A History of the Central Limit Theorem

July 31, 2019

Seminararbeit

Fanni Plenar 1630098

Technical University of Vienna

Contents

1 Introduction

The "central limit theorem", CLT, is a collective term for theorems about the convergence of distributions, densities or discrete probabilities. The term itself was first used by George Pólya, in his article from 1920.

The most well-known version of the CLT is about the convergence of the normed sums of (X_k), a sequence of independent and identically distributed random variables on a common probability space with expectations a_k.
Define we $b_n = \mathrm{Var} \sum_{k=1}^{n} X_k$.

$$r \in R : P(\frac{\sum_{k=1}^{n}(X_k - a_k)}{\sqrt{b_n}} \leq r) \to \Phi(r) \text{ for } n \to \infty,$$

where $\Phi(r)$ is the distribution function of the standard normal distribution

$$\Phi(r) = \int_{-\infty}^{r} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

The CLT had a very long history, until it get his place in mathematics. In the next pages we will learn about how it changed between 1810 and 1935.

But, before I start, we need to talk about Abraham de Moivre's approximations to binomial distributions, even if it doesn't fit the characterization of the CLT, it still had an impact on the later approaches.

1.1 De Moivre' approximation

In 1733, De Moivre found an approximation to binomial distributions.

De Moivre wanted to find an approximation to $P(|Z - [\frac{n}{2}]| \leq t)$ which is the same as $\sum_{i=-t}^{t} P(Z = [\frac{n}{2}]+i)$ for a large number of n fair trials.

In his work he used Jakob Bernoulli's "Law of Large Numbers", where Bernoulli showed that for n identical and independent trials, if h_n is the relative frequency of a perticular event occuring with the probability p then

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

De Moivre needed an approximation for $P(Z = [\frac{n}{2} + i)$ which is the probability of $[\frac{n}{2}] + i$ "successes" for a large number of n fair trials, where the fairness of the trials means that $p = \frac{1}{2}$. So he started to work with

$$P(Z = [\tfrac{n}{2}] + i) = 2^{-n} \begin{pmatrix} n \\ [\frac{n}{2}]+i \end{pmatrix}.$$

First he approximated

$$\frac{\begin{pmatrix} n \\ [\frac{n}{2}] \end{pmatrix}}{2^n} \approx \frac{2}{\sqrt{2\pi n}} \text{ and } log\frac{\begin{pmatrix} n \\ [\frac{n}{2}]+i \end{pmatrix}}{\begin{pmatrix} n \\ [\frac{n}{2}] \end{pmatrix}} \approx -2\frac{i^2}{n}$$

That follows:

$$P(Z = [\tfrac{n}{2}] + i) \approx \tfrac{2}{\sqrt{2\pi n}} e^{-2\frac{i^2}{n}}$$

This equality could be considered as a local limit theorem, but that wasn't de Moivre's main goal. It was to find an approximation

$$P(|Z - [\tfrac{n}{2}]| \leq t) \approx 2\tfrac{2}{\sqrt{2\pi n}} \sum_{i=0}^{t} e^{-2\frac{i^2}{n}} \approx \tfrac{4}{\sqrt{2\pi n}} \int_0^t e^{-2\frac{x^2}{n}} dx.$$

2 The beginning of the History

The history of the CLT starts with Pierre-Simon Laplace, who didn't have a concrete theorem and mostly used his approach to the CLT as a tool to solve other mathematical problems. A few author tried to dicuss Laplace's work for example Robert Leslie Ellis in 1844. In 1856 Anton Meyer even presented a proof for the special case of the CLT, for two-valued random variables, his paper was accepted for publication, but the publication failed and Meyer died in a short time. An other author, who had an influence on later authors was Siméon Denis Poisson.

Later Peter Gustav Lejeune Dirichlet and Augustin Louis Cauchy both published articles, which could be considered as a proof of the CLT.

In this stage of its history, it was connected to error theory. It wasn't a mathematical problem of its own, the authors mostly used it as a tool to solve other problems.

2.1 Laplace

Laplace's work in probability theory is really important. He published his "Théorie analytique des probabilitiés"(TAP) in 1812, which includes typical problems,stochastic models, and analytic methods.

Laplace worked with sums of independent random variables since the beginning. He also developed the "Laplacian method" for approximating integrals. His basic idea was, that if $f(x)$ depends on a very large parameter such that the function f has a single, very sharp peak and only a small interval around this maximum results as appreciable for the integral, then f asymptotically equal to a function $f(a)e^{-\alpha(x-a)^{2k}+\cdots}$, if f has its maximum at $x = a$. Laplace used this method for example in the case of the Gamma function.

He had his first approach to the CLT in 1810, after almost forty years work, but he didn't state a theorem in his work. We can demonstrate his approach to the CLT in the special case of identically distributed random variables $X_1 \dots X_n$, although he worked with errors of observation, presupposing they are mutual independent, with $\forall j$: $EX_j = 0$ and $P(X_j = \frac{k}{m}) = p_k$, for $m \in N$ $k \in \{-m, -m+1, \dots, m-1, m\}$, to calculate

$$P_j := P(\textstyle\sum_{l=1}^{n} X_l = \frac{j}{m}) \text{ for } j \in \{-nm, -nm+1, \dots, nm-1, nm\}.$$

Laplace used the generating function $T(t) = \sum_{k=-m}^{m} p_k t^k$, where P_j is equal to the coefficient of t^j after the multiplication of $[T(t)]^n$. But he used a trick, he worked with e^{ix}, where $i = \sqrt{-1}$, instead of t. Then from the introduction of a special case of characteristic functions:

$$\tfrac{1}{2\pi} \textstyle\int_{-\pi}^{\pi} e^{-itx} e^{isx} dx = \delta_{ts} \ (t, s \in (Z)),$$

follows that the coefficient to t^j is:

$$P_j = \tfrac{1}{2\pi} \textstyle\int_{-\pi}^{\pi} e^{-ijx} \Big[\sum_{k=-m}^{m} p_k e^{ikx} \Big]^n dx.$$

We know that $e^{ikx} = \sum_{l=0}^{\infty} \frac{(ikx)^l}{l!}$, what means

$$\sum_{k=-m}^{m} p_k e^{ikx} = \sum_{l=0}^{\infty} (ix)^l \sum_{k=-m}^{m} p_k \frac{k^l}{l!}.$$

$\sum_{k=-m}^{m} p_k k = 0$, since the expectation is 0 . Then we define m such that $m^2\sigma^2 = \sum_{k=-m}^{m} p_k k^2$, the other terms $i^l x^l$ also have constant coefficients $A_l \ \forall l \in \{3, 4, \ldots\}$, so we get:

$$P_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ijx} \Big[1 - \frac{m^2\sigma^2x^2}{2} + \sum_{l=3}^{\infty} A_l (ix)^l\Big]^n dx$$

Then we can find z(x), such that:

$$log z(x) = log[1 - \frac{m^2\sigma^2x^2}{2} + \sum_{l=3}^{\infty} \frac{(ikx)^l}{l!}]^n$$

$$z(x) = e^{-\frac{nm^2\sigma^2x^2}{2}} (1 + \sum_{l=3}^{\infty} \frac{n(ikx)^l}{l!}).$$

And with $z(x)$

$$P_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itx} z(x) dx,$$

where if we consider $y = \sqrt{n}x$, then we get:

$$P_j = \frac{1}{2\pi\sqrt{n}} \int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} e^{-ij\frac{y}{\sqrt{n}}} e^{-\frac{m^2\sigma^2y^2}{2}} (1 + \sum_{l=3}^{\infty} \frac{(iky)^l}{\sqrt{n}^{l-2} l!}) dy,$$

that means, if $n \to \infty$, then

$$P_j \approx \frac{1}{2\pi\sqrt{n}} \int_{-\infty}^{\infty} e^{-ij\frac{y}{\sqrt{n}}} e^{-\frac{m^2\sigma^2y^2}{2}} dy = \frac{1}{m\sigma\sqrt{2\pi n}} e^{\frac{j^2}{2m^2\sigma^2n}}.$$

The last equality was showed by Laplace.

This can be used to find $P(r_1\sqrt{n} \leq \sum X_l \leq r_2\sqrt{n})$, which can be approximated as the sum of $P(\sum X_l = \frac{j}{m})$ for all $\frac{j}{m} \in [r_1\sqrt{n}; r_2\sqrt{n}]$, what could be approximated with integration, like at de Moivre's distribution:

$$P(r_1\sqrt{n} \leq \sum X_l \leq r_2\sqrt{n}) \approx \int_{r_1}^{r_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx.$$

So we became the integral form of the CLT.

In his work Laplace trusted in the power of series expansions, and didn't determine the errors of approximations. For Laplace the CLT wasn't a mathematical problem itself, but a tool, what could solve other problems, for example:

- The comet problem

  At this problem he observed the "randomness" of 97 comets. He used the CLT to calculate the probability of all angles of inclination falls within a certain interval.

- The problem of foundation method of least squares

  He used the CLT at this problem too, but his arguments were only valid for an "infinitely large" number of observation, which in this case was unrealistic.

- The problem of risk in the game of chance

  Here Laplace, with the help of the CLT, dealt with a sequence of games, each with two possible outcomes "gain and "loss".

2.2 Poisson

Poisson wrote two article about the CLT, one in 1824 and the other in 1829. His work on the CLT was important for two main reasons. Firstly, he created a new concept "choses", which could be an early form of random variables, and used it to formulate and prove his theorems. Secondly, he also used counterexamples to discuss the validity of his theorem.

In his version of the CLT he considered X_1,...X_s to be a great number of choses, whose density functions f_n decrease sufficiently fast. He also supposed that for the absolute values $\rho_n(\alpha)$ of the characteristic functions of X_n, which he defined

$$\rho_n(\alpha)cos\phi_n := \int_a^b f_n(x)cos(\alpha x)dx \text{ and } \rho_n(\alpha)sin\phi_n := \int_a^b f_n(x)sin(\alpha x)dx,$$

there exist a function $r(\alpha)$ independent of n with $0 \leq r(\alpha) < 1$ $\forall \alpha \neq 0$ and it is valid that

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

Then for arbitrary γ_1, γ_2,

$$P\Big(\gamma_1 \leq \frac{\sum_{n=1}^{s}(X_n - EX_n)}{\sqrt{2\sum_{n=1}^{s} VarX_n}} \leq \gamma_2\Big) \approx \frac{1}{\sqrt{\pi}} \int_{\gamma_1}^{\gamma_2} e^{-u^2} du.$$

The difference between the two side tends to zero if s tends to infinity.

As we can see Poisson used the distribution function of a normal distribution with expectation 0 and variance $\frac{1}{2}$. If we would like to make his approximation a little more familiar, with the standard normal distribution, we can reform the left side of the approximation, using $u = \frac{v}{\sqrt{2}}$:

$$\frac{1}{\sqrt{\pi}} \int_{\gamma_1}^{\gamma_2} e^{-u^2} du = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{2}\gamma_1}^{\sqrt{2}\gamma_2} e^{\frac{-v^2}{2}} dv.$$

And we become the CLT in a more familiar form:

$$P\Big(\gamma_1\sqrt{2} \leq \frac{\sum_{n=1}^{s}(X_n - EX_n)}{\sqrt{\sum_{n=1}^{s} VarX_n}} \leq \gamma_2\sqrt{2}\Big) \approx \frac{1}{\sqrt{2\pi}} \int_{\sqrt{2}\gamma_1}^{\sqrt{2}\gamma_2} e^{\frac{-v^2}{2}} dv.$$

Poisson also believed that his CLT could be used also for discrete random variables.

2.3 Dirichlet

In 1846, Dirichlet discussed linear combinations of random errors, this discusson could be a rigorous proof of the CLT.

The discussed errors were considered to have symmetric densities, which are concentrated on a fixed interval $[-a, a]$, this also assumes that the expectations are zero. He also presupposed that for the linear combination $\alpha_1 x_1 + \ldots + \alpha_n x_n$ the sequence of α_v has a positive lower and a positive upper bound. For his proof, to be useful for non-identically distributed observational errors too, there had to be a C, for which was valid that $\forall x \in [-a; a]$: $C > |f_v'(x)|$ for every density functions f_v.

His main result was:

$$\Big|P\Big(-\lambda\sqrt{n} \le \sum_{v=1}^{n} \alpha_v x_v \le \lambda\sqrt{n}\Big) - \frac{2}{\sqrt{\pi}} \int_0^{\frac{\lambda}{r}} e^{-s^2} ds\Big| \to 0 \ (n \to 0)$$

where

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

he defined k_v in his proof as:

$$k_v := \frac{1}{2} \int_{-a}^{a} z^2 f_v(z) dz$$

So he found a limit for the error of the approximation, which was actually far from the optimal, but it also wasn't his intention. He just wanted to show that his modification of the Laplace's method of approximation could be used also to calculate the probabilities of linear combination of random errors.

As we can see, in his formula he uses the integral form of the CLT with a few differences between this and the modern form of the CLT, where we usually use the normal distribution with variance 1. But if we would divide the sum of errors with $(\frac{1}{2}r\sqrt{n})$ and in the integral, which could be defined also for $s \in [-\frac{\lambda}{r}, \frac{\lambda}{r}]$, because of the symmetric densities, use $\frac{x}{\sqrt{2}}$ instead of s, what also means that it's defined for $x \in [-\frac{\sqrt{2}\lambda}{r}; \frac{\sqrt{2}\lambda}{r}]$, we would get a more familiar form of the CLT:

$$\Big|P\Big(\frac{-2\lambda}{r} \le \frac{\sum_{v=1}^{n} \alpha_v x_v}{\sqrt{\sum_{v=1}^{n} k_v \alpha_v^2}} \le \frac{2\lambda}{r}\Big) - \frac{1}{\sqrt{2\pi}} \int_{\frac{-\sqrt{2}\lambda}{r}}^{\frac{\sqrt{2}\lambda}{r}} e^{\frac{-x^2}{2}} dx\Big| \to 0 \text{ for } (n \to 0)$$

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

2.4 Cauchy

In 1853, Cauchy established upper bounds for the error of a normal approximation to the distribution of a linear combination of identically distributed independent errors. He wrote it in a discussion with Bienaymé on least squares.

His conditions were similar as Dirichlets. So the errors ϵ_j had symmetric densities f_j, which vanished for arguments beyond the compact interval $[-k; k]$. He added that for the linear combination $\sum_{j=1}^{n} \lambda_j \epsilon_j$ should be valid that λ_j should have the "order of magnitude" of $\frac{1}{n}$ or less, which means:

$$\exists \alpha, \beta > 0 \text{ independent of n such that } \forall j \in \{1, \ldots, n\} \exists \gamma(j) \ge 1 \text{ with}$$
$$\alpha \le n^{\gamma(j)} |\lambda_j| \le \beta,$$

and $\Lambda := \sum \lambda_j^2$ should be of order $\frac{1}{n}$.

Cauchy used the notation $c := \int_0^k x^2 f(x) dx$ and he get for $v > 0$:

$$\Big|P\Big(-v \le \sum_{i=1}^{n} \lambda_i \epsilon_i \le v\Big) - \frac{2}{\sqrt{\pi}} \int_0^{\frac{v}{2\sqrt{c\Lambda}}} e^{-\theta^2} d\theta\Big| \le C_1(n) + C_2(n, v) + C_3(n)$$

where the functions C_1, C_2 and C_3 tends to 0 if n increases.

We can get upper bounds for the absolute error of the approximation of the CLT If we consider a sequence of independent random variables X_j, distributed like the errors before, and $\lambda_j = \frac{1}{n}$, $v = \frac{a}{\sqrt{n}}$ $(a > 0)$, $c = \frac{1}{2} Var X_1$:

$$\Big|P\Big(-a\sqrt{n} \leq \sum_{i=1}^{n} X_i \leq a\sqrt{n}\Big) - \frac{2}{\sqrt{\pi}} \int_0^{\frac{a}{2\sqrt{c}}} e^{-x} dx\Big| \leq C_1(n) + C_2(n, \tfrac{a}{\sqrt{n}}) + C_3(n) \to 0$$
for $n \to \infty$.

We can also, like at Dirichlet's case, form a more familiar formula with the standard normal distribution, if we divide the sums of random variables with $(\sqrt{2nc})$, and in the integral we use $y := x\sqrt{2}$. So we get the formula:

$$\Big|P\Big(\frac{-a}{\sqrt{2c}} \leq \frac{\sum_{i=1}^{n} X_i}{\sqrt{nVarX_1}} \leq \frac{a}{\sqrt{2c}}\Big) - \frac{1}{\sqrt{2\pi}} \int_{-\frac{a}{\sqrt{2c}}}^{\frac{a}{\sqrt{2c}}} e^{\frac{-y}{2}} dy\Big| \to 0 \text{ for } n \to \infty.$$

And since we consider the errors to be independent and identically distributed, we can consider $\sqrt{nVarX_1}$ as the variance of $\sum_{i=1}^{n} X_i$.

3 The founders of "St. Peterburg school"

The founders of "St. Peterburg school", especially Pafnutii Lvovich Chebyshev, Andrei Andreevich Markov, and Aleksandr Mikhailovich Ljapunov, all had an influence on the history of the CLT.

Chebyshev and Markov both worked with moments, and in their work they both used the CLT to illustrate their methods in moment theory, while Ljapunov worked with it as a mathematical object of its own and he was the first, who rigorously proved the CLT.

3.1 Chebychev

In 1887, Chebyshev published an article with an uncomplete proof of the CLT, where the method he used was somewhat different from the authors before him. The French translation of this article was published three years later, in 1890.

In his work he presented the CLT in the following form:
Let u_i be a sequence of "independent quantités" with zero expectations and nonnegative densities ϕ_i, also with moments of arbitrary high order. Under the assumption that for each order for all "quantités" an upper and a lower bound of the moments existed , he stated that $\forall t_1 < t_2 \in R$:

$$\lim_{n\to\infty} P\Big(t_1 \leq \frac{\sum_{i=1}^{n} u_i}{\sqrt{2\sum_{i=0}^{n} Eu_i^2}} \leq t_2\Big) = \frac{1}{\sqrt{\pi}} \int_{t_1}^{t_2} e^{-x^2} dx.$$

As we can see, he also used the distribution function of a normal distribution with variance $\frac{1}{2}$, but if we use in the integral $y = \sqrt{x}$ and make a few changes in the probability, then we become the CLT in the well-known form, what we mostly use today. To make it a little less complicated we can define $r_1 := \sqrt{2}t_1$ and $r_2 := \sqrt{2}t_2$, then with these changes we get:

$$\lim_{n\to\infty} P\Big(r_1 \leq \frac{\sum_{i=1}^{n} u_i}{\sqrt{\sum_{i=0}^{n} Eu_i^2}} \leq r_2\Big) = \frac{1}{\sqrt{2\pi}} \int_{r_1}^{r_2} e^{-y^2} dy.$$

Actually we can consider $\sqrt{\sum_{i=0}^{n} Eu_i^2}$ as the root of the variance of $\sum_{i=0}^{n} u_i$, since they are independent and for all j is valid that $Eu_j = 0$ what means $Varu_j = Eu_j^2$.

Chebychev didn't proved the CLT rigorously, but his theorem is still important. One of the two reason for its importance that he stated his theorem for "quantités" and not for errors as the other authors before him. The other is that he explicitly stated conditions for the validity of the assertion and so he was the first to expressed the CLT as a limit theorem proper.

3.2 Markov

Although Markov became Chebychev's successor in teaching probability theory in 1882, he wasn't too active in this field and only around 1898 started to work on a moment theoretic proof of the CLT. Actually his proof of the CLT was just a corollary of more general moment theoretic results.

He wrote an article in 1898, where he definied the CLT for "independent quantities" $u_1, u_2 \ldots$. He stated three conditions which these these quantitiés obeyed .
Firstly, they had to have zero expectations.
Secondly, there had to be a constant C_m $\forall m$ such that $|Eu_k^m| < C_m$ $\forall k \in N$
And lastly, Eu_k^2 had to have a positive lower bound.
And he got $\forall \alpha < \beta$:

$$P\Big(\alpha\sqrt{2\sum_{i=0}^{n} Eu_i^2} \leq \sum_{i=1}^{n} u_i \leq \beta\sqrt{2\sum_{i=0}^{n} Eu_i^2}\Big) \to \frac{1}{\sqrt{\pi}} \int_{\alpha}^{\beta} e^{-x^2} dx.$$

As we can see this is the same form as the one that Chebychev used, with a little difference in the conditions. They both considered upper and lower bounds for the moments in each order, but Markov presupposed, in his third condition, that Eu_k^2 doesn't tend to 0 if k grows.

In his article he didn't state a complete proof about the convergence of the moments of the normed sums to the normal distribution, which would be important for his approach to the CLT. He proved that in a letter exchange with Vasilev, this proof was published in 1899. The main result of this theorem was that under particular conditions:

$$\Big(\frac{\sum_{i=1}^{n}(X_i - EX_i)}{\sqrt{2\sum_{i=1}^{n}\sigma_i^2}}\Big)^m \to \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^m e^{-t^2} dt.$$

In his earlier works the CLT wasn't an independent research subject, it was mostly a corollary to other moment theoretic results. But Ljapunov's proof of the CLT had an impact on Markov, and after he retired from teaching, he started to work on probability theory more seriously. In 1908 he could also prove the CLT under the so called Ljapunov condition with moment methods.

3.3 Ljapunov

Ljapunov was influenced by Chebychev, but he barely worked with moments, he considered Chebyshev's and Makov's work on the CLT to be complicated and he tried to find more general conditions for the CLT.

In 1900, he proved the CLT for the so called "Ljapunov condition". He let x_1,x_2,... be an infinite sequence of independent random variables ("variables independentes"), with $Ex_i =: \alpha_i$, $E(x_i - \alpha_i)^2 =: a_i$ and $E|x_i^3| =: l_i$.
And he also defined $A_n := \frac{\sum_{i=1}^{n} a_i}{n}$ and $L_n^3 := \max_{1 \leq i \leq n} l_i$.
Then he proved that under the condition

$$\frac{L_n^2}{A_n} n^{-\frac{1}{3}} \to 0 \ (n \to \infty)$$

for all $z_1 < z_2$

$$\Big|P(z_1\sqrt{2nA_n} < \sum_{i=1}^{n}(x_i - \alpha_i) < z_2\sqrt{2nA_n}) - \frac{1}{\sqrt{\pi}} \int_{z_1}^{z_2} e^{-z^2} dz\Big| < \Omega_n,$$

where Ω_n is independent of z_1, z_2 and

$$\Omega_n \to 0 \text{ for } n \to \infty.$$

As we can see in his formula, he didn't used the distribution function of the standard normal distribution, just like the other authors he also used the distribution function of the normal distribution with expectation zero and variance half.

In 1901, he could weaken his condition, with

$$\frac{(d_1+d_2+...+d_n)^2}{(a_1+a_2+...+a_n)^{2+\delta}} \to 0$$

where $d_i := E|x_i - \alpha_i|^{2+\delta}$ with an arbitrary small $\delta > 0$.

Ljapunov took the CLT seriously as a distinct mathematical object. His proofs, as we can see in the example of Markov, had an impact on authors in Russia and also in Western Europe.

4 The CLT in the twenties

After the First World War probability theory began to be more important and the CLT became an object of study within mathematics itself.

In the twenties a lot of authors started to work with the CLT, like Richard von Mises, George Pólya, Paul Lévy and Felix Hausdorff. Jarl Waldemar Lindeberg also worked with the CLT, he proved it for the "Lindeberg condition". We also need to talk about Sergei Natanovich Bernshtein, whose "lemma fondamental" was also important in the history of the CLT.

4.1 Von Mises and Pólya

In 1919, von Mises published his article "Fundamental Limit Theorems of Probability Theory", in German "Fundamentalsätze der Wahrscheinlichkeitrechnung", where he formulated and proved his local and integral CLTs, althought his results were obsolete in the one-dimensional case. He also created the term "distribution", which also have German translation, "Verteilung", for a monotonically increasing, right continuous function, which has the limit 0 as x tends to $-\infty$ and 1 as x tends to ∞.

The CLT received its name from an article Pólya wrote in 1920, this article should be recognized as a response to von Mises. The two mathematicans had an exchange of letters, where Pólya critized von Mises's treatment of the CLT, mostly because it was inferior to Ljapunov's and Markov's work.

4.2 Lindeberg

Lindebergs most important result in his mathematical work was his proof of the CLT.

In 1920, he proved the CLT under a very weak condition, he did this without knowing about Ljapunov's works. In this work the discussed random variables, "quantites" X_k, which were mutually independent and had the distribution U_k, with $EX_k = 0$, $VarX_k = EX_k^2 = \sigma_k^2$ and with finite absolute moment of third order. He also presupposed

$$\frac{1}{r_n^3} \sum_{k=1}^{n} \int_{-\infty}^{\infty} |x|^3 dU_k(x) \to 0 \text{ for } (n \to 0),$$

where he defined

$$r_n := \sqrt{\sum_{k=1}^{n} \sigma_k^2}.$$

After certain modifications he could weaken his conditions, so X_k didn't necessary have finite absolut moment of third order, and he published his results in 1922.

In his work he considered $(U_k)_{k\in\{1\dots n\}}$ to be the distribution functions of n mutually independent random variables $(X_k)_{k\in\{1\dots n\}}$ with $EX_k = 0$, $VarX_k = EX_k^2 = \sigma_k^2$, also he presupposed that $\sum_{k=1}^{n} \sigma_k^2 = 1$.
He defined U to be the distribution of the sum of all random variables

$$U(x) := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} U_n(x-t_1-t_2-\dots-t_{n-1})dU_{n-1}(t_{n-1})\dots U_1(t_1),$$

and a function s

$$s(x) = \begin{cases} |x|^3 \text{ if } |x|<1 \\ x^2 \text{ else} \end{cases}.$$

He proved that $\forall\epsilon>0$, even if it is taken arbitrarily small, $\exists\eta>0$ such that

$$\left|U(x) - \int_{-\infty}^{x}\frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}dt\right| < \epsilon$$

if

$$\sum_{k=1}^{n}\int_{-\infty}^{\infty} s(x)dU_k < \eta.$$

Since U is the distribution of the sum of all random variables, it is equal to $P(\sum_{k=1}^{n} U_k < x)$. And with $a_k = EU_k = 0$ and $b_n = \sqrt{\sum_{k=1}^{n}\sigma_k^2} = 1$, We can wrote:

$$U(x) = P(\frac{\sum_{k=1}^{n}(U_k - a_k)}{b_k} < x).$$

We also can see that he used $\frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}$ in the integral, which is the density function of the standard normal distribution.

He used an entirely new method for his arguments.

4.3 Hausdorff

Hausdorff was mainly interested in the integral version of the CLT. He studied Ljapunov's and von Mises's work. He also studied Lindeberg proof of the CLT, he was mostly interested in his method. Later he deduced a theorem, which is a version of the CLT, with the name " Ljapunov's limit theorem", the translation of "Grenzwerthsatz von Liapunoff". For his theorem he presupposed "variables" $X_1,\dots X_2$ and for all j: $EX_j = 0$, $EX_j^2 = a_j^2$ and $E|X_j^3| = c^3$. He considered Φ_n to be the distribution function of $\sum_{k=1}^{n}\frac{X_k}{b_n\sqrt{2}}$, where $b_n^2 = \sum_{j=1}^{n} a_j^2$, and $d_n = (\sum_{j=1}^{n} c_j^3)^{\frac{1}{3}}$, then

$$|\Phi_n - \Phi| \le \mu\Big(\frac{d_n}{b_n}\Big)^{\frac{3}{4}},$$

where μ is a "numerical constant" and $\Phi(x) = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{x} e^{-t^2}dt$.

As we can see, he also used the the distribution function of a normal distribution with variance $\frac{1}{2}$ and expectation 0, $\Phi_{0;\frac{1}{2}}$, instead of the standard normal distribution.

He also noticed a sufficient condition for the convergence of Φ_n to Φ

$$\frac{d_n}{b_n} \to 0 \text{ for } (n\to 0).$$

If we look closer to this condition, we can find out that the "Ljapunov condition" from 1901, with $\delta = 1$ implies it. To prove that, we show that $a_i^2 = EX_i^2 = E(X_i - EX_i)^2$ and $c_i^3 = E|X_i|^3 = E|X_i - EX_i|^3$, since $EX_i = 0$, and then we use the "Ljapunov condition", which says:

$$\frac{(c_1^3+c_2^3+...+c_n^3)^2}{(a_1^2+a_2^2+...+a_n^2)^3} \to 0.$$

With that

$$\Big(\frac{d_n}{b_n}\Big)^{\frac{3}{4}} = \Big(\frac{(c_1^3+c_2^3+...+c_n^3)^{\frac{1}{3}}}{(a_1^2+a_2^2+...+a_n^2)^{\frac{1}{2}}}\Big)^{\frac{3}{4}} = \frac{(c_1^3+c_2^3+...+c_n^3)^{\frac{1}{4}}}{(a_1^2+a_2^2+...+a_n^2)^{\frac{3}{8}}} = \Big(\frac{(c_1^3+c_2^3+...+c_n^3)^2}{(a_1^2+a_2^2+...+a_n^2)^3}\Big)^{\frac{1}{4}} \to 0.$$

4.4 Lévy

In his earlier works Lévy used counterexamples to discuss the CLT.

Lévy created certain probability laws, he called them "laws of type $L_{\alpha,\beta}$. In this laws he worked with constants $c_0 > 0$, c_1 and with characteristic functions in the form $e^{\psi(t)}$, where

$$\psi(t) = -(c_0 + sgn(t)c_1 i)|t|^\alpha$$

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 also showed that $\forall\beta$,$\alpha \neq$ 1,2 exists a probability density function f with a characteristic function ϕ such that

$$\Big(\phi\Big(\frac{t}{n^{\frac{t}{\alpha}}}\Big)\Big)^n \to e^{\psi(t)}.$$

In 1922, he had a version of the CLT, as a special case of his theorem on the convergence to distributions of type $L_{\alpha,\beta}$. In his version of the CLT he considered a sequence of independent random variables (X_k), with distribution functions F_k, which have expectation 0 and variance 1. He also presupposed

$$\forall\epsilon > 0 \exists a > 0 \in N : \int_{|\eta|\leq a} \eta^2 dF_k(\eta) \geq 1 - \epsilon,$$

although it was only important, because it was a condition for "the laws of $L_{\alpha,\beta}$". Then he considered a sequence $(m_k)_{k\in N>0}$ with:

$$\frac{max_{1\leq k\leq n} m_k^2}{\sum_{k=1}^n m_k^2} \to 0 \text{ for } n \to \infty.$$

Then

$$\lim_{n\to\infty} P\Big(\frac{\sum_{k=1}^n m_k X_k}{\sum_{k=1}^n m_k^2} \leq x\Big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$

Unfortunately, Lévy, wasn't lucky with the CLT, since he always had priority conflict with other authors about the publication of similar, sometimes the same, results. He had his first such conflict, about his version above, with Lindeberg.

4.5 Bernshtein

In 1922, Bernshtein published an article, where he used a lemma, the so called "lemme fondamental", with that the CLT could be used for "almost independent random variables" and it also had an influence in the history of the martingale theorems. But Bernshtein didn't give any proof, also the wording of this lemma wasn't entirely clear, so his article didn't had any impact.

The version with the proof was published in 1926.

5 Necessary and sufficient conditions for the CLT

In 1935, both Paul Lévy and Willy Feller proved that there are conditions, which are necessary and also sufficient for the CLT.

5.1 Lévy

As I mentioned before Lévy already worked on the CLT in the twenties.

He tried to get the newest results on the CLT, rather than a well-organized theory, so he often didn't prove the assertions, which he used in his discussion. He also used his newly created analytical tools of concentration and dispersion.

He created "dispersion" to compere the size of a random variable to the overall sum, he also created its inverse and called it "concentration". This two new term were very useful in discussing the convergence of series of random variables.

He defined the concentration $f_X(l)$, which is the maximum probability to an interval length $l > 0$, of a random variable X as

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

And he defined the dispersion $\phi_X(\gamma)$, the minimum interval length to the probability $\gamma \in [0, 1[$, of a random variable X as

$$\phi_X := \inf\{x \in R_0^+ | f_X \geq \gamma\}.$$

In his theorems he let $\gamma \in]0, 1[$ to be an arbitrary, but fixed probability and he considered L_n to be the dispersion of $\sum_{k=1}^n X_k$ assigned to γ. Lévy always assumed $L_n \neq 0$ from a certain number n, although he later showed that this is almost evident.

As I mentioned before he used the dispersion to compere the size of random variables. He called X_k "invidually negligible" in terms of the dispersion of the total sum, if

$$\forall\epsilon\colon\ P\Big(|X_k| > \epsilon L_n\Big) \to 0 \text{ for } (n \to \infty).$$

He also expressed that "all terms are invidually small" if

$$\forall\epsilon > 0\colon\ \lim_{n\to\infty} P\Big(\max_{1\leq n} |X_k| > \epsilon L_n\Big) = 0.$$

He used concentration and dispersion in the case of the CLT too.

First he proved the "classical case" of the CLT, where he considered the (X_k) to be a sequence of identically distributed random variables then

$$P\Big(\frac{\sum k=1^n X_k}{\sqrt{n}} \leq x\Big) \to \Phi(x) \text{ for } n \to \infty,$$

where Φ is the standard normal distribution, if and only if $EX_1^2 = 1$ and $EX_1 = 0$.

In this case, he only had to show that if the distribution of $\frac{\sum k=1^n X_k}{\sqrt{n}}$ tends to 0 then $EX_1^2 < \infty$, because of the properties of the CLT. .

After that, he started to work with the general case, with not identically distributed random variables, where he assumed that the random variables are negligible in terms of the dispersion of the total sum. In this case he proved that the necessary and sufficient conditions are that $\forall \epsilon_1, \epsilon_2 > 0$ and $\forall n \in N\ \exists X(n)$ such that

$$\frac{X(n)}{\sqrt{Var \sum_{k=1}^{n} Y_{nk}}} \leq \epsilon_1, \text{ where } Y_{nk} := \begin{cases} X_k, & \text{if } |X_k| \leq X(n) \\ 0, & \text{else} \end{cases}$$

$$\text{and } \sum_{k=1}^{n} P(|X_k| > X(n)) \leq \epsilon_2$$

5.2 Feller

Feller started to work on probability theory only around 1934. He knew how to deal with characteristic functions and he get some benefit from this knowledge, since he used the characteristic functions as his main tool in his theorem. He also used some auxiliary theorems, which he proved in his article.

His ideas were easy to understand, since he explicitly presented his methods and the characteristic functions were also familiar for his audience.

For the distribution functions V_k of the random variables X_k he also presupposed the negligibility with respect to the respective convolution function W_n. It means that $\exists a_n, b_k\ \forall x \neq 0$:

$$\max_{1 \leq k \leq n} |V_n(a_n x + b_k) - E(x)| \to 0 \text{ for } n \to 0$$

with

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

it could be written also in the form:

$$\forall \epsilon\text{: } \max_{1 \leq k \leq n} P(|X_k - b_k| > \epsilon a_n) \to 0.$$

He found out that for a sequence of distributions V_k, which all have zero median, the sufficient and necessary condition for using the CLT is that

$$\forall \delta > 0\text{: } \lim_{n \to \infty} \frac{1}{p_n^2(\delta)} \sum_{v=1}^{n} \int_{|x| \leq p_n(\delta)} x^2 dV_(x) = \infty,$$

where

$$p_n(\delta) = \min\{r \in R_0^+ | \sum_{v=1}^{n} \int_{|x|>r} dV_(x) \leq \delta\}.$$

In his article Feller also wrote a separate discussion of the Lindeberg condition.

5.3 Results

Their results are similar.

If we want to see both of them results we can consider X_k to be a sequence of independent random variables, with distribution functions V_k, which all have the median 0.

Feller's main result was: $\exists a_n \in R^+$ and $\exists b_k \in R$ such that

$$P\Big(\tfrac{1}{a_n}\textstyle\sum_{k=1}^{n}(X_k - b_k) \le x\Big) \to \Phi(x)$$

$$\text{and } \max_{1\le k\le n} P(|X_k - b_k| > \epsilon a_n) \to 0 \ (\forall \epsilon > 0)$$

as $n \to \infty$ if and only if

$$\forall \delta > 0 \forall \eta > 0 \exists n(\delta,\eta) \forall n \ge n(\delta,\eta): \ \frac{p_n^2(\delta)}{\sum_{k=1}^{n}\int_{|x|\le p_n(\delta)} x^2 dV_k(x)},$$

$$\text{where } p_n(\delta) = \min\{r \in R_0^+ | P(|X_k| > r) \le \delta\}.$$

Lévy in his version considered L_n to be the dispersion of $\sum_{k=1}^{n} X_k$ assigned to an arbitrary, however fixed, probability $\gamma \in]0;1[$. So his theorem looks like: $\exists a_n \in R^+$ and $\exists b_n \in R$ such that

$$P\Big(\tfrac{1}{a_n}\textstyle\sum_{k=1}^{n}(X_k - b_k) \le x\Big) \to \Phi(x)$$

$$\text{and } \max_{1\le k\le n} P(|X_k| > \epsilon L_n) \to 0 \ (\forall \epsilon > 0)$$

as $n \to \infty$ if and only if

$$\forall \delta > 0 \forall \eta > 0 \exists n(\delta,\eta) \forall n \ge n(\delta,\eta) \exists X(n) > 0:$$

$$\frac{X^2(n)}{\sum_{k=1}^{n}\Big(\int_{|x|\le X(n)} x^2 dV_k(x) - (\int_{|x|\le X(n)} x dV_k(x))^2\Big)}$$

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

In their theorems they both had criterion about the negligibility of the random variables, and both criterion implies

$$\max_{1\le k\le n} P(|X_k| > \epsilon a_n) \to 0,$$

which could be proved in Feller's case with the help of the zero median property of all distributions and in Lévy's case with the asymptotically equality of the orders of magnitude of a_n and L_n.

5.4 Priority

As I mentioned before Lévy always had priority conflict in his works about the CLT. In the case of the sufficient and necessary conditions of the CLT, he had such problems again.

Feller's results were given more attention, as Lévy said, the reason for that was that Feller published his work earlier. Later, Le Cam studied the cronology of the publication of their articles and he found out that although Lévy published his work only in December, he made his work, in the form of a " prepint" available for the audience earlier than Feller, what means that he is entitled to the priority. And it was also not certain that Feller's article was published earlier, since we don't know the exact delivery date.

Lévy often wasn't acknowledged for his work. For example, Gnedenko and Kolmogorov didn't mention him in connection with this subject. Even Cramér, who usually had high praise for his work, mentions only Feller in a discussion about necessary and sufficient conditions of the CLT. In the end he realized that there aren't any meaningfully speak about "priority" for two works, which are so different in style and methods.

6 Conclusion

In this study we discussed the history of the central limit theorem.

It started with Laplace, who used it as a tool to solve other mathematical problems. Poisson also had an influence on the history of the CLT, he gave counterexamples to it and also created a new concept, "choses", for random variables. We also talked about Dirichlet's discussion of the linear combination of observational errors and Cauchy's upper bounds for the error of the approximation to the distribution of a linear combination of errors, which both could be considered as a proof of the CLT.

Chebyshev expressed the CLT proper and he used it for "quantites". Markov also gave proofs for the CLT, although his first proof was more likely a corollary of other moment theoretic results. The first one, who considered the CLT as a mathematical problem on its own, was Ljapunov. He also proved it for the so called "Ljapunov condition". After Ljapunov's work, Markov also proved the CLT under the "Ljapunov condition" with moment methods.

After the First World War the CLT became a mathematical problem itself. More author started to work with it. It got its name in 1920, from an article wroted by Pólya. Lindeberg also proved it, for even weaker conditions than Ljapunov, although he didn't knew about his work. Bernshtein's "lemme fondamental" was important too.

The CLT also have sufficient and necessary conditions, Lévy and Feller both found these conditions, in nearly the same time, but with different methods. Feller used more "traditional" methods, while Lévy used his newly invented concentration and dispersion. In the end, they get similar results.

7 References

- Fischer H.: *A History of the Central Limit Theorem*, 2011, Springer