0, \quad \frac{1}{r_n^2} \sum_{k=1}^{n} \int_{|x| \leq r_n t} x^2 \, \mathrm{d}F_k(x) \geq 1 - \eta. \tag{2.10}$$

Whereas Lévy's first condition (2.9) aims at a certain uniformity among the single distribution functions, his second condition (2.8) requires each single variance to be small compared with the variance of the entire sum. Lindeberg's condition (2.10) stresses both aspects at the same time; the uniformity required, however, is weaker than in Lévy's condition. In fact, this last condition can be deduced from both Lévy's conditions together.

We will see more about Lévy and the CLT in the next section, as he continued his studies on this topic later in his life.

Bernstein

In 1922, the same year in which Lindeberg's and Lévy's fundamental contributions to the CLT appeared, Sergei Natanovich Bernstein published a paper containing the so-called "*Lemme fondamental*", which generalizes the assertion of the CLT towards "almost independent" random variables, and can also be applied to sums of random variables which form Markov chains[3]. The statement was as follows: let $S_n = u_1 + \cdots + u_n$, $\mathbb{E}[S_n^2] = B_n$, $\mathbb{E}[u_1^2] + \cdots + \mathbb{E}[u_n^2] = B_n'$ (it is always supposed that, for simplicity of notation, $\mathbb{E}[u_i] = 0$). If, for each arbitrary set of already known values $u_1, \ldots, u_{i-1}$, the absolute values of the mathematical expectations of u_i and u_i^2 do not exceed α_i and β_i respectively, and at the same time the mathematical expectation of $|u_i^3|$ remains below c_i, then

$$\mathbb{P}(z_0\sqrt{2B_n} < S_n < z_1\sqrt{2B_n}) \xrightarrow[n\to\infty]{} \frac{1}{\sqrt{\pi}} \int_{z_0}^{z_1} \mathrm{e}^{-z^2} \mathrm{d}z,$$

presupposing that

$$\frac{1}{\sqrt{B_n}} \sum_{i=1}^{n} \alpha_i, \quad \frac{1}{B_n} \sum_{i=1}^{n} \beta_i, \quad \frac{1}{B_n^{\frac{3}{2}}} \sum_{i=1}^{n} c_i$$

tend to 0 together with $\frac{1}{n}$.

[3] A Markov chain (or Markov process) is a stochastic process describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. More formally, a discrete-time Markov chain is a sequence of random variables (X_k) satisfying the Markov property: $\mathbb{P}(X_{n+1} = x | X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} = x | X_n = x_n)$, if both conditional probabilities are well defined, that is, if $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) > 0$.

Gnedenko and Kolmogorov have hinted at the fact that Bernstein's lemma together with the additional remark in the particular case of independent variables yields very general sufficient conditions for the convergence of the distributions of normed sums to the normal distribution.

2.2 Lévy and Feller on Normal Limit Distributions around 1935

In 1935, both Paul Lévy and William Feller proved that there are necessary and sufficient conditions for the convergence of distributions of suitably normed sums of independent random variables to the normal distribution.

Lévy

In the field of strong laws of large numbers, a significant part was played by necessary and sufficient conditions for the almost sure convergence[4] of a series of independent random variables. Lévy's own version was as follows: let (X_k) be a sequence of independent random variables. For the existence of a sequence of real numbers a_k such that $\sum_{k=1}^{\infty}(X_k - a_k)$ almost surely converges, it is necessary and sufficient that there exists a sequence (Y_k) being equivalent to the sequence (X_k), for which $\sum_{k=1}^{\infty} \operatorname{Var}(Y_k)$ converges. Lévy applied the ideas associated with almost sure convergence to the CLT again: the mutually analogous references, on the one hand between the divergence of the sums of all variances and the almost sure divergence of the series of the random variables, on the other hand between the divergence of the sums of all variances and the validity of the assertion of the CLT, apparently led Lévy to consider equivalent random variables in conjunction with the CLT as well. In this way, he arrived at the following theorem. Let (X_k) be a sequence of independent random variables. If there exists a sequence (Y_k) of bounded random variables being equivalent to (X_k) such that $\max_{1\leq k\leq n}|Y_k|$ and $\frac{d_n^2}{\sum_{k=1}^{n}\operatorname{Var}(Y_k)} \to 0$, then one can find constants A_n and $B_n > 0$ such that the distribution of $\sum_{k=1}^{n}\frac{X_k}{B_n} - A_n$ tends to the standard normal distribution. Lévy's result was more general than Lindeberg's theorem ((2.2) and (2.3)) and he proceeded to find even necessary conditions for the convergence to the normal distribution by further refining the basic ideas which had led to the just-stated theorem. He considered normed sums of random variables, each of them additively composed of one part being "very small in relation

[4]To say that the sequence X_n converges almost surely (or almost everywhere, or with probability 1, or strongly towards X) means that $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$.

to the total sum" and one possibly sizable but normally distributed part. In order to be able to compare the size of an individual random variable to the overall sum, Lévy created two new concepts, "dispersion" and its inverse, "concentration", which proved to be especially useful in discussing the convergence of series of random variables. He defined the **concentration $f_X(l)$** of the random variable X assigned to the interval of length $l > 0$ as follows:

$$f_X(l) := \sup_{-\infty<a<\infty} \mathbb{P}(a < X < a + l),$$

whereas he called **dispersion** of a random variable X a function $\varphi_X : [0, 1[\to \mathbb{R}_0^+$ with

$$\varphi_X(\gamma) := \inf\{x \in \mathbb{R}_0^+ | f_X(x) \geq \gamma\}.$$

Roughly speaking, the dispersion is the minimum interval length related to a particular probability and the concentration is the maximum probability related to a particular interval length. Lévy considered sequences of random variables (X_k) and (η_k), where all variables within each sequence were assumed to be independent. He presupposed that

$$X_k = a_k + b_k\xi_k + \eta_k + \eta_k',$$

where a_k and b_k were constants, ξ_k obeyed a Gaussian law, and η_k' met the condition that $\mathbb{P}(\eta_k' \neq 0)$ was convergent, L_n denoted the dispersion of $\sum_{k=1}^n X_k$, assigned $\gamma = \frac{1}{2}$ and the variables η_k were assumed to be bounded such that $\max_{1\leq k\leq n} \frac{|\eta_k|}{L_n}$ tends to 0. Under the additional assumption that $L_n \to \infty$, Lévy claimed that the distribution of the suitably normed sum of the X_k would tend to a Gaussian law. Later on, he noticed that the assumptions for the part $\eta + \eta'$ were equivalent to the condition

$$\mathbb{P}\left(\max_{1\leq k\leq n} |X_k| > \epsilon L_n\right) \xrightarrow[n\to\infty]{} 0 \quad \forall \epsilon > 0,$$

which turned out to be even sufficient for convergence to the normal distribution if all random variables X_k could be considered small in the sense

$$\max_{1\leq k\leq n} \mathbb{P}(|X_k| > \epsilon L_n) \xrightarrow[n\to\infty]{} 0 \quad \forall \epsilon > 0.$$

Lévy also proved a "classical" version of the CLT: let (X_k) be a sequence of i.i.d. random variables. The distribution $s_n := \frac{\sum_{k=1}^n X_k}{\sqrt{n}}$ for $n \to \infty$ tends to the standard distribution Φ if and only if $\mathbb{E}[X_1^2] = 1$ and $\mathbb{E}[X_1] = 0$. Given the properties of the CLT, he only had to show that, under the given assumptions,

$$\mathbb{P}\left(\frac{\sum_{k=1}^n X_k}{\sqrt{n}} \leq x\right) \to \Phi(x) \Rightarrow \mathbb{E}[X_1^2] < \infty.$$

For arbitrarily large positive X, he considered sequences of random variables (X'_k) and (X''_k), where $X_k = X'_k + X''_k$ and

$$X'_k := \begin{cases} X_K & \text{if } |X_k| \leq X \\ 0 & \text{else.} \end{cases}$$

Furthermore, Lévy introduced the denotations $\epsilon := \mathbb{P}(|X_1| > X)$, $S'_n := \sum X'_k, S''_n := \sum X''_k, m := \sqrt{\mathrm{Var}(X'_1)}$. He made plausible that S'_n for sufficiently large n could be represented with an arbitrarily small error by a sum of $(1-\epsilon)\,n$ "nonzero" terms, and, accordingly, S''_n by a sum of $\epsilon\,n$, each distributed in the same way as X'_1. Lévy asserted (and proved) that the law of large numbers, after suitable norming, implied the convergence of the distribution of the sum S_n to the Gaussian law, so it was a sufficient condition for the CLT. He also showed that, in the special case of i.i.d. random variables under the assumption of the convergence to the normal distribution (the CLT thesis), they also obeyed the law of large numbers. This way, the law of large number is also a necessary condition for the CLT. His main assertion can be expressed as follows: let L_n be the dispersion of $\sum_{k=1}^n X_k$ assigned to an arbitrary, however fixed, probability $\gamma \in]0,1[$, There exist sequences $(a_n > 0)$ and (b_k) of real numbers such that

$$\mathbb{P}\left(\frac{1}{a_n}\sum_{k=1}^{n}(X_k - b_k) \leq x\right) \xrightarrow[n\to\infty]{} \Phi(x)$$

and

$$\max_{1\leq k\leq n} \mathbb{P}(|X_k| > \epsilon L_n) \to 0 \quad \forall \epsilon > 0$$

if and only if

$$\begin{gathered} \forall \delta > 0\ \forall \eta > 0\ \exists n(\delta,\eta) : \forall n \geq n(\delta,\eta),\ \exists X(n) > 0 : \\ \frac{(X(n))^2}{\sum_{k=1}^n \left(\int_{|x|\leq X(n)} x^2 \mathrm{d}V_k(x) - \left(\int_{|x|\leq X(n)} x \mathrm{d}V_k(x)\right)^2\right)} < \eta, \end{gathered} \tag{2.11}$$

and

$$\sum_{k=1}^{n} \mathbb{P}(|X_k| > X(n)) < \delta.$$

Feller

William Feller started working on probability theory around 1934 and used characteristic functions as his main tool for proving the theorem. He generally based his considerations on a sequence of distribution functions (V_n), all continuous on the right, and the respective convolution functions $W_n = V_1 * V_2 * \cdots * V_n$. He tried to address

the following problem: given a sequence of distribution functions $(V_n(x))$, do there exist two number sequences (a_n) and (c_n) such that $W_n(a_n x + c_n) \to \Phi(x)$, where $\Phi(x)$ is the standard normal distribution, and, if this is the case, how can such number sequences be determined? Feller presupposed the negligibility[5] of the V_k with respect to the total convolution W_n. He basically demanded that there exist suitable b_k such that, for each $x \neq 0$,

$$\max_{1\leq k\leq n} |V_k(a_n x + b_k) - E(x)| \xrightarrow[n\to\infty]{} 0,$$

where

$$E(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{else.} \end{cases}$$

This demand is equivalent to the condition that for the random variables X_k obeying the distributions V_k:

$$\max_{1\leq k\leq n} \mathbb{P}(|X_k - b_k| > \epsilon\, a_n) \xrightarrow[n\to\infty]{} 0 \quad \forall \epsilon > 0. \tag{2.12}$$

In his words, the sequence $(V_k(x+b_k))$ together with the norming factors (a_n) belongs to $\Phi(x)$ if the limit relation $W_n(a_n x + c_n) \to \Phi(x)$, with $c_n = \frac{1}{a_n}\sum_{k=1}^n b_k$ and the condition (2.12) are simultaneously met. The solution Feller found was: let (V_k) be a sequence of distributions, all with zero median. For each $\delta > 0$ let

$$p_n(\delta) := \min\left\{ r \in \mathbb{R}_0^+ \,\middle|\, \sum_{\nu=1}^n \int_{|x|>r} \mathrm{d}V_\nu(x) \leq \delta \right\}.$$

Then the presupposition

$$\forall \delta > 0 \quad \lim_{n\to\infty} \frac{1}{p_n^2(\delta)} \sum_{\nu=1}^n \int_{|x|\leq p_n(\delta)} x^2 \mathrm{d}V_\nu(x) = \infty$$

is necessary and sufficient for the existence of sequences $(a_n > 0)$, (b_k) such that the sequence $(V_k(x + b_k))$ together with the norming factors a_n belongs to $\Phi(x)$. Feller's main results, written in a more modern way in order to compare it with Lévy's work, was as follows. Let (X_k) be a sequence of independent random variables whose distributions V_k all have median 0. Then there exist sequences $(a_n > 0)$ and (b_k) of real numbers such that

$$\mathbb{P}\left(\frac{1}{a_n} \sum_{k=1}^n (X_k - b_k) \leq x \right) \xrightarrow[n\to\infty]{} \Phi(x)$$

and

$$\max_{1\leq k\leq n} \mathbb{P}(|X_k - b_k| > \epsilon\, a_n) \to 0 \quad \forall \epsilon > 0$$

[5] Given a function $\mu : \mathbb{N} \to [0,1]$, we say μ is negligible if for all polynomials p, there exists $n_0 \in \mathbb{N}$, such that $\forall n \geq n_0,\ \mu(n) \leq \frac{1}{p(n)}$.

if and only if

$$\forall \delta > 0 \; \forall \eta > 0 \; \exists n(\delta, \eta) : \forall n \geq n(\delta, \eta), \; \frac{p_n^2(\delta)}{\sum_{k=1}^{n} \int_{|x| \leq p_n(\delta)} x^2 \mathrm{d}V_k(x)} < \eta, \tag{2.13}$$

where $p_n(\delta) = \min\{r \in \mathbb{R}_0^+ | \mathbb{P}(|X_k| > r) \leq \delta\}$.

Despite the formal conformity of (2.11) and (2.13), a direct proof for the equivalence of these two conditions seems to be rather difficult. Still, the equivalence of Lévy's and Feller's assertions concerning the convergence to the normal distribution can be quite readily seen, but we do not report it in this thesis.

Chapter 3

The Central Limit Theorem today

As we have already seen, interest in limit theorems in Probability Theory originally arose from their statistical applications, though today there results have important uses in many other fields as well. Let us consider a random experiment, in a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and let us focus in a real-valued random variable $X : \Omega \to \mathbb{R}$ associated with it. A problem that comes naturally is determine, or better, estimate the distributions of X. Broadly speaking, if $X \in \mathcal{L}^1$, we use limit theorems to estimate its mean, i.e. $\mu = \mathbb{E}[X]$.

Remark 3.1. Note that, by estimating the mean of $X \in \mathcal{L}^1$, this means you can calculate $\mathbb{E}[f(X)]$, with $f : \mathbb{R} \to \mathbb{R}$ an arbitrary borel function such that $f(X) \in \mathcal{L}^1$. In particular, for all $B \in \mathscr{B}$, if $f = \mathbb{1}_B$, you obtain

$$\mathbb{E}[f(X)] = \mathbb{E}[\mathbb{1}_{\{X \in B\}}] = \mathbb{P}(X \in B) = \mathbb{P}^X(B).$$

Therefore knowing the expected values of $\mathbb{E}[f(X)]$ is equal to knowing the distribution of X.

Remark 3.2. Note that, in order to estimate the mean of any random variable $X \in \mathcal{L}^1$, you have to know how to estimate the probability of any event $A \in \mathcal{A}$. In fact, it is sufficient to choose $X = \mathbb{1}_A$, then $\mathbb{P}(A) = \mathbb{E}[\mathbb{1}_A]$.

Sequences of random variables and sample meaning

Let us find the mean μ of $X : \Omega \to \mathbb{R}$, a generic random variable in $\mathcal{L}^1$. A classical Statistic's procedure is to do a large number of repetitions of the same aleatoric experiment, taking notes of the value assumed by the random variable X, and then calculate its sample meaning. More theoretically, you suppose to repeat the aleatoric experiment infinitely many times, obtaining a sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$, corresponding to the hypothetical values assumed by X in each experiment. It is reasonable

to suppose $X_1, \ldots, X_n, \ldots$ independent and identically distributed (i.i.d.), or only that they each have mean equal to μ. Naturally, we consider the sample mean:

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

Limit theorems study the asymptotic behaviour of the sample meaning $\overline{X}_n$ for $n \to \infty$. In particular, the Law of large numbers establishes when $\overline{X}_n \to \mu$. The central limit theorem investigates the distribution of $\overline{X}_n$ for $n \to \infty$.

3.1 Law of Large Numbers

In this entire chapter, $(X_n)_n$ is a sequence of real-valued random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We assume that these variables are in $\mathcal{L}^1$ and have the same expected value μ:

$$\mathbb{E}[X_n] \coloneqq \mu, \ \forall n.$$

We define

$$\overline{X}_n \coloneqq X_1 + \cdots + X_n, \ \forall n.$$

Definition 3.3. *Let $(X_n)_n$ be a sequence of real-valued random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We assume that these variables are in $\mathcal{L}^1$ and have the same expected value: $\mathbb{E}[X_n] = \mu$, for every n.*

- *The sequence $(X_n)_n$ is said to satisfy the **weak law of large numbers** if*

$$\overline{X}_n \xrightarrow{P} \mu.$$

- *The sequence $(X_n)_n$ is said to satisfy the **strong law of large numbers** if*

$$\overline{X}_n \xrightarrow{a.s.} \mu.$$

A classic version of the weak law of large numbers comes directly from the Chebychev's inequality (1.15) is as follows. Let $(X_n)_n$ be a sequence of real-valued random variables in $\mathcal{L}^2$ with mean μ and variance σ^2. Suppose $X_1, \ldots, X_n, \ldots$ uncorrelated. Then

$$\overline{X}_n \xrightarrow{P} \mu.$$

In fact, $\mathbb{P}(|\overline{X}_n - \mu| \geq \lambda) \leq \frac{\sigma^2}{n\lambda^2}$, which converges to zero for $n \to \infty$.

Let us now consider two versions of the strong law of large numbers, which hold under the additional assumption that the random variables $X_1, X_2, \ldots, X_n, \ldots$ are i.i.d.. First, we state the strong law of large numbers due to Kolmogorov (1930).

Theorem 3.4 (**Strong law of large numbers by Kolmogorov**). *Let $(X_n)_n$ be a sequence of real-valued random variables i.i.d. in $\mathcal{L}^1$, with $\mathbb{E}[X_n] = \mu \; \forall n$. It is true that*

$$\overline{X}_n \xrightarrow{a.s., L^1} \mu.$$

Theorem 3.5 (**Strong law of large numbers**). *Let $(X_n)_n$ be a sequence of real-valued random variables i.i.d. in $\mathcal{L}^2$, with $\mathbb{E}[X_n] = \mu$. It is true that*

$$\overline{X}_n \xrightarrow{a.s., L^2} \mu.$$

Proof. Without loss of generality we can suppose $\mu = 0$. In fact, if $\mu \neq 0$, it suffices to consider $Y_i = X_i - \mu$ and note that

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu \quad \iff \quad \overline{Y}_n \xrightarrow{\text{a.s.}} 0,$$

where $\overline{Y}_n = \frac{Y_1 + \cdots + Y_n}{n}$.

Convergence in L^2. Now, let $\mu = 0$. Regarding the L^2-convergence of $(\overline{X}_n)_n$ towards $\mu = 0$, we first note that $\overline{X}_n \in \mathcal{L}^2$, since it is a linear combination of random variables in $\mathcal{L}^2$. It remains to show that

$$\mathbb{E}\left[|\overline{X}_n - 0|^2\right] = \mathbb{E}\left[\overline{X}_n^2\right] \xrightarrow{n\to\infty} 0. \tag{3.1}$$

To this end, we compute the mean and variance of $\overline{X}_n$. We have

$$\mathbb{E}[\overline{X}_n] = \frac{1}{n}\mathbb{E}[X_1] + \cdots + \frac{1}{n}\mathbb{E}[X_n] = \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = \mu = 0.$$

Moreover,

$$\mathrm{Var}(\overline{X}_n) \underset{X_1,\ldots,X_n indep.}{=} \mathrm{Var}\left(\frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n\right) = \frac{1}{n^2}\mathrm{Var}(X_1) + \cdots + \frac{1}{n^2}\mathrm{Var}(X_n) = \frac{\sigma^2}{n}$$

Since $\mathrm{Var}(\overline{X}_n) = \mathbb{E}[\overline{X}_n^2]$, the limit in (3.1) follows from the fact that $\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$.

Almost sure convergence. Given that $(\overline{X}_n)_n$ converges to zero in L^2, we know that there exists a subsequence that converges almost surely. However, this is not sufficient, as we must prove that the entire sequence converges almost surely. To this end, we construct a particular subsequence that converges almost surely. After that, we will consider the asymptotic behavior of all the other terms in the sequence. We divide the rest of the proof into two steps.

Step 1: subsequence converging to zero almost surely. Consider the subsequence $(\overline{X}_{n^2})_n$, we show that it converges to zero almost surely:

$$\overline{X}_{n^2} \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{3.2}$$

As previously seen, we have that $\mathbb{E}[\overline{X}_{n^2}] = \mathrm{Var}(\overline{X}_{n^2}) = \sigma^2/n^2$. Therefore

$$\sum_{n=1}^{\infty} \mathbb{E}[\overline{X}_{n^2}^2] = \sum_{n=1}^{\infty} \frac{\sigma^2}{n^2} < \infty.$$

Since $\overline{X}_{n^2}^2 \geq 0$, we can exchange the series with the expected value, obtaining

$$\mathbb{E}\left[\sum_{n=1}^{\infty} \overline{X}_{n^2}^2\right] < \infty.$$

This implies that $\sum_n \overline{X}_{n^2}^2 < \infty$ almost surely. The fact that the series is convergent implies that the general term of the series is infinitesimal:

$$\overline{X}_{n^2}^2 \xrightarrow[n\to\infty]{\text{a.s.}} 0,$$

which is equivalent to (3.2).

Step 2: the entire sequence converges to zero almost surely. For every natural number $n \geq 1$, let $p_n \in \mathbb{N}$ be such that

$$p_n^2 \leq n < (p_n + 1)^2. \tag{3.3}$$

We have that

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{p_n^2} X_i + \frac{1}{n}\sum_{i=p_n^2+1}^{n} X_i = \frac{p_n^2}{n}\overline{X}_{p_n^2} + \frac{1}{n}\sum_{i=p_n^2+1}^{n} X_i.$$

Thus,

$$\overline{X}_n - \frac{p_n^2}{n}\overline{X}_{p_n^2} = \frac{1}{n}\sum_{i=p_n^2+1}^{n} X_i. \tag{3.4}$$

Now we show that the series with general term $\overline{X}_n - \frac{p_n^2}{n}\overline{X}_{p_n^2}$ is almost surely finite. We proceed as in Step 1, starting by considering the series of expected values:

$$\sum_{n=1}^{\infty} \mathbb{E}\left[\left(\overline{X}_n - \frac{p_n^2}{n}\overline{X}_{p_n^2}\right)^2\right] \overset{(3.4)}{=} \sum_{n=1}^{\infty} \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=p_n^2+1}^{n} X_i\right)^2\right] = \sum_{n=1}^{\infty} \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=p_n^2+1}^{n} X_i\right)$$

$$= \sum_{n=1}^{\infty} \frac{1}{n^2}\left(\sum_{i=p_n^2+1}^{n} \mathrm{Var}(X_i)\right) = \sum_{n=1}^{\infty} \frac{n - p_n^2}{n^2}\sigma^2 \overset{(3.3)}{\leq} \sum_{n=1}^{\infty} \frac{(p_n+1)^2 - p_n^2}{n^2}\sigma^2$$

$$= \sum_{n=1}^{\infty} \frac{2p_n + 1}{n^2}\sigma^2 \overset{(3.3)}{\leq} \sum_{n=1}^{\infty} \frac{2\sqrt{n} + 1}{n^2}\sigma^2 \overset{1\leq\sqrt{n}}{\leq} \sum_{n=1}^{\infty} \frac{3}{n^{3/2}}\sigma^2 < \infty.$$

Since the series of expected values is convergent, reasoning as in Step 1, we deduce that $\sum_n \left(\overline{X}_n - \frac{p_n^2}{n}\overline{X}_{p_n^2}\right) < \infty$ almost surely, so the general term must be infinitesimal:

$$\overline{X}_n - \frac{p_n^2}{n}\,\overline{X}_{p_n^2} \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

Now, from Step 1 we know that $\overline{X}_{p_n^2} \xrightarrow[n\to\infty]{\text{a.s.}} 0$. Furthermore, using (3.3), we have that $\frac{p_n^2}{n} \to 1$. We conclude that

$$\overline{X}_n \xrightarrow{\text{a.s.}} 0.$$

□

3.2 Central Limit Theorem

Let $(X_n)_n$ be a sequence of real random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. There are several versions of the Central Limit Theorem. This is the classical one, with the usual assumptions $X_1, X_2, \ldots, X_n, \ldots$ i.i.d. and in $\mathcal{L}^2$. Let

$$\mathbb{E}[X_n] = \mu \quad \text{Var}(X_n) = \sigma^2.$$

Then

$$\mathbb{E}[\overline{X_n}] = \mu \quad \text{Var}(\overline{X_n}) = \frac{\sigma^2}{n}.$$

Under the assumption $\sigma > 0$ (for the case $\sigma = 0$ see (3.8)), the central limit theorem provides information about the distribution of the *standardized sample mean*:

$$\overline{Z_n} = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}}, \quad \forall n.$$

Theorem 3.6 (**Central Limit Theorem**). *Let $(X_n)_n$ be a sequence of real-valued random variables i.i.d. in $\mathcal{L}^2$, with $\mathbb{E}[X_n] = \mu$ and Var$(X_n) = \sigma^2$ $\forall n$. If $\sigma > 0$, we have*

$$\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1). \tag{3.5}$$

Remark 3.7. We can rewrite (3.5) as follows:

$$\frac{\overline{X_n} - \mu}{1/\sqrt{n}} \xrightarrow{d} Y \sim \mathcal{N}(0, \sigma^2).$$

This means that for a large n

$$\overline{X_n} \approx \mu + \frac{1}{\sqrt{n}} Y.$$

So $\overline{X_n} = \frac{X_1 + \cdots + X_n}{n}$ has approximately law $\mathcal{N}(\mu, \frac{\sigma^s}{n})$.

Remark 3.8. Note that if $\sigma = 0$ then $X_n = \mu$ a.s. for every n, so $\overline{X_n} = \mu$ a.s., for every n. In this case it follows directly that

$$\frac{\overline{X_n} - \mu}{1/\sqrt{n}} \xrightarrow{d} Y \sim \mathcal{N}(0, 0).$$

Proof. To demonstrate convergence in distribution, we use Lévy's continuity theorem[1]. Let φ denote the characteristic function of $X_n - \mu$ (which does not depend on n, since $X_1, \ldots, X_n, \ldots$ are identically distributed). Now, let's calculate the characteristic function of

$$\overline{Z}_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} (X_k - \mu).$$

For any $u \in \mathbb{R}$, we have

$$\varphi_{\overline{Z}_n}(u) = \mathbb{E}\left[e^{iu\frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n}(X_k-\mu)}\right] \overset{\text{indep.}}{=} \prod_{k=1}^{n} \mathbb{E}\left[e^{iu\frac{1}{\sigma\sqrt{n}}(X_k-\mu)}\right] = \left(\varphi\left(\frac{u}{\sigma\sqrt{n}}\right)\right)^n. \tag{3.6}$$

Recall that $X_k \in \mathcal{L}^2$, so $X_k - \mu \in \mathcal{L}^2$. Therefore, the characteristic function $\varphi \in C^2(\mathbb{R})$. In particular, the following second-order Taylor expansion holds:

$$\varphi(u) = \varphi(0) + \varphi'(0)u + \frac{1}{2}\varphi''(0)u^2 + o(u^2), \quad \text{for } u \to 0.$$

Moreover, we recall that

$$\varphi(0) = 1, \quad \varphi'(0) = i\mathbb{E}[X_k - \mu] = 0, \quad \varphi''(0) = -\mathbb{E}[(X_k - \mu)^2] = -\mathrm{Var}(X_k) = -\sigma^2.$$

Thus,

$$\varphi(u) = 1 - \frac{\sigma^2}{2}u^2 + o(u^2), \quad \text{for } u \to 0. \tag{3.7}$$

Using (3.6) in (3.7), we obtain

$$\varphi_{\overline{Z}_n}(u) = \left(1 - \frac{1}{2}\frac{u^2}{n} + o\left(\frac{u^2}{n}\right)\right)^n \xrightarrow[n\to\infty]{} \mathrm{e}^{-\frac{1}{2}u^2}.$$

We have thus demonstrated that the characteristic function of $\overline{Z}_n$ converges to the characteristic function of a random variable $Z \sim \mathcal{N}(0,1)$. By Lévy's continuity theorem, we conclude that $\overline{Z}_n \xrightarrow{d} Z$. □

Remark 3.9 (**Convergence rate of $\overline{X_n}$ towards μ**). The law of large numbers states that $\overline{X}_n$ converges to μ. To find the convergence rate is equal to ask whether there exists $\alpha > 0$ such that

$$n^\alpha \left(\overline{X}_n - \mu\right) \longrightarrow Y \neq 0. \tag{3.8}$$

If we interpret the convergence mentioned above as convergence in law, the central limit theorem states that (3.8) holds with $\alpha = \frac{1}{2}$. In other words, the rate of convergence of $\overline{X}_n$ towards μ is of the order of $\frac{1}{\sqrt{n}}$.

[1] **Lévy's continuity theorem.** Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of real distributions and let $(\varphi_n)_{n\in\mathbb{N}}$ be the sequence of the corresponding characteristic functions. It holds: i) if $\mu_n \xrightarrow{d} \mu$ then μ pointwise converges to the CHF φ of μ, that is $\varphi_n(\eta) \xrightarrow[n\to\infty]{} \varphi(\eta)$ for all $\eta \in \mathbb{R}$; ii) conversely, if φ_n pointwise converges to a function φ continuous in 0, then φ is the CHF of a distribution μ and it holds $\mu_n \xrightarrow{d} \mu$.

Remark 3.10 (**Why does Z has normal distribution?**). Suppose we only know that $\overline{Z}_n$ converges in distribution to some random variable Z (whose distribution is still unknown). Now, consider the following subsequences of the sequence $(X_n)_n$:

$$X_n^p := X_{2n}, \quad X_n^d := X_{2n+1},$$

that is, the subsequence of the random variables with even indices and the subsequence with odd indices. Let $\overline{Z}_n^p$ and $\overline{Z}_n^d$ be the corresponding standardized sample means. By the central limit theorem, both $(\overline{Z}_n^p)_n$ and $(\overline{Z}_n^d)_n$ also converge in distribution to a standard normal random variable:

$$\overline{Z}_n \xrightarrow[n\to\infty]{d} Z, \qquad \overline{Z}_n^p \xrightarrow[n\to\infty]{d} Z^p, \qquad \overline{Z}_n^d \xrightarrow[n\to\infty]{d} Z^d,$$

with Z, Z^p, and Z^d having a $\mathcal{N}(0,1)$ distribution. Note that the random variables Z, Z^p, and Z^d may be defined on different probability spaces; the only requirement is that they follow a $\mathcal{N}(0,1)$ distribution. However, for future convenience, we assume they are defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and that Z^p and Z^d are independent. Now, we note that

$$\overline{Z}_{2n} = \frac{\overline{Z}_n^p + \overline{Z}_n^d}{\sqrt{2}}.$$

We know that

$$\overline{Z}_{2n} \xrightarrow[n\to\infty]{d} Z.$$

Furthermore, since $\overline{Z}_n^p$ and $\overline{Z}_n^d$ are independent, just as Z^p and Z^d are, it can be shown that

$$\frac{\overline{Z}_n^p + \overline{Z}_n^d}{\sqrt{2}} \xrightarrow[n\to\infty]{d} \frac{Z^p + Z^d}{\sqrt{2}}.$$

So the random variables

$$Z \qquad \text{and} \qquad \frac{Z^p + Z^d}{\sqrt{2}}$$

have necessarily the same distribution. Then the thesis follows from the remark below.

Remark 3.11. Let $Z \in \mathcal{L}^2$. Let also X and Y be random variables defined on the same probability space, independent and distributed as Z. If the random variables Z and $\frac{X+Y}{\sqrt{2}}$ have the same law, then $Z \sim \mathcal{N}(0, \sigma^2)$, for some $\sigma \geq 0$.

Proof. We note that $\mathbb{E}[Z] = 0$ and $Z \sim \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$, with $(X_i)_i$ a sequence of i.i.d. random variables. In fact the characteristic function of Z is $\varphi_Z(\eta) = \mathrm{e}^{-\frac{\sigma^2\eta^2}{2}}$, which is equal to

$$\varphi_{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i}(\eta) = \mathbb{E}\left[\mathrm{e}^{i\eta\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i}\right] = \mathbb{E}\left[\prod_{i=1}^n \mathrm{e}^{i\eta\frac{1}{\sqrt{n}}X_i}\right] = \left(\mathrm{e}^{i\eta\frac{1}{\sqrt{n}}X_i}\right)^n$$

by imposing $n := 2^m$. Then, we conclude by using the CLT.

Remark 3.12 (**Theorem of Berry-Esseen**). From the central limit theorem we know that

$$\overline{Z_n} = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1). \tag{3.9}$$

Recall that convergence in distribution is equivalent to pointwise convergence of the cumulative distribution functions at the continuity points of the limiting cumulative distribution function. In this case, the limiting cumulative distribution function is continuous everywhere, as the cumulative distribution function of Z is given by

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \mathrm{d}z, \quad \forall x \in \mathbb{R}.$$

So (3.9) is equal to

$$\lim_{n\to\infty} F_{\overline{Z}_n}(x) = \Phi(x), \quad \forall x \in \mathbb{R}.$$

The Berry-Esseen theorem provides a more precise estimate of the convergence of $F_{\overline{Z}_n}$ towards Φ. This is a particularly advanced result, for which we do not provide the proof.

Theorem 3.13 (Berry-Esseen Theorem)**.** *Under the same hypothesis of* (3.5)*, suppose also* $X_1 \in \mathcal{L}^3$*, then*

$$|F_{\overline{Z}_n}(x) - \Phi(x)| \leq C \frac{\mathbb{E}[|X_1|^3]}{\sigma^3\sqrt{n}},$$

for some $C > 0$*. An open problem is the determination of the optimal constant* C*. Currently, it is known that* $0.4097 < C < 0.4748$*.*

Finally, let us see a multidimensional variant of the central limit theorem (without demonstration).

Theorem 3.14 (**Vector central limit theorem**)**.**
Let $(X^{(n)})_n$*, with* $X^{(n)} = (X_1^{(n)}, \cdots, X_d^{(n)})$*, a succession A sequence of i.i.d. random vectors in* $\mathbb{R}^d$ *in* $\mathcal{L}^2$*. Let* $\mu = \mathbb{E}[X^{(n)}] \in \mathbb{R}^d$, $Q = \mathrm{Var}(X^{(n)})) \in \mathbb{R}^{d\times d}$*, and*

$$\overline{X}_n = \frac{X^{(1)} + \cdots + X^{(n)}}{n},$$

then it holds that

$$\frac{\overline{X}_n - \mu}{1/\sqrt{n}} \xrightarrow{d} Y \sim \mathcal{N}(0, Q).$$

Bibliography

[1] H. Fischer: "A History of the Central Limit Theorem". First edition. Springer: New York, 2010.

[2] A. Pascucci: "Teoria della Probabilità". Springer: Milano, 2020. Available at: https://link.springer.com/book/10.1007/978-88-470-4000-7

Ringraziamenti

Alla fine di questo elaborato, mi sembra necessario riservare un piccolo spazio per ringraziare tutte le persone che mi hanno sostenuto durante questo percorso.
Per prima cosa vorrei ringraziare il mio relatore Andrea Pascucci, per la disponibilità e il suo aiuto nella stesura di questa tesi.
Grazie ai miei genitori e tutta la mia famiglia per esserci sempre, per avermi permesso di studiare senza mettermi fretta e continuare a farlo.
Grazie alla mia Rossi per avermi accompagnata dalla prima liceo in ogni circostanza della mia vita, sei fondamentale per me e un grande conforto in qualsiasi momento.
Grazie Peppe, per quello che sei ora e per quello che sei stato dal primo giorno in cui ti ho conosciuto, so che posso sempre contare su di te. Amarti è così facile.
A Fra, grazie per essere una persona così solare e per avermi accolto (o un po' trascinato, senza insistere mai), in un gruppo tutto nuovo di bellissime persone.
Grazie a tutte le mie amichette di nuoto per aver alleggerito il primo periodo fuori casa e in particolare grazie a te Cice per essere un'altra piccola sorellina.
Vorrei inoltre ringraziare il mio prof del liceo, Alberto Serra, senza il quale non sarei mai finita ad iscrivermi a questa facoltà, e Foca, che mi ha accompagnata in questo percorso non facile.
Grazie anche ai miei amici matematici e di unione, per aver alleggerito questi tre anni intensi; alle mie coinquiline e mia sorella, per gli scleri giornalieri; e infine grazie a tutti gli amici di Forlì che quelle poche volte che torno mi fanno sentire sempre a casa. Vi voglio bene.