Statistical Science
1986, Vol. 1, No. 1, 78–96

The Central Limit Theorem around 1935

L. Le Cam

Abstract. A long standing problem of probability theory has been to find necessary and sufficient conditions for the approximation of laws of sums of random variables by Gaussian distributions. A chapter in that search was closed by the 1935 work of Feller and Lévy and by a beautiful result of Cramér published in early 1936. We review the respective contributions of Feller and Lévy mentioning as necessary contributions of Laplace, Poisson, Lindeberg, Bernstein, Kolmogorov, and others, with an effort to place them in the context of the authors' times and in a modern content.

Key words: Central Limit Theorem, Gaussian distributions, characteristic functions, martingales.

1. INTRODUCTION

In the beginning there was de Moivre, Laplace, and many Bernoullis, and they begat limit theorems, and the wise men saw that it was good and they called it by the name of Gauss. Then there were new generations and they said that it had experimental vigor but lacked in rigor. Then came Chebyshev, Liapounov, and Markov and they begat a proof and Polyá saw that it was momentous and he said that its name shall be called the Central Limit Theorem.

Then came Lindeberg and he said that it was elementary, for Taylor had expanded that which needed expansion and he said it twice, but Lévy had seen that Fourier transforms are characteristic functions and he said "let them multiply and bring forth limit theorems and stable laws." And it was good, stable, and sufficient, but they asked "Is it necessary"? Lévy answered, "I shall say verily unto you say that it is not necessary, but the time shall come when Gauss will have no parts except that they be in the image of Gauss himself, and then it will be necessary." It was a prophecy, and then Cramér announced that the time had come, and there was much rejoicing and Lévy said that it must be recorded in the bibles and he did record it, and it came to pass that there were many limit theorems and many were central and they overflowed the chronicles and this was the history of the central limit theorem.

This is indeed the story of the central limit theorem, albeit a short one. The name central limit theorem now covers a large variety of different results. Polyá (1920) had bestowed it upon results to the general effect that, under some restrictions, sums of independent random variables, suitably standardized, have cumulative distribution functions close to that given by the famous de Moivre–Laplace formula

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt.$$

Here we shall be concerned mostly with the part of the history of the central limit theorem that took place between 1920 and 1937. The developments of that time involved some of the giants in our field: Bernstein, Lindeberg, Lévy, Feller, and Kolmogorov to mention only a few names.

We have paid special attention to Lévy's contributions for a number of reasons. One of them is that one of Lévy's (1935b) major papers on the subject has not been reproduced in his Collected Works (1976). Lévy complained to the end of his life that he did not receive due credit for that work, all the credit for obtaining necessary conditions for the central limit theorem being claimed by, and usually granted to, Feller.

As we shall see the matter of priorities is considerably more complex. It will be discussed in Section 5.

To describe the situation we shall first have to review various statements of the central limit theorem and classify them according to their formal structure. This is done in Section 2. Then, in Section 3, we give a short review of the contributions of Laplace, Liapounov, Lindeberg, and Bernstein. This brings the story up to the late 1920s or early 30s at which time there were truly major contributions by Kolmogorov and Lévy among others with a final settlement of the problem by Cramér in 1936. They are described in Section 4. Section 6 is about Feller's 1935 paper and Section 7 is about Lévy's paper of the same year.

L. Le Cam is Professor of Statistics and Mathematics at the University of California, Berkeley. His mailing address is Department of Statistics, University of California, Berkeley, CA 94720.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science*. www.jstor.org

The large number of papers published after 1935 makes a short review of the situation rather difficult. For the interested reader, we would recommend the recent books by Araujo and Giné (1980) and Pollard (1984), the papers by Mačys (1968), Zaitsev and Arak (1984), and Dudley and Philipp (1983) as well as the Maurey-Schwartz seminars of the Ecole Polytechnique (1972–1981).

About terminology, the distribution with density $(1/\sqrt{2\pi})e^{-x^2/2}$ was introduced by de Moivre (1738). It and its variations by changes of location and scale are often called "normal." This is an unfortunate appellation as anyone who had to deal with medical problems can testify: the patients ("abnormal") may have "normal" distributions while the "normal controls" have non-normal ones. We have followed Lévy and called the distribution "Gaussian." This is not because Gauss had that much to do with it. He was not involved in the proofs or statements of central limit theorems. He did however publish a paper (Gauss, 1809) that contains a description of Legendre method of least squares. Either that paper was not refereed or the referee did not do a competent job. Gauss' argument is perfectly circular: "Everyone knows that the average of the observations is the best estimate of the expectation. The de Moivre curve is the only one for which that is true for the location parameter. Therefore, the observations must follow that distribution and, therefore, the method of least squares is best." In addition, Gauss uses a version of Bayes' theorem, without giving any credit to Bayes or to Laplace who expounded about it much earlier (Laplace, 1778). The "proof" of Bayes formula by Gauss cannot even be considered adequate by the standards of his time or earlier ones.

Gauss took pride publishing "few things, but polished ones" (*pauca, sed matura*). That cannot be said for the paper in question, even though Gauss complained that it took 3 years to translate it into Latin. It is therefore quite fitting that, in accordance with the Stephen Stigler (1980) law of eponymy, the distribution introduced by de Moivre be called Gaussian.

2. WHAT IS THE CENTRAL LIMIT THEOREM?

The name "central limit theorem" is at present used for a variety of results about the behavior of the distributions of sums of random variables, or random elements that take values in sundry spaces, such as Banach spaces (Araujo and Giné, 1980) or groups (Parthasarathy, 1967). Here we shall concentrate largely on real valued independent random variables and on approximations of the distributions by "normal" ones. The appellation "central" is due to Polyá (1920) who used it because of the central role of the theorem in probability theory, not as the modern French do, because it describes the behavior of the center of the distribution as opposed to its tails.

One of the most commonly used forms of the theorem is as follows. Let $X_1, \ldots, X_n$ be random variables with sum $S = \sum_{j=1}^n X_j$.

THEOREM 1. *Let the variables X_j be independent with expectations zero and variances σ_j^2. Let s be the standard deviation of the sum S and let F be the cumulative distribution of S/s. Let*

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt.$$

Then whenever $\sum_j E\{|X_j/s|^2 I[|X_j| > \varepsilon s]\} < \varepsilon$ one has $\sup_x |F(x) - \Phi(x)| \le 5\varepsilon$.

This statement is close to the statement given by Lindeberg in 1922, except that we have given an explicit bound that is better than that available to Lindeberg. It can be obtained through the Berry-Esseen procedure (see for instance Loéve, 1977, p. 294). For improvements see Zolotarev (1966). Note that it is not truly a "limit theorem: but an "approximation theorem": the distance between the two cumulative distributions is bounded by a function calculable in terms of the individual variables.

In this respect, to describe the contributions of various authors, it is convenient to classify theorems according to their logical form. There are three main categories.

A. Approximation theorems in which a distance between the distribution of a sum and the approximating distribution is bounded by suitable expressions, as in Theorem 1 above, or in the Berry–Esseen theorem (Berry, 1941; Esseen, 1945).

B. Limit theorems for triangular arrays in which one considers a double sequence $\{X_{n,j}: j = 1, 2, \ldots, k_n; n = 1, 2, \ldots\}$ and the sum $S_n = \sum_j X_{n,j}$. One takes the limit of the distributions of the S_n.

C. Normed sums in which one considers a single sequence $\{X_j: j = 1, \ldots, 2\}$ and tries to find constants a_n and c_n such that, if $S_n = \sum_{j=1}^n X_j$, the distribution of $(S_n - c_n)/a_n$ tends to a limit.

As we shall see, both Feller and Lévy, in 1935, used the "normed sums" formulation, thus treating only a rather particular case.

Another classification of theorems can be obtained according to the method of "norming." Theorem 1 uses the so-called "classical norming" by expectations and standard deviations. The reliance on classical norming persisted for a long time. It is used in Kolmogorov's paper of 1933 (Kolmogorov, 1933a) and in Khinchin's booklet (Khinchin, 1933). It is noticeably

absent in Lévy's paper of 1931. The possibility of working without any moments assumptions and using other norming constants is mentioned briefly by Bernstein in 1926 (see Bernstein, 1926, Remarque, p. 23). Lévy (1931) attributes to Khinchin the idea that allows him to dispense with moments.

Theorem 1 uses as distance the vertical distance between cumulative distribution functions. One can use the weaker Lévy distance. Some authors, in particular Khinchin (1938), use distances between densities. We shall not go into the details of all possibilities and stick with the vertical distance between cumulatives.

Theorem 1 should be compared to Theorem 2 below, the "final" version of the central limit theorem as it appears in Lévy's monograph of 1937 (Lévy, 1937a). It is stated there in an intuitive form. It is a wonderful exercise to translate it into an (ε, δ) framework. It can be done. A recent version is that of Zolotarev (1967) for the classical situation, Mačys (1968), for a situation covering the general Gaussian case and Zolotarev (1970) for a more general problem.

THEOREM 2. *In order that a sum $S = \sum_j X_j$ of independent variables have a distribution close to Gaussian it is necessary and sufficient that, after reducing medians to zero, the following conditions be satisfied*:

1. *Each summand that is not negligible compared to the dispersion of the entire sum has a distribution close to Gaussian.*
2. *The maximum of the absolute value of the negligible summands is itself negligible compared to the dispersion of the sum.*

Some of the terms used there need explanation. Note that there is no mention of moments, expectations, or variances. Here the "dispersion" of a sum S can be measured by its interquartile range, for instance. For more general results, using infinitely divisible distributions instead of just Gaussian approximations, one needs to use Paul Lévy's concentration function $C(\tau) = \sup_x P[x \leq S \leq x + \tau]$ or its inverse, called dispersion function, $D(\alpha)$, equal to the infimum of the length of intervals that contain S with probability α or more.

To measure how close a distribution is to another, Lévy uses, for the Gaussian case, the Kolmogorov vertical distance $\rho(F_1, F_2) = \sup_x |F_1(x) - F_2(x)|$ between cumulatives. A term X_j of the sum can be called "negligible" within ε if $\text{Prob}[|X_j| > \varepsilon L] < \varepsilon$ for the interquartile range L of S.

The main difference between Theorem 1 and Theorem 2 is that Theorem 1 just gives a sufficient condition, whereas Theorem 2 asserts that the conditions are necessary and sufficient. There is a not so subtle difference in that, in Theorem 1, $\sup_j \text{Prob}[|X_j| > \varepsilon s]$ must be small, whereas in Theorem 2 absolutely no conditions are imposed, except the independence. This is because Lévy could then use a famous theorem of Cramér (Lévy, 1925):

THEOREM 3. *If the sum $X + Y$ of two independent variables has a Gaussian distribution, then so do X and Y.*

The validity of this result had been conjectured by Lévy in 1928 at the occasion of a hassle with Fréchet (see Fréchet (1928) and Lévy (1929, 1930)). The hassle revolved around two different matters. One was the validity of the theorems stated by Lévy in his 1925 book. Fréchet gave a "counterexample" in the form of a convergent series $\sum_{n=1}^{\infty} \varepsilon_n/n$ where ε_n are identically distributed, say uniformly on $[-1, +1]$. Fréchet attributes the example to Hausdorff. However, Poisson in 1824 had already considered a similar situation with ε_n distributed according to the symmetric exponential density $\frac{1}{2}e^{-|x|}$. Lévy argues in his reply that in the case of a convergent series each term contributes a non-negligible portion of the dispersion of the sum and that under such circumstances one cannot expect the central limit theorem to hold.

The other aspect of Fréchet's criticism is that, in the theory of observational errors, one cannot assume that the different causes of error operate in an additive manner. He proposes a different "law of composition" in which the "total error" is the maximum of the individual independent contributions to it.

It is interesting to note that Fréchet never gave up on these matters. I witnessed exchanges between Fréchet and Lévy in the late forties, with Lévy answering in a gentle but somewhat annoyed tone: "But, Monsieur Fréchet, we have gone over that ground many times since 1928." As we shall see, in the next section, Fréchet was echoing opinions previously expressed by Bertrand (1889), Poincaré (1912), and Borel (1924).

Lévy had conjectured the validity of Theorem 3 quoted above, but he had been unable to prove it. Cramér, in January 1936, obtained a proof of a stronger result:

Let $\varphi(z) = Ee^{zX}$. If φ is an entire nonvanishing function of the complex variable z such that, $\log^+$ denoting the positive part of the logarithm,

$$\limsup_{|z| \to \infty} \frac{1}{|z|^2} \log^+ |\varphi(z)| < \infty,$$

then X has a Gaussian distribution.

To prove the result in that form Cramér used a rather deep theorem of Hadamard. Lévy quickly remarked that, under the conditions of Theorem 3, one can use $|\log|$ instead of $\log^+$ and that a simple theo-

rem on harmonic functions yields the conclusion of the theorem.

The fact that using Cramér's theorem, Lévy could produce necessary and sufficient conditions for Gaussian approximation prodded him to write his famous monograph "Théorie de l'Addition des Variables Aléatoires" (Lévy, 1937a).

3. FROM LAPLACE TO BERNSTEIN

De Moivre, a French mathematician, exiled to England because of religious persecutions, is usually credited with a proof that binomial distributions can be approximated by Gaussian ones. The result, obtained using a formula originally proved by de Moivre but now called Sterling's formula, occurs in his "Doctrine of Chances" of 1733. This is a very special result. Laplace, quoting Lagrange occasionally, wrote many papers where he apparently tried to extend de Moivre's work. Finally, in 1810, he published a paper stating and "proving" a central limit theorem in a very general form. Laplace's proof is certainly valid for a sum of bounded independent random variables whose values are integer multiples of some number ε (lattice variables). The proof uses what is now called the characteristic function, or Fourier transform, Ee^{itX}, for t real. Laplace passes from lattice variables to continuous ones by a swift wave of his hand. That is not too serious, as one can readily check. Laplace also sweeps away the passage to unbounded random variables, saying in effect: "It is easy as I shall show you in the example of the density $\frac{1}{2}e^{-|x|}$." Unfortunately, that sweep is a bit too drastic. Poisson, in 1824, gave counterexamples that include the Cauchy distribution and convergent series of the type used by Fréchet in 1928. In that same paper Poisson extends Laplace's results to sums $\gamma_1 X_1 + \gamma_2 X_2 + \cdots + \gamma_n X_n$ where the X_j are independent, bounded in absolute value by some constant, but not necessarily identically distributed.

The factors γ_j are nonrandom. They are assumed to be bounded away from zero and such that the variances γ_j^2 var X_j remain bounded. Poisson makes further assumptions that are not stated very clearly. Except for this and for the fact that he takes unconstitutional liberties with limits under integral signs, his proof is quite correct. One does get the impression that he understood very well what was going on.

The first rigorous proof is usually credited to Liapounov (1900), some 90 years after Laplace's work. It is interesting to note that Liapounov follows Laplace step by step. He discretizes and uses only variables with a bounded range. However, for these, he gets, after some horrendous calculations, bounds for the distance $\sup_x |F(x) - \Phi(x)|$ that allow him to conclude that the theorem is still true if $EX_j = 0$ and if $\sum_j E|X_j|^3/[\sum_j EX_j^2]^{3/2}$ tends to zero. The result was quickly improved by Markov (1900) and by Liapounov (1901) himself. Markov appears to be the first to try to replace the independence condition on the variables X_j. He gives a theorem for what are now called "Markov chains" (1908). The work of Markov was to be extended to a much wider class of problems by Serge Bernstein. He published a short note in 1922 claiming that the work had been carried out in 1917–1918. A paper with complete proofs appeared in 1926 (Bernstein, 1926).

The proofs used by these authors are of two kinds. Some (e.g., Chebyshev and also Markov) made use of the method of moments. The other proofs rely on the Fourier transforms also called characteristic functions of the type $\varphi(t) = Ee^{itX}$, $i = \sqrt{-1}$, t real. They are definitely in line with Laplace's work of 1810. Bernstein deals with vector valued variables; Laplace had already pointed out the possibility of such extensions and in fact had obtained bivariate Gaussian limits (Laplace, 1810b).

It is true that Laplace's proofs are somewhat incomplete. However, for the cases he considered (identically distributed or with distributions selected from a finite family) there is absolutely no difficulty in making the proof entirely rigorous. The same applies to Poisson's proof of 1824. It is therefore very curious that it took so long before they got rewritten rigorously by Liapounov. There were a few attempts in the meantime, for instance that of Glaisher (1872). However any one of the powerful analysts of the 19th century (e.g., Cauchy (1853), who knew about characteristic functions and stable laws) should have been able to rewrite Laplace's proofs. It is even more remarkable that neither Joseph Bertrand nor Henri Poincaré, considered as the leading French probabilists of their time, could do it. Bertrand and Poincaré wrote treatises on the calculus of probability, a subject neither of the two appeared to know. Except for some faint praise for Gauss' circular argument, Bertrand's book consists mainly of repeated claims that his predecessors made grievous logical mistakes. Poincaré gives a rather thorough discussion of Gauss' argument and of Bertrand's criticism of it. One such criticism was that the density of the observations might not be of the shift kind $f(x - \theta)$ but more general, $f(x, \theta)$. He shows that if the average of the observations is the maximum likelihood estimate, then, in general $f(x, \theta)$ need not be Gaussian but an exponential family. He goes on to criticize the "principle of the arithmetic mean" explaining that outlying observations should be given less weight than the other ones.

Poincaré then says that the best justification for the Gaussian law is that, when one sums many small independent variables, the distribution of the sum is nearly Gaussian. He gives two "proofs," one by the method of moments, another by Laplace transforms.

However these "proofs" contain major gaps. For instance he does not prove and does not even mention that if Laplace transforms converge, so do the distributions. Poincaré does not give many references. The names of Bertrand and Gauss are mentioned without any indication of where the material can be found. In fact, except for some references to his own works, the only reference given by Poincaré is to some algebraic work of Frobenius.

Borel, who succeeded Poincaré as the leading French probabilist, says in the 1924 edition of his book that he

> will not insist on the theoretical discussions surrounding the Laplace–Gauss distribution nor on the mathematical developments to which they have led. Indeed, it does not appear that the results so obtained have an importance in keeping with the analytical efforts they require . . .

Borel goes on to add that "it might be possible to prove certain theorems but they would not be of any interest, since, in practice, one could not verify whether the assumptions are satisfied." Such opinions are expressed again in Borel's 1950 rewrite of his *Eléments*.

Why the French chose to ignore their compatriot Laplace and give first billing to the German mathematician Gauss is hard to explain. Stigler (1980) gives sound reasons for the avoidance of terms such as Laplace's distribution. However this does not explain away total lack of references. Czuber in his excellent book of 1891 is much fairer, except that he seems to think that Laplace treated only the case of symmetrically distributed variables. That is not the case, see for instances page 335 of his *Théorie Analytique des Probabilités* (3rd edition (1820) or Laplace (1810a, p. 322)).

In 1919, when Lévy was asked to give lectures on probability at the Ecole Polytechnique, he relied on Poincaré's book. He was blissfully unaware of the results of the Russian school. What is perhaps more surprising is that he was unaware of the works of Laplace or Cauchy, because, says he (Lévy, 1970), "They are not mentioned in the books by Bertrand, Poincaré, or Borel." He reinvented the technique of characteristic functions, only to be told by Polyá, then in Zurich, that Cauchy had used them before and had given the formulas for the characteristic functions of (symmetric) stable distributions.

With all of this, it was a great surprise when Lindeberg in 1920 and 1922 gave a perfectly elementary proof of the central limit theorem, essentially in the form of our Theorem 1, Section 2 (but not with the explicit bound given here). Lindeberg's proof is very simple. It applies just as easily to Euclidean valued or even Hilbert valued random vectors. Lévy, who reproduces a form of it in his 1925 *Calcul des probabilities*, was to use it very effectively to obtain his central limit theorem for martingales in 1934 (Lévy, 1934b and 1935a). In spite of this, the proof does not appear in standard textbooks (one exception is Thomasian (1969)) and some famous probabilists had difficulties with it. Feller (1971, p. 256 footnote) says:

> Lindeberg's method appeared intricate and was in practice replaced by the method of characteristic functions developed by Lévy. That streamlined modern techniques permit presenting Lindeberg's method in a simple and intuitive manner was shown by H. F. Trotter, *Archiv. der Mathematik*, Vol. 9 (1959) pp. 226–234. Proofs of this section utilize Trotter's idea.

That is surely a big of an exaggeration! Actually Trotter's method differs from Lindeberg mostly by a change in terminology: he uses "convolution operators" instead of "convolution of cumulative distribution functions" as did Lindeberg. The return to "convolution operators," instead of the sums of random variables by Lévy (1931) is unfortunate: It renders difficult the application of the method to the martingales (Lévy, 1934b). In addition, the use of "convolution operators" is not that far from the use of characteristic functions since Fourier transforms are just what is needed to represent the convolution algebra of absolutely continuous measures by an algebra of functions under pointwise multiplication.

Briefly, Lindeberg's method is as follows. Consider a sum $S_n = X_1 + X_2 + \cdots + X_n$ and another sum $T_n = Y_1 + Y_2 + \cdots + Y_n$. Let f be a bounded function. Then

$$Ef(S_n) - Ef(T_n) = \sum_{k=1}^{n} \{Ef(R_k + X_k) - Ef(R_k + Y_k)\},$$

where $R_k = (\sum_{j<k} X_j) + (\sum_{j>k} Y_j)$. If f has two derivatives, the second one satisfying a Lipschitz condition $|f''(u) - f''(v)| \leq A|u - v|$, expand each of $f(R_k + X_k)$ and $f(R_k + Y_k)$ around R_k getting for instance

$$f(R_k + X_k) = f(R_k) + X_k f'(R_k) + \frac{X_k^2}{2} f''(R_k) + \frac{X_k^2}{2}[f''(R_k^*) - f''(R_k)],$$

where R_k^* is in between R_k and $R_k + X_k$. Now if the X_k and Y_k are all independent with $EX_k = EY_k = 0$

and $EX_k^2 = EY_k^2 = \sigma_k^2$, the difference of expectations $Ef(R_k + X_k) - Ef(R_k + Y_k)$ will contain only the "third order" terms

$$E\frac{X_k^2}{2}[f''(R_k^*) - f''(R_k)]$$

and a similar term with X_k replaced by Y_k. This immediately gives a bound

$$|Ef(S_n) - Ef(T_n)| \leq \tfrac{1}{6} A \sum E\{|X_k|^3 + |Y_k|^3\}.$$

Lindeberg takes for the Y_j independent Gaussian variables. Smoothing the indicator function of the interval $(-\infty, x]$ and assuming $\sum \sigma_j^2 = 1$ one gets a bound of the type

$$\sup_x |F(x) - \Phi(x)| \leq C\left\{\sum_k [E(X_k)^3 + |Y_k|^3]\right\}^{1/4}.$$

To get Lindeberg's theorem in its general form, it is sufficient to use a standard truncation argument already used by Liapounov in 1900. Note that Lindeberg's argument can be used for random vectors, interpreting the absolute value sign in $|X_k|^3$ as a norm. In that form, it extends to Hilbert space or any Banach space the topology of which can be obtained from functions whose second derivative satisfies a Lipschitz condition.

It thus seems that after Lindeberg's paper of 1922, or at least after the publication of Lévy's book of 1925, the case could have been considered closed, except perhaps for refinements on the bound given above. It is clear from Lévy's book (1925) that he considered Lindeberg's proof simpler than and superior to his own, using characteristic functions. However he continued using these preferentially because they give easily a number of results on stable laws, until he was stung by criticism from Borel: "The results obtained by this procedure have not been commensurable with the analytic effort they require" (Borel, 1924, p. 125). It is most curious that characteristic functions that have front billing in Lévy's book of 1925 are conspicuously absent in his work of 1930 to 1935, even though to whom Borel's barbs were directed is not entirely clear.

The case was not closed, however, since all the theorems available gave only *sufficient* conditions for approximation by Gaussian distributions. One knew, from the examples given by Poisson in 1824 that the Gaussian approximation did not always hold for sums of arbitrary independent variables. To describe the ascertainment of necessary and sufficient conditions will bring us to another chapter and to a discussion of matters of priority between Feller and Lévy. However, before getting into this it is necessary to report briefly on some developments that took place in France and Russia in the early thirties.

4. MORE LIMIT THEOREMS

In two papers published in 1931 and 1933, Kolmogorov contributed a major result that has not received the attention it deserves. He proves what one would now call an "invariance principle" from the name bestowed by Kac (1949) and Donsker (1951) on functional central limit theorems. Kolmogorov considers independent variables X_j with $EX_j = 0$, $EX_j^2 = \sigma_j^2$ and $E|X_j|^3 \leq \varepsilon\sigma_j^2$. He forms the trajectory $W = \{\sum_{j\leq k} \sigma_j^2, \sum_{j\leq k} X_j;\ k = 1, 2, \dots\}$ as a graph in the plane. Taking two smooth curves $\{a(t);\ t \in [0, T]\}$ and $\{b(t);\ t \in [0, T]\}$, Kolmogorov shows that, for ε small, the probability that W lies between the two curves differs little from a number obtainable from the solution of the heat equation that vanishes on the two curves. He also mentions that one can obtain similar results for variables that are not independent but form Markov chains. Kolmogorov's results are partially reproduced in Khinchin's booklet "Asymptotische Gesetze der Wahrscheinlichkeitsrechnung" of 1933. It is most remarkable that in spite of this they have largely been ignored.

That period also saw the study of convergence of series of independent random variables and that of stochastic processes with independent increments. Some major contributions are those of Khinchin (Khinchin and Kolmogorov, 1925), Kolmogorov (1928 and 1932), and Lévy (1931 and 1934a).

The matter of whether an infinite series $\sum_{n=1}^{\infty} X_n$ of independent variables can be made to converge almost surely by addition of nonrandom terms may seem to be remote from the central limit theorem itself. It was treated by Khinchin and Kolmogorov (1925), by Kolmogorov (1928), and then by Lévy (1931). Lévy had a copy of Kolmogorov's paper. However, he says (Lévy, 1970, p. 87), that he had read only a few pages because he "did not yet know that Kolmogorov would become one of the greatest mathematicians of his generation." In any event Lévy rediscovers by his own methods Kolmogorov's three-series theorem (see for instance Loéve (1977, p. 240), but he goes on studying in detail what happens if the series $\sum_{n=1}^{\infty} X_n$ diverges. This brings him back to the central limit theorem and to two statements. He considers an infinite sequence $\{X_j;\ j = 1, 2, \dots\}$ of independent variables. Another sequence $\{X_j' : j = 1, 2, \dots\}$ is called equivalent to $\{X_j\}$ if $\sum_j \Pr[X_j \neq X_j'] < \infty$. Lévy credits Khinchin with that concept and proceeds to state:

Let $S_n = X_1 + X_2 + \cdots + X_n$. In order that S_n be of the form $A_n + B_n\xi_n$ where A_n and B_n are nonrandom and where ξ_n tends in distribution to the standard

Gaussian distribution $\mathcal{N}(0, 1)$ it is sufficient that there be a sequence equivalent to the original one to which one can apply Lindeberg's theorem.

Lévy then says that the condition used in this theorem cannot be necessary, because some of the X_j could be exactly Gaussian and as large as one pleases, while Lindeberg's condition implies that each X_j must be negligible compared to the dispersion of the entire sum. He states another result as follows:

The above conclusion remains valid if $X_j = a_j + b_j\xi_j + \eta_j + \eta_j'$ where a_j and b_j are nonrandom, ξ_j is $\mathcal{N}(0, 1)$, the $|\eta_j|$ have an upper bound that is infinitely small compared to the dispersion of S_n and $\sum_j \Pr[\eta_j' \neq 0] < \infty$. (For what is meant by "negligible" and "dispersion" see Section 2 above.) Here, the triplets (ξ_j, η_j, η_j') are independent of each other, but the variables constituting a particular triplet need not be independent.

Lévy adds: "It is perhaps not impossible to obtain, in this realm of ideas, a necessary and sufficient condition, but its practical application would be difficult."

This paper of Lévy's is noteworthy for several reasons. Except for Bernstein's (1926, p. 23) footnote, the statements seem to be the first where a central limit theorem is given without any moment conditions whatsoever. The results were obtained by direct methods without any appeal to the characteristic functions he had promoted so vigorously before. They also give the first inkling that he had hopes to obtain necessary and sufficient conditions for the approximation of distributions of sums by Gaussian distributions without imposing any restrictions on how small the terms of the sum might be, while conjecturing that if they are individually small, their maximum must also be small.

A casual reader of the papers of that period might be puzzled by some of the formulations used to say that in $S = X_1 + X_2 + \cdots + X_n$ no single term has much influence on the behavior of the total sum. This may be interpreted in various ways. To avoid complications with centerings and other problems, consider for instance the case where each of the X_j: $j = 1, 2, \ldots, n$ has a symmetric distribution around zero. Let S_k be the sum $\sum_j [X_j: j \neq k]$. Let F and F_k be the respective distributions of S and S_k. The statement might be interpreted as a condition to the effect that the Kolmogorov distance $\sup_k \| F - F_k \|$ is small. This is what Lévy seems to say in his 1929 reply to Fréchet (Lévy, 1929, p. 2). That is also what Feller seems to say (Feller, 1935, p. 523, lines 3–6). Actually that kind of condition is only a necessary and sufficient condition for the approximation of F by the accompanying infinitely divisible law whose characteristic function is $\exp\{\sum_j [\varphi_j - 1]$ with $\varphi_j(t) = Ee^{itX_j}$.

The condition for approximation by a Gaussian distribution is that, if $\sup_k \| F - F_k \|$ is small, then $\sup_j |X_j|$ must also be small compared to the interquartile range of S (or, as shown in Lévy (1937b), small compared to $|S|$).

Another major accomplishment of the early thirties is the characterization of "processes with independent increments." This was started by de Finetti (1929), followed by Kolmogorov (1932), and finished by Lévy (1934a). Here, again, Lévy's paper is noteworthy for its direct methods, avoiding the use of characteristic functions except for a description of the distribution of the process at a given time.

Lévy describes the process as a sum of independent terms. Besides a nonrandom component, the process is a sum of discontinuities that occur at fixed time points, a Gaussian process with continuous trajectories, and a series of terms obtained by selecting time points according to Poisson processes and placing them at these points jumps with independently chosen sizes. From this description Lévy concludes that, when there are no fixed discontinuities, the characteristic function of the process $X(t)$ at time t has the form

$$\log \varphi(s) = ias - \frac{\sigma^2}{2} s^2 + \int [e^{isx} - 1 - isxu(x)]M(dx).$$

Here M is the "Lévy measure" that gives for each set A the expected number $M(A)$ of jumps with sizes in A. The function u can be some continuous function, such that $0 \leq u \leq 1$, equal to one near zero and to zero outside a bounded interval. Lévy observes that the characteristic function φ determines both σ^2 and M and notes the implication that if X and Y are independent, infinitely divisible with a Gaussian sum $X + Y$ then each of X and Y must be Gaussian. He proceeds to state the conjecture that this result remains true without the infinitely divisible restriction on X and Y. He also says that if $X + Y$ is infinitely divisible, then X and Y must be. That was soon shown to be incorrect, but we have not been able to track down the author of the first counter example. (It is trivial to show that if X is a variable taking value 1 with probability $p < \frac{1}{4}$ and zero with probability $1 - p$ one can add a suitable independent Y to make $X + Y$ infinitely divisible.)

In two other papers of 1934, Lévy returns to the central limit theorem, but this time for sums $S_n = X_1 + \cdots + X_n$ where the entries X_j are not independent. They are martingale differences in the sense that given X_j: $j \leq k - 1$, the conditional expectation $E_{k-1}X_k$ is zero. They are assumed to have conditional variances $E_{k-1}X_k^2 = \sigma_k^2$ and are subjected to other conditions bounding their size. This is very similar to the assumptions made by Bernstein (1926, p. 21) in his "fundamental lemma." One essential difference is

that Bernstein assumes that the σ_k^2 are close to nonrandom quantities. Lévy does not make that assumption at all. What he does is to introduce a clock where "time" is measured by the successive sums $\sum_{j\le k} \sigma_j^2$. He calls $S(t)$ the sum $\sum_j [X_j : j \le \nu]$ where ν is the first integer for which $\sum_{j\le\nu} \sigma_j^2 \ge t$ and proceeds to show that, when such integers exist, $S(t)/\sqrt{t}$ has a distribution tending to $\mathcal{N}(0, 1)$ at $t \to \infty$. He also gives indications on what happens if one stops at nonrandom integers n, obtaining "mixed normal" limits.

Lévy's proof is patterned on Lindeberg's, replacing X_ν by a variable αX_ν so that $\sum_{j<\nu} \sigma_j^2 + E_{\nu-1}(\alpha X_\nu)^2$ is exactly t and then proceeding downward from ν to replace the variables X_j by $\sigma_j \xi_j$ with ξ_j independent $\mathcal{N}(0, 1)$ variables. Here again the man who so strongly supported the use of characteristic functions between 1922 and 1930 avoids their use entirely.

At this point the stage was set for a search for conditions that would be not only sufficient but also necessary. Part, but only part, of the answer was given by Feller and by Lévy in papers published in 1935. We shall now take a look at these papers.

5. A MATTER OF PRIORITY

Feller says in his book (1971, Vol. II, p. 256):

> Special cases and variants had been known before, but Lindeberg gave the first general form containing theorem 1. The necessity of Lindeberg's condition with the classical norming was proved by Feller . . .

In his *Souvenirs*, Lévy (1970, p. 108) says:

> However, in the meantime, W. Feller had obtained and published the same result. The independence of our researches cannot be contested and he did not contest it. However, his work having been published before mine, it is to him alone that the merit of having established the theorem that is in a certain sense the final one on the Gaussian distribution is generally attributed. Nonetheless, I am convinced that I found everything without any other useful indications than some sentences of Poincaré. I shall never have had any luck with the Gaussian distribution.

As we shall see both Feller and Lévy treated only a special case in 1934–1935, namely the case of "normed sums" (see Section 2). Both authors assumed that, after norming, the individual summands are "uniformly asymptotically negligible," according to the terminology of Loève (1977) or "infinitesimal" according to that of Gnedenko and Kolmogorov (1954). It is only after Cramér obtained Theorem 3 (Section 2) in January 1936 that the role of the asymptotic negligibility could be fully ascertained. Both Lévy and Feller were aware of the problem arising from the possibility that a sum of a few independent terms could have a near Gaussian distribution without any of the terms being near Gaussian. However, Feller (1935, p. 531) dismisses the problem as "not belonging to the calculus of probability."

As to who did what first, some dates are as follows: Lévy's paper (1935b) was written in October 1934. It was presented at a meeting of the Société Mathématique de France, November 28, 1934. Typesetting took place February 9, 1935. Order to print the issue was given in September 1935 and that issue of the journal was on sale in December 1935. Lévy had also explained the results in a "Notice sur mes travaux scientifiques" distributed privately to colleagues in June 1935. Feller's paper was received by the *Mathematische Zeitschrift* on the 5th of May 1935. That issue of the Zeitschrift was "abgeschlossen" on the 8th of November 1935. As far as could be ascertained the word "abgeschlossen" is used and was used to mean that the issue was complete and ready for the printer. Records on when the issue was actually distributed are missing. However distribution may have been swift. For instance Bernstein's paper of 1926 is included in a volume of *Mathematische Annalen* dated 1927, but the particular "Heft" is dated 1926. It was "abgeschlossen" December 15, 1926 and received at Berkeley January 19, 1927. (Remember, in those day they used boats!).

In his second paper on the central limit theorem, Feller (1937) partly acknowledges Lévy's claim to priority. The German is hard to translate accurately, but here is an approximation:

> I am happy to note, according to a kind communication from Mr. P. Lévy, that his paper, although published later, was submitted and presented to the Société Mathématique de France substantially before mine (October 1934 versus May 1935).

Even though the methods used by Feller and Lévy differ considerably, one could inquire about the possibility of influence through communications between Paris and Stockholm. There was indeed a substantial amount of correspondence between France, Sweden, and Russia on matters of probability. However, Feller had just moved from Kiel to Stockholm. He was new to probability, having previously worked on measure theory, differential geometry, partial differential equations, and some other mathematical subjects. He seems to have known of Lévy's work only through Lévy's 1925 book, ignoring in particular Lévy's work

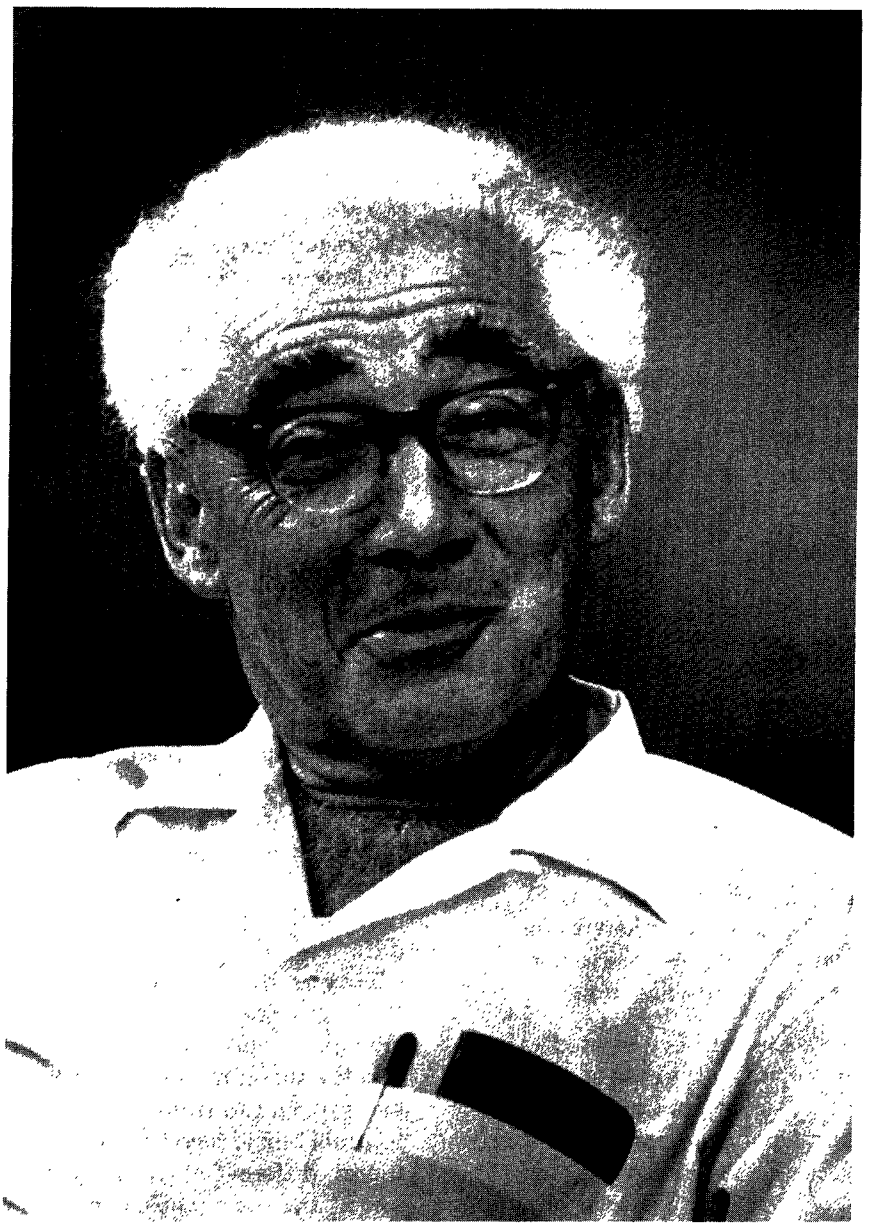
of 1931 and 1934. He also did not appear to know of Bernstein's paper of 1926.

6. FELLER'S PAPER OF 1935

Feller (1935) considers an infinite sequence $\{X_j: j = 1, 2, \dots\}$ of independent random variables and addresses himself to the question:

> Given such a sequence, when do there exist sequences of numbers a_n and c_n such that if
>
> $$S_n = \sum_j [X_j: j \leq n]$$
>
> then
>
> $$\frac{1}{a_n}[S_n - c_n]$$
>
> tends in distribution to $\mathcal{N}(0, 1)$. If so, how can the constants a_n and c_n be computed.

Feller does not answer that question in its general form. He treats only the case where the a_n, if they exist, must be such that for each k the variables X_k/a_n tend in probability to zero. The justification for



W. Feller (picture courtesy of the Department of Statistics, Stanford)

this restriction is given as follows:

> The only case that belongs to the domain of problems on limit theorems is the one in which, as n increases, the influence of the individual components on the distribution of the sum goes to zero, in other words, that the convergence does not arise from the overwhelming influence of individual components that themselves tend to $\mathcal{N}(0, 1)$.
>
> (Feller, 1935, p. 523, lines 3–8.)

Feller refers later (p. 524, lines 2–5) to the problems created by the fact that the a_n might conceivably stay bounded. The problem is mentioned again on page 531, lines 10–20. Feller observes that, in that case, one would obtain a series of independent, perhaps non-Gaussian random variables, say $\{Y_n: n = 1, 2, \dots\}$ with a sum $S = \sum_{n=1} Y_n$ that has a Gaussian distribution. He dismisses the investigation of that possibility as "not belonging to the calculus of probability."

In the above pseudo-quotations, we have reformulated Feller's questions and comments in terms of random variables. Actually Feller does not use that language. He writes about "convolutions of distribution functions." There is some mention of random variables, but only in footnotes, giving the impression that Feller did not think that such concepts belonged in a mathematical framework. This was a common attitude in the mathematical community. It is true that Kolmogorov in 1933 had given to such language a perfectly rigorous mathematical foundation, but it seems to have taken time for this to penetrate. One can even claim that Lévy in his 1925 book (Note, p. 325) had given a mathematical description that is sufficient for the study of sequences of random variables. This, as well as the relevant work of Fréchet (1915) and Daniell (1918), had remained largely unnoticed.

Under his negligibility condition, Feller proceeds to give necessary and sufficient conditions for the existence of the numbers a_n and c_n. The answer is as follows:

THEOREM. *Assume that the X_n have medians equal to zero. For $\delta > 0$, let $p_n(\delta)$ be the smallest number such that*

$$\sum_{j=1}^{n} [\Pr| X_j | > p_n(\delta)] \leq \delta.$$

Then there are numbers a_n and c_n such that the distribution of $(1/a_n)$ $(S_n - c_n)$, $S_n = \sum_{j=1}^{n} X_j$ tends to $\mathcal{N}(0, 1)$ if and only if for each $\delta > 0$ one has

$$\lim \frac{1}{p_n^2(\delta)} \sum_{j=1}^{n} EX_j^2 I[| X_j | < p_n(\delta)] = \infty.$$

Feller proceeds to state that, under this condition, there exist sequences $\{\delta_n\}$, $\delta_n \to 0$ such that if $Y_{n,j} = X_j I[| X_j | < p_n(\delta_n)]$ then a_n^2 can be taken equal to the sum $a_n^2 = \sum_{j=1}^{n} \operatorname{var} Y_{n,j}$ of the variances of the truncated variables $Y_{n,j}$. He also gives further possible choices, all depending on the preliminary choice of certain sequences of truncation constants.

Under the negligibility assumption, this does answer the question he posed, at least formally. However, it does not provide a very usable recipe for selecting the constants a_n for any given value of n. A possible recipe could be as follows: Let $\{X_j': j = 1, 2, \dots\}$ be an independent copy of the sequence $\{X_j: j = 1, 2, \dots\}$. Let $Z_j = X_j - X_j'$. For $a \in (0, \infty)$, let $s_n^2(a) = \sum_{j=1}^{n} E \min[1, Z_j^2/a^2]$. Take a_n such that $s_n^2(a_n) = 2$. Once the a_n have been chosen it is a simple matter to recenter S_n. One takes for c_n the sum $\sum_{j=1}^{n} EX_j I[| X_j | \leq a_n]$.

Now let us consider what could be called "new" in Feller's, paper if one does not take into account Lévy's paper (1935b). The sufficiency of the conditions can be readily derived from Lindeberg's work (1922) if one takes into account Lévy's statement of 1931 quoted in Section 4. However, Feller's result is stated in a more analytic and precise language that that of Lévy. The main claim made by Feller in 1971 is that he proved the *necessity* of the conditions. He did, for the case he considered, and also provided information on the selection of the constants a_n and c_n.

Feller's proof is oppressively analytical. His apparent refusal to use the language of random variables and expectations makes the formulas and derivations awkward and heavy. However the principle of the operation is rather simple. He uses characteristic functions and Lévy's convergence theorem for them. We have already noted that his references to the literature are meager. He does mention Lévy's book of 1925 several times. However, even there, the references are not always correct. For instance he claims that Lévy proved a "special case (of the convergence theorem for characteristic functions) under the assumption of existence of second moments." Such moments are not even mentioned in Lévy's (1925, p. 197) statement. Feller later apologized for the error in 1937. As I said before, Feller was new to the calculus of probability. This was his first paper on the subject. He does not seem to have had time to survey the previous literature.

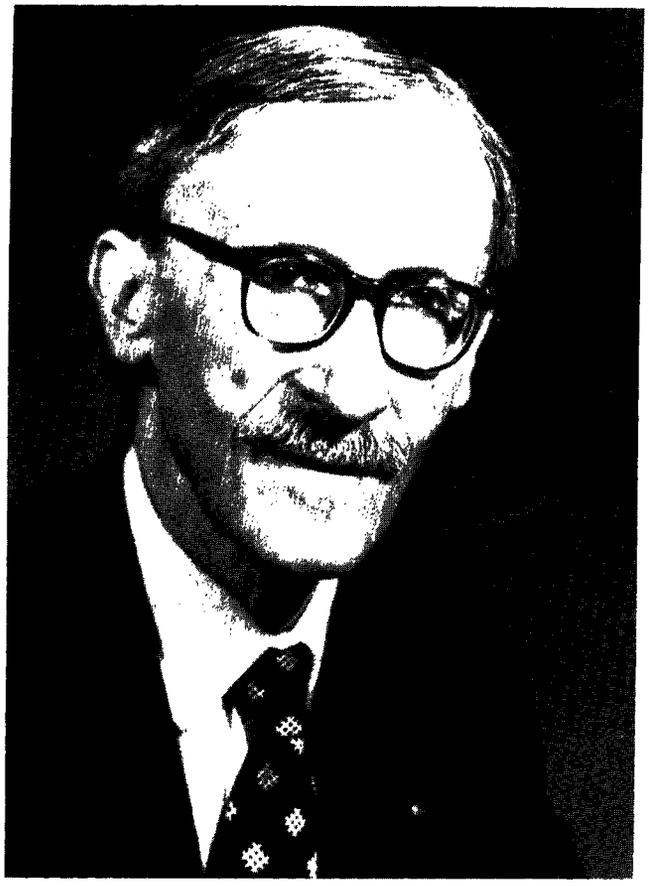
Finally, let us note that Feller treats only the case called "normed sums" (see Section 2). It is true that his analytic method can be readily expanded to cover the case of "triangular arrays." However Feller's paper is replete with statements about the possible behavior of the coefficients a_n and other matters that would not have any validity (or any relevance) in the case of

"triangular arrays." Thus, in this respect at least, one cannot say that he gave the "final" solution to the central limit problem. As we shall now see, neither did Lévy in his paper (1935b).

7. LÉVY'S PAPER OF 1935

This is a sizeable paper. It is much harder to read than Feller's paper, even for a French native whose German is most deficient. Feller's paper is replete with formulas and precise, heavy analytical derivations. Lévy's contains few formulas. The author just discourses along, scattering a wealth of ideas on his way. However, in many respects, Lévy's paper is more interesting than Feller's.

After a brief introduction, Lévy proceeds to obtain stable distributions by an "elementary" procedure. "Elementary" means that he does not use characteristic functions at all. In fact the only characteristic functions that are mentioned are on page 357 for stable laws and on page 380 as an aside after a statement to the effect that if the limiting distributions of normed sums of identically distributed variables are not stable, they must still be infinitely divisible, with characteristic functions given by the formula Lévy (1934a) had previously derived in 1934.



P. Lévy

After this aside on stable laws, Lévy proceeds to give necessary and sufficient conditions for the convergence to a Gaussian distribution of normed sums of independent and *identically distributed* variables. Then he tackles the general case of independent summands. However, not satisfied with that, he gives a detailed study of convergence to the Gaussian law for what we now call "martingales." (The name "martingale" was introduced in probability theory by Ville (1939) who took it from the jargon used in Monaco to describe gambling systems. It was adopted by J. L. Doob (1953) who used it for sequences of random variables whose expectation given the past is the previous variable, as should occur for the fortune of a gambler in a fair game.)

As we shall see, Lévy's statements and proofs are correct for independent summands. His direct martingale central limit theorem is also quite correct. However, his converse theorem, for the martingale case, suffers from some difficulties. Here one should keep in mind Lévy's own statement in his *Souvenirs*, page 107: " . . . being in too much of a hurry, I did not wait for my ideas to reach full maturity and my paper was badly written." From the story told in the *Souvenirs*, one gathers that Lévy dealt with the independent identically distributed case in September 1934, but wrote about the general case, including martingales, in early October 1934. He had obtained and published earlier (Lévy, 1935a) central limit theorems for martingales. The sufficiency proof is reproduced (Lévy, 1935b), but the necessity statements are added. They are not entirely correct.

One of the techniques introduced by Lévy in his study of sums of independent variables is a splitting of the summands according to the size of their absolute values. One could summarize the result as follows. Let $X_{n,j}$ have median zero. Let $X'_{n,j} = X_j I[|X_{n,j}| \leq \tau_n]$ *and let* $X''_{n,j} = X_{n,j} I[|X_{n,j}| > \tau_n]$. If $\sup_{j \leq n} \Pr[|X_{n,j}| > \tau_n]$ tends to zero then the two sums $\sum_j X'_{n,j}$ and $\sum_j X''_{n,j}$ behave asymptotically as if they were independent.

In all the arguments, Lévy assumes that Lindeberg's theorem is known. He concentrates on the role of "large values" of the X_j and shows that $\sum_{j \leq n} X_j$ can be normalized to have distribution close to $\mathcal{N}(0, 1)$ only if $\max_{j \leq n} |X_j|$ becomes small compared to the dispersion of the sum $S_n = \sum_{j \leq n} X_j$. This is done through a computation on the tail probabilities of S_n: If S_n is approximately $\mathcal{N}(0, \sigma_n^2)$, then the tail probabilities $P[|S_n| > x]$ are approximately

$$\frac{2\sigma_n}{x} \exp\left\{-\frac{1}{2}\frac{x^2}{\sigma_n^2}\right\}.$$

Now take $\tau_n = \xi\sigma_n$ in the definition of the variables $X''_{n,j}$ described above. Choose ξ so that $\sum_{j \leq n} P[|X_j| > \tau_n]$ is, say, approximately $1/10$. Lévy argues that the

dispersion of $\sum_{j\le n} X''_{n,j}$, and therefore that of S_n is too large to be compatible with the Gaussian exponential bounds. He uses that argument for the case of identically distributed variables and repeats it with a modification for the general case of independent summands. However, in the latter case, he needs to apply it to sums of type $\sum_j \{X_j: m < j \le n\} = S_n - S_m$ where both S_m and S_n are approximately Gaussian. This does not allow him to deal with what he calls "intermittent convergence," that is convergence for selected subsequences $\{n_\nu: \nu = 1, 2 \dots\}$. Also, in the general case, Lévy deals only with "individually negligible terms," thus using the same restriction as Feller.

It is most peculiar that Lévy would resort to a delicate (but "elementary"!) evaluation of the tail probabilities for the sums S_n. Lévy must have known that sums of individually negligible terms have distributions tending to infinitely divisible limits. The result is an easy consequence of his splitting of X_j as $X'_{n,j} + X''_{n,j}$. That is an argument he had used forcefully in his 1934 study of processes with independent increments (Lévy, 1934a). In that same paper, he had given the general formula for the characteristic function of infinitely divisible distributions. He had proved the appropriate uniqueness theorem for the entries in that formula. For the measure M, now called the Lévy measure of the process with independent increments, Lévy had given the interpretation recalled in Section 4. The same kind of interpretation is readily available for sums of independent variables. Lévy's discussion of the effects of large values indicates that he probably knew of such an interpretation. He certainly could have established it without any trouble. From all this information, the necessity for the Gaussian limits of the negligibility of $\sup_{j\le n} |X_j|$ compared to the dispersion of S_n is an utter triviality, if the summands are already assumed to be individually negligible.

The documents available to us are insufficient to establish why Lévy did not proceed along the lines we just indicated. However they suggest possibilities. Besides being stung by Borel's criticism, Lévy was fascinated by the general case, without any negligibility restrictions. This is suggested by his paper of 1931 and by his "hypothetical Lemma III" (Lévy, 1935b, p. 381) that says that if X and Y are independent and $X + Y$ Gaussian then so are X and Y (Cramér's Theorem, see Section 2). Lévy elaborated at length on this point, explaining that if the hypothetical Lemma III is correct, the nonindividually negligible terms must be approximately Gaussian. Removing them, the rest must obey the "law of large numbers." This last appellation occurs throughout the paper. It has nothing to do with what one usually calls laws of large numbers, but is used to mean that $\sup_{j\le n} |X_j|$ is negligible compared to the dispersion of S_n. Lévy (1935b, p. 388, footnote) also says, "extension to the case of intermittent convergence ... is immediate if Lemma III is true, but seems rather difficult to establish without using that Lemma." As already mentioned "intermittent" means convergence along subsequences $\{n_\nu: \nu = 1, 2, \dots\}$ and Lévy was perhaps thinking of the irrelevant difficulty created by the fact that a sum $\sum_j \{X_j: n_{\nu-1} < j \le n_\nu\}$ need not be negligible compared to the dispersion of S_{n_ν}.

Another explanation for Lévy's refusal to use characteristic functions is that he was thinking of a proof applicable to the martingale case and that he considered that "characteristic functions are mostly useful for sums of independent variables" (see Lévy, 1970, p. 76). This may be so, because the remainder of Lévy's paper deals with dependent variables $\{X_j: j = 1, 2, \dots, n\}$ where X_n satisfies either a boundedness restriction and the condition that $E[X_n | X_1, \dots, X_{n-1}] = 0$, or a symmetry condition, given the past variables $X_1, X_2, \dots, X_{n-1}$.

In previous papers, Lévy (1935a) had given the martingale limit theorems described in Section 4. The theorems are repeated in the 1935 paper, using a truncation procedure to eliminate the boundedness restriction. This gives a "martingale central limit theorem" that has become a prototype for most others. Note that the martingale case had already been treated by Bernstein in 1926. However Bernstein assumed that the conditional variances σ_j^2 are very close to nonrandom quantities. Lévy makes no such assumption.

There is no difficulty with the direct part of the martingale central limit theorem. However Lévy goes further and tries to extend to that case the proof of necessity he had obtained for the independent case. In brief, he assumes that, given the values of the $X_1, \dots, X_{j-1}$, the next variable X_j has a symmetric distribution around zero. He considers constants $a(t)$ and lets $X'_{t,j} = X_j I[|X_j| \le a(t)]$. Define $\sigma^2_{t,j} = E\{[X'_{t,j}]^2 | X_1, \dots, X_{j-1}\}$ and let $k(t)$ be the first integer n such that $\sum_j [\sigma^2_{t,j}: j \le n] \ge t$. Let $\mathscr{S}(t) = \sum_j \{X_j: j \le k(t)\}$.

Lévy asserts that if the $k(t)$ exist and if $\mathscr{S}(t)/\sqrt{t}$ tend to $\mathscr{N}(0, 1)$ as $t \to \infty$, then $1/\sqrt{t} \sup_j \{|X_j|: j \le k(t)\}$ must tend to zero in probability. His proof of this assertion is very mysterious. He asserts, without any explanations, that under the circumstances,

$$\mathscr{S}\left(\frac{i+1}{m}\, t\right) - \mathscr{S}\left(\frac{it}{m}\right)$$

must also be approximately Gaussian, for any integer m and any $i = 0, 1, 2, \dots, m - 1$. He also uses a symmetry argument, as if, given the past, $\mathscr{S}(\tau, t) = \sum_j \{X_j: \tau < j \le k(t)\}$ were symmetrically distributed around zero for all stopping times $\tau < k(t)$. This is not necessarily true as shown by examples given by Lévy himself in the next page of his paper (p. 397). It is however possible to carry out a proof if one adds to Lévy's condition the assumption that the sums

$(1/\sqrt{t})\ \mathcal{L}(\tau, t)$ have uniformly bounded conditional medians.

Next, Lévy elaborates on a number of examples where the sums under consideration are *not* taken "at sections with constant t." He shows that they need not be normally distributed in the limit. For instance they may be "mixed normal," that is, conditionally Gaussian given the value of a random variance term. He gives a very interesting example where the X_j have symmetric distributions given the past, but where he drives the distributions toward Gaussian ones and then makes them drift away, repeating such cycles infinitely often. This gives a counterexample to the theorem he had just stated, but he does not notice it.

Lévy also states other conjectures to the effect that, under the conditional symmetry assumption or similar ones, the distribution of the sums can tend to Gaussian only if two particular conditions are either simultaneously fulfilled or simultaneously violated. The conditions are: 1) that the sums be taken "at sections with constant t" and 2) that $\sup_j\{|X_j|: j \leq k_n(t)\}$ be negligible compared to the dispersion of $\mathcal{L}(t)$. As already mentioned, one needs supplementary conditions to insure the validity of Lévy's converse theorem. The range of validity of Lévy's other conjectures does not seem to have been studied.

Almost as soon as the papers by Feller and Lévy appeared, Cramér (1936) proved the validity of Lévy's hypothetical Lemma III. This prompted Lévy to write his 1937 monograph. There, sums of independent random variables are treated in detail and so are stochastic processes with independent increments. A martingale central limit theorem is proved, but the converse is not mentioned. In the meantime Feller also returned to the evaluation of his norming constants (Feller, 1937). Cramér published a *Cambridge Tract* that is small, but packed with information. Much earlier in 1924, Borel had started editing a monumental *Traité de Calcul des Probabilités*, with many contributing authors. The publication continued, but that treatise had essentially no influence on the development of the field, while Cramér's booklet and Lévy's monograph were bibles to generations of probabilists.

ACKNOWLEDGMENTS

This reseach was partially supported by National Science Foundation Grants MCS80-2698 and DMS 8403239. For help with the subtleties of the German language, I am most indebted to Dr. Imke Janssen and to Professor Erich Lehmann. Details concerning the publication of Lévy's 1935 paper were communicated to me by Henri Villat through the help of Jacques Neveu. Checks on the publication of the *Mathematische Zeitschrift* were carried out by Walter Kaufmann-Buhler of Springer-Verlag. I am also indebted to Professor Harald Cramér for his recollections of the 1935 period and mention of the fact that Feller and Lévy met at the Geneva Conference of 1937. Although the two great men discussed matters of common interest, each remained convinced that he had priority over the other. Finally I am indebted to Jacques Neveu and Bernard Bru who pointed out the work of Poisson and transmitted a copy of it to me.

REFERENCES

ARAUJO, A. and GINÉ, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables.* Wiley, New York.

BERNSTEIN, S. (1926). Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Math. Ann.* **97** 1–59.

BERRY, A. C. (1941). The accuracy of Gaussian approximation to the sum of independent variates. *Trans. Am. Math. Soc.* **49** 122–136.

BERTRAND, J. (1889). *Calcul des probabilités.* Paris, Gauthier-Villars.

BOREL, E. (1924). *Eléments de la Théorie des Probabilités*, 3rd ed. Hermann, Paris.

CAUCHY, A. (1853). Sur les résultats moyens d'observations de même nature et sur les résultats les plus probables. *C. R. Acad. Sci. Paris* **37** 198–206 (see also *Collected Works*).

CRAMÉR, H. (1936). Sur une propriété de la loi de Gauss. *C. R. Acad. Sci. Paris* **202** 615–616.

CRAMÉR, H. (1937). *Random variables and probability distributions.* Cambridge Tracts No. 36, Cambridge University Press.

CZUBER, E. (1891). *Theorie der Beobachtungsfehler.* Teubner, Leipzig.

DANIELL, P. J. (1918). A general form of integral. *Ann. Math.* **19** 279–294 (See also Integrals in an infinite number of dimensions. *Ann. Math.* **20** 281–288).

DE FINETTI, B. (1929). Sulla funzione a incremento aleatorio. *Atti. Acad. Naz. Lincei* **6** 163–168, 325–329, 548–553.

DE MOIVRE, A. (1738). *The Doctrine of Chances*, 2nd ed. The 3rd ed. (1756) has been reprinted by Chelsea, New York (1967).

DONSKER, M. (1951). An invariance principle for certain probability limit theorems. *Mem. Am. Math. Soc.* **6** 1–12.

DOOB, J. L. (1953). *Stochastic Processes.* Wiley, New York.

DUDLEY, R. M. and PHILIPP, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **62** 509–552.

ESSEEN, C. G. (1945). Fourier analysis of distributions functions—A mathematical study of the Laplace-Gaussian law. *Acta Math.* **77** 1–125.

FELLER, W. (1935). Über den Zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung. *Math. Z.* **40** 512–559.

FELLER, W. (1937). Über den Zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung. II. *Math. Z.* **42** 301–312.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd ed. Wiley, New York.

FRÉCHET, M. (1915). Sur l'intégrale d'une fonctionelle étendue à un ensemble abstrait. *Bull. Soc. Math. France* **43** 248–265.

FRÉCHET, M. (1928). Sur l'hypothèse de l'additivité des erreurs partielles. *Bull. Sci. Math. Ser. 2* **52** 203–216.

GAUSS, C. F. (1809). *The Heavenly Bodies Moving about the Sun in Conic Sections.* Reprint, Dover Pub., New York, 1963.

GLAISHER, J. W. L. (1872). On the law of facility of errors of observations, and on the method of least squares. *Mem. R. Astronomical Soc.* **39** 75–124.

GNEDENKO, B. V. and KOLMOGOROV, A. N. (1954). *Limit Distri-*

butions for Sums of Independent Random Variables. Addison-Wesley, Reading, MA.

KAC, M. (1949). On deviations between theoretical and empirical distribution functions. *Proc. Natl. Acad. Sci. U.S.A.* **35** 252–257.

KHINCHIN, A. Y. (1933). Asymptotische Gesetze der Wahrscheinlichkeitsrechnung. *Ergebn. Math.* **2** 1–77.

KHINCHIN, A. Y. (1938). *Mathematical Foundations of Statistical Mechanics.* GTTI, Moscow-Leningrad. (English translation by G. Gamow, Dover, New York, 1949).

KHINCHIN, A. Y. and KOLMOGOROV, A. N. (1925). Über Konvergenz von Reihen deren Glieder durch den Zufall bestimmt werden. *Mat. Sbornik* **32** 668–677.

KOLMOGOROV, A. N. (1928). Über die Summen durch den Zufall bestimmter unabhangiger Grossen. *Math. Ann.* **99** 309–319.

KOLMOGOROV, A. N. (1931). Eine Verallgemeinerung des Laplace-Liapounoffschen Satzes. *Izv. Akad. Nauk SSSR, Ser. Mat.* 959, 962.

KOLMOGOROV, A. N. (1932). Sulla forma generale di un processo stocastico omogeneo. *Atti. Acad. Naz. Lincei Cl. Sci. Fis. Mat. Nat.* **6** 868–869.

KOLMOGOROV, A. N. (1933a). Über die Grenzwertsätze der Wahrscheinlichkeitsrechnung. *Izv. Akad. Nauk SSSR, Ser, Fiz. Mat.* 363–372.

KOLMOGOROV, A. N. (1933b). Grundbegriffe der Wahrscheinlichkeitsrechnung. *Ergeb. Math.* **2**

LAPLACE, P. S. (1778). Mémoire sur les probabilités. *Mém. Acad. R. Sci. Paris.* reproduced in *Oeuvres de Laplace* **9** 383–485.

LAPLACE, P. S. (1810a). Mémoire sur les formules qui sont fonctions de très grands nombres et sur leur application aux probabilités. *Mém. Acad. Sci. Paris* **10** reproduced in *Oeuvres de Laplace* **12** 301–345.

LAPLACE, P. S. (1810b). Mémoire sur les intégrales définies et leur application aux Probabilités. *Mém. Acad. Sci. Paris* reproduced in *Oeuvres de Laplace* **12** 357–412.

LÉVY, P. (1925). *Calcul des probabilités.* Gauthier-Villars, Paris.

LÉVY, P. (1929). Sur quelques travaux relatifs à la théorie des erreurs. *Bull. Sci. Math.* **53** 1–21.

LÉVY, P. (1930). La théorème fondamental de la théorie des erreurs. *Ann. Inst. H. Poincaré* **1** 163–175.

LÉVY, P. (1931). Sur les séries dont les termes sont des variables éventuelles indépendantes. *Studia Math.* **3** 119–155.

LÉVY, P. (1934a). Sur les intégrales dont les éléments sont des variables aléatoires indépendantes. *Ann. Scuola Norm. Pisa* 337–366.

LÉVY, P. (1934b). L'addition de variables aléatoires enchainées et la loi de Gauss. *Bull. Soc. Math. France* **62** 42–43.

LÉVY, P. (1935a). Propriétés asymptotiques des sommes de variables aléatoires enchainées. *C. R. Acad. Sci. Paris* **199** 627–629.

LÉVY, P. (1935b). Propriétés asymptotiques des sommes de variables indépendantes on enchainées. *J. Math. Pures Appl.* 347–402.

LÉVY, P. (1937a). *Théorie de l'Addition des Variables Aléatoires.* Gauthier-Villars, Paris.

LÉVY, P. (1937b). Complément à un théorème sur la loi de Gauss. *Bull. Sci. Math.* **61** 115–128.

LÉVY, P. (1970). *Quelques Aspects de la Pensée d'un Mathematicien. Souvenirs Mathématiques. Considérations Philosophiques.* Blanchard, Paris.

LÉVY, P. (1976). *Oeuvres.* D. DUGUÉ (ed). Gauthier-Villars, Paris.

LIAPOUNOV, A. M. (1900). Sur une proposition de la théorie des probabilités. *Bull. Acad. Sci. St. Petersburg* **5** 359–386.

LIAPOUNOV, A. M. (1901). Nouvelle forme du théorème sur la limite des probabilités. *Mem. Acad. Sci. St. Petersburg* **8** 1–24.

LINDEBERG, J. W. (1920). Über das Exponentialgesetz in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae* **16** 1–23.

LINDEBERG, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Math. Z.* **15** 211–225.

LOÉVE. M. (1977). *Probability Theory,* 4th ed. Springer-Verlag, New York.

MAČYS, Y. Y. (1968). Sur la convergence des répartitions de sommes de variables aléatoires indépendantes vers les lois de la classe I_0 de Linnik. *C. R. Acad. Sci. Paris* **267** 316–317.

MARKOV, A. (1900). *The Calculus of Probability.* Akad. Nauk, Petrograd.

MARKOV, A. (1908). Extension des théorèmes limites du calcul des probabilités aux sommes des quantités liées en chaine. *Mem. Acad. Sci. St. Petersburg* **8** 365–397.

MAUREY, B. and SCHWARTZ, L. (1972–1976). *Séminaire Maurey-Schwartz Espaces L^p, Applications Radonifiantes et Géométrie des Espaces de Banach.* Ecole Polytechnique-Centre de Mathématiques, Palaiseau, France.

MAUREY, B. and SCHWARTZ, L. (1977–1978). *Séminaire sur la Géometrie des Espaces de Banach.* Ecole Polytechnique-Centre de Mathématiques, Palaiseau, France.

MAUREY, B. and SCHWARTZ, L. (1978–1981). *Séminaire d'Analyse Fonctionnelle.* Ecole Polytechnique-Centre de Mathématiques, Palaiseau, France.

PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces.* Academic Press, New York.

POINCARÉ, H. (1912). *Calcul des probabilités,* 2nd ed. Gauthier-Villars, Paris.

POISSON, S. D. (1824). Sur la probabilité des résultats moyens des observations. Additions à la *Connaissance des Temps pour l'Année.* 1827. Bachelier, Paris, 1824.

POLLARD, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

POLYÁ, G. (1920). Über den Zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentproblem. *Math. Z.* **8** 171–180.

STIGLER, S. (1980). Stigler's law of eponymy. *Trans. N. Y. Acad. Sci. Ser. II* **39** 147–158.

THOMASIAN, A. (1969). *The Structure of Probability Theory with Applications.* McGraw-Hill, New York.

VILLE, J. (1939). *Etude Critique de la Notion de Collectif.* Gauthier-Villars, Paris.

ZAITSEV, A. Y. and ARAK, T. V. (1984). On the rate of convergence in Kolmogorov second uniform limit theorem. *Theory Prob. Appl.* **28** 351–374.

ZOLOTAREV, V. M. (1966). An absolute estimate of the remainder in the central limit theorem. *Teor. Verojatn. Primen* **11** 108–119.

ZOLOTAREV, V. M. (1967). A generalization of the Lindeberg-Feller theorem. *Theory Prob. Appl.* **12** 608–618.

ZOLOTAREV, V. M. (1970). Théorèmes limites généraux pour les sommes de variables indépendantes. *C. R. Acad. Sci. Paris* **270** 899–902.