Commentary on Fasz. 64

S. D. Chatterji

The main body of the comments is divided into 7 sections followed by two appendices which complement the discussion contained in section 3 (as regards the strong law of large numbers) and that in section 4 (as regards LINDEBERG's central limit theorem).

§ 1 Introduction

During his period of full professorship in Bonn (1921–1935) HAUSDORFF had given semester long courses on probability theory twice, once in 1923 and later in 1931 (both over the summer semesters). However, during the early part of his career in Leipzig (i. e. between 1895 – 1910) he had lectured on probability theory and related topics (like mathematical insurance theory, mathematical statistics, political arithmetic) several times (at least 8 semestrial lectures). Besides, he had published two long research articles on probability theory in 1897 and 1901 ([H 1897a], [H 1901a], this volume, pp. 443–590). However, his thinking on probability theory in terms of measure and integration seems to have begun in earnest while he was writing his magnum opus *Grundzüge der Mengenlehre* ([H 1914a], vol. II of this edition) in Greifswald during 1913–1914. In the 10th chapter of the *Grundzüge* („Inhalte von Punktmengen") HAUSDORFF states his point of view on this explicitly:

> ... daß manche Theoreme über das Maß von Punktmengen vielleicht ein vertrauteres Gesicht zeigen, wenn man sie in der Sprache der Wahrscheinlichkeitsrechnung ausdrückt. (loc. cit. p. 416)

In the same chapter, HAUSDORFF gives the first correct measure-theoretical proof of BOREL's famous law of normal numbers (loc. cit. pp. 419–422) which is now (as in these Lecture Notes) rightly considered to be a special form of the strong law of large numbers of probability theory. From a study of HAUSDORFF's Nachlass (some of which is reproduced in this volume elsewhere) and these Lecture Notes, it would seem that HAUSDORFF's ideas and techniques in probability theory never progressed much beyond what we can glimpse here. We shall discuss in this essay the several novelties in HAUSDORFF's presentation in these Lecture Notes as compared to others of the pre-1920 era; we shall also indicate the severe theoretical limitations from the modern point of view of the developments given here.

HAUSDORFF's Lecture Notes are subdivided into nine sections followed by a set of 11 Exercises presented as worked out examples. The section headings can be approximately translated as follows:

1. Elementary probabilities.

2. Moments of elementary distributions.

3. Games of chance. Insurance calculations.
4. General probabilities.
5. General distributions and their moments. Additive set functions and Stieltjes integrals.
6. Distribution of a pair of variables.
7. The exponential law. Lyapounov's limit theorem.
8. The moment problem. The second limit law.
9. Comparison between theory and experience. Method of least squares.

We shall now give critical summaries of the nine sections evaluating them from a modern perspective but at the same time comparing their contents with what was the contemporary 1920's presentation of the topics involved. In our exposition, we have generally followed HAUSDORFF's notations with occasional exceptions which have been indicated.

§ 2 Finitely additive theory

The first three sections of the Notes concern elementary probability theory where only the probabilities of events depending on finitely many (say n) other events are involved; eventually (and interestingly) statements are sought when $n \to \infty$. Mathematically, this theory needs only the finite additivity of the underlying probability measure and the chief difficulties are inherently combinatorial. The related literature is rich and vast, dating back to PASCAL and FERMAT (17th century) continuing through JACOB BERNOULLI (his famous book *Ars Conjectandi* published posthumously in 1713) and LAPLACE (whose main treatise dates from 1812); it has continued to flourish unabated all through the 19th and the 20th centuries down to our own times. Most introductory courses in probability theory give a glimpse of this rich heritage and HAUSDORFF's is no exception. Here no sophisticated mathematical constructions are needed to start the theory and rigour and clarity can be achieved with a minimum of technical preliminaries. An abstract modern presentation may be based on a finitely additive probability measure defined on a suitable Boolean algebra which, without loss of generality, can be taken to be formed from certain subsets (and even of all subsets) of a fixed set; the elements A of the Boolean algebra $\mathcal{A}$ would then be interpreted as the "events" (taken as an undefined term) and $w(A)$ the associated probability. HAUSDORFF's presentation can certainly be interpreted this way and it is a point of view which would have been easy for him to accept; however, HAUSDORFF does not make any explicit statement about the axioms to be fullfilled by his events ("Ereignisse") and he lays out his axioms of probability in two simple assumptions which can be interpreted either abstractly (as above) or more concretely (as done in the examples presented) as a finitely additive probability measure defined somehow on certain

subsets of a given set. The more modern probability texts (e. g. the well known book of FELLER [F 1968]) present even this elementary theory more formally as a countably additive probability measure defined on all subsets of a finite or a denumerable set, a concept that needs hardly any elaborate development. All the older books, without exception, leave matters undefined almost as in these Lecture Notes of HAUSDORFF, passing on to interesting and illustrative examples as soon as possible. As in many other texts, HAUSDORFF develops the theory of so-called elementary distributions ("elementare Verteilungen") and the moments of the associated random variables, although the latter term is neither defined nor introduced. The standard examples of binomial, multinomial and hypergeometric distributions are introduced, the notions of independence and conditional probabilities are discussed and BAYES' theorem is presented. All of this is illustrated by the usual examples of coin-tossing, dice-throwing, Card games, withdrawl of balls from urns, roulette, lotto and the like. BAYES' theorem is accompanied by a presentation of what is often known as LAPLACE's law of succession (cf. HAUSDORFF's text, pp. 604–608). Along with moments, the formalism of cumulants (called "die logarithmischen Momente") is also given. The cumulants were much used by probabilists and statisticians of the period around 1900 and HAUSDORFF had included some theory of these in his article [H 1901a] under the name "kanonische Parameter"; the cumulants (also called semi-invariants) were first introduced by THIELE in 1889; see HALD [Ha 1998] pp. 344–349 for their origin and a brief report on [H 1901a], this volume pp. 577–578.

In section 2, various elementary forms of the weak law of large numbers (theorems I, II, III, III*) are proved as well as theorem IV which is equivalent to a strong law; indeed, HAUSDORFF forewarns the reader that this last theorem is of a nature different from the others and that it will be taken up formally later in section 4. The 4th moment calculation leading to equation (16) of section 2 of the Notes, was one of HAUSDORFF's main original steps in his proof of BOREL's law of normal numbers as given in his *Grundzüge*; it is taken up in a slightly more elaborate form later in section 4.

§ 3 Countably additive theory

Sections 4, 5, 6 of the Notes develop the mathematical machinery needed in order to dicuss general probability distributions. Here, the axiom of countable additivity (i. e. σ-additivity) is explicitly formulated ("Axiom (γ)" in section 4), the notions of $A^\infty = \limsup A_n,\ A_\infty = \liminf A_n$, for a sequence of events $\{A_n\}$ are defined and a clear formulation of the so-called Borel-Cantelli lemma is proved. Evidently, the modern appelation "Borel-Cantelli" is not used by HAUSDORFF; he simply says that if $\sum_n w(A_n) < \infty$ the $w(A^\infty) = 0$, ($w(A)$ being the probability of the event A) and he gives the correct one line proof using countable subadditivity (cf. (8) of section 4). He uses this next to prove a form of the strong law of large numbers (signalled before in section 2). He adds further that if the events A_n are independent then $\sum_n w(A_n) = \infty$ implies

that $w(A^\infty) = 1$. He ends this discussion with a reference to BOREL's famous 1909 paper which he qualifies as "prinzipiell ganz unklar". We shall explain some of the lacunae in BOREL's discussion in appendix A where we present HAUSDORFF's elegant and concise proof of the strong law in the context of modern work.

Sections 5 and 6 are devoted essentially to the theory of measure and integration in $\mathbb{R}$ and $\mathbb{R}^2$. Section 5 begins with the statement that the previous considerations (specially those of section 4 concerning σ-additivity) remain somewhat uncertain ("schweben insofern noch in der Luft") in so far as the realizability of the σ-additive axiom remains to be established. Let us recall that already in the *Grundzüge*, HAUSDORFF had shown that the Lebesgue measure (with translation invariance) cannot be defined for all subsets of $\mathbb{R}^n$ if one were to demand σ-additivity; this, of course, goes back to VITALI as well in 1905. Indeed, HAUSDORFF had proved further that even finitely additive extensions of the Lebesgue measure to all subsets of $\mathbb{R}^n$ were impossible if $n \geq 3$ and if one insisted on rotation invariance. Hence the importance of establishing the existence of suitable σ-additive probability measures defined on appropriate σ-algebra (HAUSDORFF calls them "abgeschlossenes Mengensystem" or "Borelsches System"), a fact clearly underlined by HAUSDORFF. All through the development of the foundations of probability theory in the 20th century this point has rightly attracted the close attention of many mathematicians starting with WIENER who established the so-called Wiener measure an $C([0,1])$ in 1920–23; this was preceded by DANIELL (1918–1919) and more explicitly followed by KOLMOGOROV in 1933, to name just a few of the most important. HAUSDORFF never seemed to have noted these developments, neither in these Notes nor elsewhere.

What HAUSDORFF establishes in section 4 is that given a monotone non-decreasing left-continuous function $\varphi : \mathbb{R} \to [0, \infty[$ with $\varphi(-\infty) = 0,\ \varphi(\infty) = \mu < \infty$, there exists a σ-additive non-negative measure Φ defined on a σ-algebra containing all intervals such that for $-\infty < \alpha < \beta < \infty$

$$\Phi([\alpha, \beta[) = \varphi(\beta) - \varphi(\alpha) \tag{3.1}$$

A study of HAUSDORFF's proof shows that he actually proves the following more general theorem: let $\mathcal{I}$ be a semi-ring of subsets of any set M i. e. $\mathcal{I}$ is stable under finite intersections and if A, B are in $\mathcal{I}$ then $A \setminus B$ is a finite disjoint union of sets from $\mathcal{I}$; suppose also that M is the denumerable union of sets $A_n \in \mathcal{I}$, $n = 1, 2, \ldots$; if $\Phi : \mathcal{I} \to [0, \infty[$ is a σ-additive set function such that $\sum_n \Phi(A_n) < \infty$, then Φ can be extended to a σ-additive (finite) measure defined on a σ-algebra $\mathcal{M}$ of subsets of M with $\mathcal{M}$ containing $\mathcal{I}$. The proof also gives (although this is not explicitly mentioned) that the extension is unique if we restrict ourselves to the smallest σ-algebra $\mathcal{B}$ containing $\mathcal{I}$ ($\mathcal{I} \subset \mathcal{B} \subset \mathcal{M}$). It remains now to apply this result to the case of $\mathcal{I}$ formed of all Intervalls I of the form $[\alpha, \beta[$ $(-\infty < \alpha < \beta < \infty)$; taking $\mathcal{I}$ to be sets of the form $[\alpha, \beta[\times [\alpha', \beta'[$ we get the measures treated by HAUSDORFF in $\mathbb{R}^2$ in section 5. HAUSDORFF clearly sees the analogy between the work in section 4 and that in section 5 and he no doubt must have seen the obvious generalization to $\mathbb{R}^n$;

however, he does not note the complete generality of his method of extension. If he had, he could have stated what has later been called the Hahn-Kolmogorov extension theorem, a special form of which can be given as follows: any σ-additive probability measure defined on an algebra of subsets can be uniquely extended (preserving σ-additivity) to the generated σ-algebra.

It is important to realize that before the application of the general theorem mentioned above, it must be established that (3.1) actually does define a σ-additive set function on the semi-ring $\mathcal{I}$ of all intervals of the form $[\alpha, \beta[$. HAUSDORFF does this in the now standard fashion by using the so-called Heine-Borel covering lemma. Once we have a σ-additive set function Φ on the semi-ring $\mathcal{I}$, HAUSDORFF's proof consists of the following steps: for any set $X \subset M$, define

$$\overline{\Phi}(X) = \inf \left\{ \sum_{n=1}^{\infty} \Phi(I_n) \ : \ X \subset \bigcup_n I_n \right\};$$

then define

$$\underline{\Phi}(X) = \mu - \Phi(M \setminus X)$$

where $\mu = \overline{\Phi}(M)$; by the hypothesis made on Φ, μ turns out to be finite. Call a set X *measurable* if

$$\overline{\Phi}(X) = \underline{\Phi}(X).$$

It is then shown that the family of measurable sets is a σ-algebra $\mathcal{M}$, $\mathcal{M}$ contains $\mathcal{I}$ and if $A \in \mathcal{I}$ then

$$\Phi(A) = \overline{\Phi}(A).$$

The proof given is obviously an adaptation of LEBESGUE's original proof for the existence of his measure and can be easily adapted to situations where the underlying Φ is only σ-finite.

All this is, of course, standard material given in numerous modern books; in the 1920's such general discussions were seldom given in probability texts although thanks to CARATHÉODORY's well-known book (*Vorlesungen über reelle Funktionen* 1917) and other monographs (e. g. those of LEBESGUE, DE LA VALLÉE POUSSIN and others) these techniques of measure theory were becoming wide-spread in the 1920's.

In the rest of section 5, HAUSDORFF rapidly develops the integration theory of measurable real functions f of one real variable with respect to a finite positive measure φ in $\mathbb{R}$. This is done as follows: after establishing the stability of such measurable functions under standard algebraic operations and the formation of sup, inf, limit of sequences of them, it is easily shown that any such measurable function f is the uniform limit of a sequence $\{f_n\}$ of measurable functions each taking only at most denumerably many distinct values (a function of the latter type is called a "Skalenfunktion"); finally, one defines $\int f \, d\varphi$ as the limit of $\int f_n \, d\varphi$. The usual properties of the integral are worked out with special mention of the monotone convergence theorem; all this is done very efficiently and rigorously; the relation between this integral and that of the Riemann-Lebesgue-Stieltjes theory is clearly spelled out. In section 6, the work

is generalized to $\mathbb{R}^2$ and a version of FUBINI's theorem is established (without mentioning FUBINI); this is then used to introduce the convolution integral (in modern notation $f * g$: (20) section 5) and the moments etc. for one or more random variables and their sums. However, the term random variable is never introduced.

We know from NL HAUSDORFF : Kapsel 51 : Fasz. 1129 (a manuscript captioned "Maß- und Integrationstheorie" of 203 sheets composed around the latter half of the 1920's) that HAUSDORFF was planning a considerable extension of the 10th chapter of his *Grundzüge* devoted to measure and integration theory. It would thus appear that the material in sections 4, 5, 6 was destined to be used in a more elaborate work. HAUSDORFF's sustained interest in the subject is testified by the large manuscripts which he prepared for his various lectures on themes entitled *Modern Integration Theory* or *Real Functions and Measure Theory* (NL HAUSDORFF : Kapsel 13 : Fasz 43: "Der moderne Integralbegriff", Bonn, WS 1922/23, WS 1927/28, 210 sheets; Kapsel 17 : Fasz. 53: "Reelle Funktionen und Maßtheorie", Bonn, WS 1932/33, 295 sheets). The last contains much material on PERRON's theory of integration and derivation theory; however, there is no treatment of integration theory in abstract sets (à la FRÉCHET - DANIELL - KOLMOGOROV etc.) which could further probability theory.

§ 4 The central limit theorem

As understood today, "a" central limit theorem states that the probability distribution of a sum $x_1 + \cdots + x_n$ of n random variables $x_1, \ldots, x_n$ "converges" as $n \to \infty$ to a certain distinguished probability distribution *provided* the sums are suitably normalized and the variables satisfy appropriate conditions. "The" central limit theorem for a sequence of *real-valued independent* random variables $x_1, x_2, \ldots$ with 0 mean and finite variances is understood to be a statement of the following type ($\mathbb{E}$ denoting mathematical expectation): let

$$a_j^2 = \mathbb{E}\, x_j^2, \quad j \geq 1; \quad b_n^2 = a_1^2 + \cdots + a_n^2, \quad 0 < b_n < \infty, \quad n \geq 1;$$

then

$$\lim_{n\to\infty} \mathbb{P}\mathrm{rob}\left\{\frac{x_1 + \cdots + x_n}{b_n} < z\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du \tag{4.1}$$

provided that the sequence $\{x_n\}$ satisfies certain further conditions. A high-water mark in the history of this theorem is provided by the theorem proved by LINDEBERG in 1922; he proved that (4.1) is valid if

$$\forall\, \varepsilon > 0, \quad \lim_{n\to\infty} \frac{1}{b_n^2} \sum_{j=1}^{n} \mathbb{E}\,(x_j^2;\ |x_j| > \varepsilon b_n) = 0 \tag{4.2}$$

The condition (4.2) is now-a-days called the *Lindeberg condition* although LINDEBERG himself did not write it in this form; in Appendix B we have given the

various equivalent forms of (4.2) given by LINDEBERG in his 1922 article (cited precisely by HAUSDORFF) and we have explained there the (standard modern) notation which we have used in (4.2). We note in passing that HAUSDORFF himself (contrary to LINDEBERG) does not use the standard normal distribution (mean 0, variance 1, as in (4.1)) but rather the normal distribution of variance $\frac{1}{2}$ (density $e^{-u^2}/\sqrt{\pi}$, $u \in \mathbb{R}$) with norming constants $\sqrt{2}\, b_n$ rather then b_n as seems to have been customary in much of the writing before 1920; this is obviously of no theoretical importance. Although HAUSDORFF cites LINDEBERG, he does not prove the central limit theorem (4.1) under the general Lindeberg condition (4.2) but rather under LYAPOUNOV's (dating from 1900; cf. reference in HAUSDORFF's Notes) which can be given as follows:

$$\lim_{n\to\infty} \frac{1}{b_n^3} \sum_{j=1}^{n} \mathbb{E}\,|x_j|^3 = 0 \tag{4.3}$$

supposing, of course, that $\mathbb{E}\,|x_j|^3 < \infty,\ j \geq 1$. Under the last hypothesis, it is easy to show that (4.3) implies (4.2). Indeed, LINDEBERG's first theorem in his 1922 paper is essentially the proof of the central limit theorem (4.1) under the Lyapounov condition (4.3). LYAPOUNOV also gave a more general condition which can be stated as follows: for some $\beta > 2$,

$$\lim_{n\to\infty} \frac{1}{b_n^\beta} \sum_{j=1}^{n} \mathbb{E}\,|x_j|^\beta = 0 \tag{4.4}$$

It is not too difficult to show (a proof is given in HAUSDORFF's Notes) that (4.3) implies (4.4) for $2 < \beta < 3$. HAUSDORFF also points out that if (4.4) holds for some $\beta > 3$ then (4.3) will hold also so that the central limit theorem proved under (4.3) will have been proved under (4.4) as well (with $\beta > 3$). HAUSDORFF concludes by remarking that in the special case of uniformly bounded random variables the central limit theorem (4.1) is valid if only $b_n \to \infty$; this remark is then used to obtain the normal approximation to binomial like probabilities illustrated by some numerical examples of the latter.

One interesting point in HAUSDORFF's presentation is that he gives the central limit theorem in the following simple finitary form: for any fixed $n = 1, 2, \ldots,\quad x \in \mathbb{R}$,

$$\left| \mathbb{P}\mathrm{rob}\left\{ \frac{x_1 + \cdots + x_n}{b_n} < x \right\} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du \right| \leq \mu\, L^{1/4}(n) \tag{4.5}$$

where

$$L(n) = \frac{1}{b_n^3} \sum_{j=1}^{n} \mathbb{E}\,|x_j|^3$$

and μ is some fixed numerical constant independent of the random variables x_n. Recall that in HAUSDORFF's work,

$$\mathbb{E}\, x_j = 0,\ j \geq 1, \quad b_n^2 = \mathbb{E}\,(x_1 + \cdots + x_n)^2, \quad d_n^3 = \sum_{j=1}^{n} \mathbb{E}\,|x_j|^3,$$

so that

$$L(n) = \left(\frac{d_n}{b_n}\right)^3;$$

further, we have used in (4.5) the norming b_n instead of $\sqrt{2}\,b_n$ and the standard normal density instead of HAUSDORFF's normal density of variance $\frac{1}{2}$ which, as we have already pointed out, does not affect (4.5) in any way. The result contained in (4.5) (implicit in LINDEBERG's paper; cf. Appendix B) although far from being the optimal result in this direction, was striking for the 1920's, not only for its remarkable elegance but also for its rigorous but simple derivation. The statement and proof given by HAUSDORFF is essentially LINDEBERG's, presented in a way that a numerical estimate for μ can be easily deduced. In Appendix B, we shall see that the best result in (4.5) replaces $L^{1/4}(n)$ by $L(n)$; this is generally referred to as the Berry-Esseen theorem; further comments on this will be found in our Appendix B.

Section 7 contains a full discussion of the basic analytic properties of the normal distribution ("das Gausssche Exponentialgesetz") including the calculation of its moments, its moment generating function, its associated orthogonal polynomials (Hermite polynomials) and the corresponding series development as well as the basic formulae concerning the gamma and the beta functions. All this was, of course, standard material for the 1920's and is dealt with efficiently. GAUSS' well-known "maximum-likelihood" derivation of the normal distribution is briefly but clearly described; HAUSDORFF, however, seems to be more inclined to consider the central limit theorem itself as a more significant justification for the use of the normal distribution. Some of the analytical material concerning the latter is repeated to some extend in section 8 for its appropriate use there.

§ 5 The "second" limit theorem

Section 8 is entirely devoted to this theme and the related moment problem. For historical reasons given hereafter, the older mathematical literature (up until 1931 at least) often used the appellation "second limit theorem" for the following statement (or any of its equivalent versions): let $F_1, F_2, \ldots$ be a sequence of monotone non-decreasing functions such that $F_n(-\infty) = 0$; $F_n(\infty) = 1$, $n \geq 1$ and such that all the moments of positive integral orders $k \geq 1$ exist for each F_n; let

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du, \quad x \in \mathbb{R};$$

if

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} x^k\, dF_n(x) = \int_{-\infty}^{\infty} x^k\, dF(x), \quad k = 1, 2, \ldots \tag{5.1}$$

then

$$\lim_{n\to\infty} F_n(x) = F(x) \tag{5.2}$$

for all $x \in \mathbb{R}$.

As mentioned before, in HAUSDORFF's Notes as well as in other older texts $F(x)$ is defined as

$$F(x) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{x} e^{-u^2}\, du;$$

obviously, this slightly different standardization of the limiting Gaussian or normal distribution function is of no theoretical significance.

The first proof that (5.1) $\Rightarrow$ (5.2) (for F standard normal) was given by CHEBYSHEV (around 1887) whose proof was completed by MARKOV (around 1898). CHEBYSHEV's proof (in the French version) appeared in his paper *Sur deux théorèmes relatifs aux probabilités* (cf. his *Oeuvres*, vol. II, pp. 480–491, Chelsea); in this paper, the first theorem concerns the weak law of large numbers and the second corresponds to the implication (5.1) $\Rightarrow$ (5.2). This seems to be the origin of the terminology "second limit theorem" used at times by MARKOV and others. Actually, CHEBYSHEV also formulated the implication (5.1) $\Rightarrow$ (5.2) in more finitary terms: supposing that in (5.1) only the moments of the order $k = 1, 2, \ldots, 2N$ were used, he gave an estimation for $|F_n(x) - F(x)|$ which then led to the result (5.1) $\Rightarrow$ (5.2) by letting $N \to \infty$. As can be guessed from these remarks, CHEBYSHEV (as well as MARKOV) based their proofs on their detailed studies of the moment problem which we do not discuss here; a clear exposition of this "method of moments", along with some historical references, can be seen in USPENSKY's book [U 1937], Appendix II, pp. 356–395. A different, direct proof was offered by PÓLYA in his paper *Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem*, Math. Zeitschrift 8 (1920), 171–181. It seems that this is the beginning of the appellation "Central limit theorem" for theorems asserting convergence to a normal distribution; a careful summary of its early history is given in HALD [Ha 1998] pp. 335–337, where further historical references can be found. HALD not only cites CHEBYSHEV, MARKOV, LYAPOUNOV, PÓLYA and later papers and books of historical relevance, he also gives a clear analysis of the contributions of the earlier papers of BESSEL, POISSON, CAUCHY and LAPLACE towards a "general" proof of the so-called central limit theorem. We cannot, however, go into this detailed analysis of the history of the central limit theorem here.

A definitive generalization of the so-called "second limit theorem" was given by FRÉCHET and SHOHAT in 1931 in their paper *A proof of the generalized second limit theorem in the theory of probability*, Trans. Amer. Math. Soc. 33 (1931), 533–543. As a conclusion to their main theorem, FRÉCHET and SHOHAT can conclude that if (5.1) holds for F, any monotone non-decreasing function with $F(-\infty) = 0,\ F(\infty) = 1$ all of whose positive integral moments are finite, then (5.2) holds whenever x is a point of continuity of F, *provided* that F is determined by its moments. This last condition, well-studied by many since CHEBYSHEV, means that if

$$\int_{-\infty}^{\infty} x^k\, dF(x) = \int_{-\infty}^{\infty} x^k\, dG(x), \quad k = 0, 1, 2, \ldots$$

(where G is any monotone non-decreasing function with $G(-\infty) = 0,\ G(\infty) = 1$, having all positive integral moments) then $F(x) = G(x)$ for all points of continuity x of G (or of F). The modern treatment of these questions is via the positive Radon measures in $\mathbb{R}$ induced by bounded monotone non-decreasing functions; although this latter point of view is more natural to us, we shall refrain from translating matters into that language here.

It is useful to recall briefly the main theorem of FRÉCHET and SHOHAT which yield the above generalized second limit theorem since this will show clearly the novelty of HAUSDORFF's approach as presented in these Notes. A minor specialization of the main theorem of FRÉCHET–SHOHAT states that if

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} x^k \, dF_n(x) = m_k, \quad k = 1, 2, \cdots$$

holds (where F_n's are as in the previous discussion and the m_k's are all finite) then there exists a subsequence $n_1 < n_2 < \cdots$ and a monotone non-decreasing function F with $F(-\infty) = 0,\ F(\infty) = 1$ such that

$$m_k = \int_{-\infty}^{\infty} x^k \, dF(x), \quad k = 1, 2, \ldots$$

and

$$\lim_{i\to\infty} F_{n_i}(x) = F(x)$$

for all points of continuity x of F.

Once the main theorem of FRÉCHET–SHOHAT is estabished, their generalized second limit theorem becomes an easy consequence. Indeed, any subsequence of $\{F_n\}$ will then have a further subsequence converging to some monotone non-decreasing function say G (with $G(-\infty) = 0\ G(\infty) = 1$) having the same moments as F so that $F(x) = G(x)$ for all points of continuity x, if F is determined by its moments; it is then evident that the sequence F_n itself must converge to F.

The above line of proof is the one generally given in modern text-books (cf. [CT 1988] pp. 264–265 for a complete presentation). This involves a preliminary discussion of the so-called HELLY–BRAY theorems which have to do with the weak convergence of positive measures in $\mathbb{R}$; these were known around 1920 but perhaps not so widely; FRÉCHET and SHOHAT refer to MONTEL (1907) and HELLY (1912) for their "selection theorems". The fact that the standard normal distribution is determined by its moments was well known since CHEBYSHEV but its proof is not trivial; HAUSDORFF's Notes offer a complete proof; other standard proofs can be read in many text-books (cf. [CT 1988] p. 285).

HAUSDORFF's Notes avoid all the preminilary work on the convergence of Stieltjes integrals and nevertheless manages to prove the complete form of the CHEBYSHEV–MARKOV version of the second limit theorem (theorem III of section 8); the theorems I and II of section 8 give a slightly more general version in so far as the limiting distribution function F (HAUSDORFF's φ) need not be

standard normal. The statements about the uniform convergence of $F_n(x)$ to $F(x)$ (if F is continuous) has been well-known at least since PÓLYA's paper of 1920 and is an easy analytical exercise.

Since HAUSDORFF's method of proof for the CHEBYSHEV–MARKOV theorem is novel and somewhat unorthodox, we indicate briefly its main line of attack, following essentially HAUSDORFF's notations. Given a monotone non-decreasing function φ, with $\varphi(-\infty) = 0,\ \varphi(\infty) = \mu_0$, possessing all moments

$$\mu_k = \int_{-\infty}^{\infty} x^k \, d\varphi(x), \quad k = 0,\, 1,\, 2, \ldots$$

(note that $\mu_0 = \varphi(\infty) = 1$ in probabilistic contexts; HAUSDORFF proceeds a little more generally), HAUSDORFF calls a point $\xi \in \mathbb{R}$ a *point of determinacy* ("Bestimmtheitsstelle") for the moment sequence $\{\mu_k\}$ if for any monotone non-decreasing function ψ with $\psi(-\infty) = 0,\ \psi(\infty) = \mu_0$, possessing the same moments μ_k i. e.

$$\mu_k = \int_{-\infty}^{\infty} x^k \, d\psi(x), \quad k = 0,\, 1,\, 2, \cdots$$

we have $\varphi(\xi) = \psi(\xi)$. As HAUSDORFF immediately points out, at a point of determinacy ξ, φ must evidently be continuous; also, if every point $\xi \in \mathbb{R}$ is a point of determinacy for $\{\mu_k\}$ then the corresponding moment problem is determinate (but not conversely, for the obvious reason that determinate moment problems corresponding to discontinuous φ are excluded). HAUSDORFF now shows that ξ is a point of determinacy for the moment sequence $\{\mu_k\}$ if and only if $\delta(\xi) = 0$ where

$$\delta(\xi) = \inf \left\{ \int_{-\infty}^{\infty} P(x) \, d\varphi(x) \ :\ P \text{ polynomial, } P \geq 0, \quad P(\xi) \geq 1 \right\}$$

It is further shown that

$$\delta(\xi) = \lim_{n \to \infty} \delta_n(\xi)$$

where $\delta_n(\xi)$ is defined like $\delta(\xi)$ but restricting the polynomials P there to be of degree $\leq 2n$. Next, an explicit formula for $\delta_n(\xi)$ is obtained:

$$\frac{1}{\delta_n(\xi)} = \sum_{\nu=0}^{n} b_\nu \, f_\nu^2(\xi)$$

where $\{f_n\}$ is the sequence of orthogonal polynomials associated with φ, rendered unique by demanding that $f_0 \equiv 1$, degree of $f_n = n$ with coefficient of x^n in f_n being 1; further

$$\frac{1}{b_n} = \int_{-\infty}^{\infty} f_n^2(x) \, d\varphi(x), \quad n \geq 0.$$

Thus, ξ is a point of determinacy for the moment problem $\{\mu_k\}$ if and only if

$$\sum_{\nu=0}^{\infty} b_\nu \, f_\nu^2(\xi) = \infty.$$

Since the f_n's and the b_n's can be explicitly calculated if φ is the normal distribution function (these calculations involving the Hermite polynomials are given in detail) it can be established that the moment problem for the normal distribution function is determinate. HAUSDORFF can now easily deduce his theorem I (section 8) to the effect that if the moments of any monotone non-decreasing sequence $\{\varphi_n\}$, $\varphi_n(-\infty) = 0$, converge to the corresponding moments $\{\mu_k\}$ of φ and if ξ is a point of determinacy (in the sense explained above) then $\varphi_n(\xi) \to \varphi(\xi)$ as $n \to \infty$. Theorem II (section 8) is now an immediate corollary in that here φ is supposed to be such that all points $\xi \in \mathbb{R}$ are points of determinacy.

The reasoning used by HAUSDORFF to achieve what is sketched above is elementary but not very short (about 7 printed pages of careful work). Some of it is standard material in the theory of orthogonal polynomials as related to the moment problem; the definition of $\delta(\xi)$, $\delta_n(\xi)$ (and of other related quantities which we have omitted) is directly related to M. RIESZ's approach to the moment problem. RIESZ's papers concerning this appeared first during 1921–1923 (cf [R 1988] pp. 216–311, "Sur le problème des moments" (3 papers)) and formed an important enrichment of the theory of moments. We know from HAUSDORFF's Nachlass that he had learned and mastered M. RIESZ's techniques around September 1920 and we find literally dozens of manuscripts on the general theme of the moment problem (dated between 1917–1924 and later); the manuscripts indicate HAUSDORFF's thorough knowledge of the theory including the classical contributions of CHEBYSHEV and STIELTJES as well as the later work of HAMBURGER. Let us recall HAUSDORFF's own important papers in this area published in 1921–1923 ([H 1921], [H 1923b]) reprinted and commented in [H 2001].

§ 6 Method of Least squares

The ninth and last section of HAUSDORFF's Notes is entirely devoted to least squares estimation of unknown parameters. This is now considered to be a topic in statistical theory; however, many of the standard probability texts of the early 20th century (for example, those of POINCARÉ, MARKOV, CZUBER, CASTELNUOVO) contained a discussion of the method of least squares. HAUSDORFF whose early training was in astronomy, had followed lectures in probability theory in Leipzig in 1890 by the well-known astronomer HEINRICH BRUNS (1848–1919); BRUNS' lecture notes (conserved in NL HAUSDORFF : Kapsel 55 : Fasz. 1162) contain a detailed presentation of the Gaussian least squares theory; BRUNS wrote a successful book on probability theory *Wahrscheinlichkeitsrechnung und Kollektivmasslehre*, Teubner, Leipzig 1906, which however

did not include least squares theory. HAUSDORFF's own lectures on probability theory (1900/1901) also contained a detailed description of the least squares theoy (NL HAUSDORFF : Kapsel 02 : Fasz. 10). The theory given in the present Notes is briefer but in some ways mathematically more complete; it seems clear that HAUSDORFF had thought about the theory on and off for a long time.

We shall now briefly outline the part of least squares theory treated here by HAUSDORFF; we use, essentially, HAUSDORFF's notation converted into matrix symbolism. This abbreviates the main formulae considerably and bring them into a form in which they often appear in current statistical writings. Actually, HAUSDORFF himself had written out his formulae of section 9 in matrix notation in another manuscript (NL HAUSDORFF : Kapsel 51 : Fasz. 1121 "Methode der kleinsten Quadrate in Matricenform"; 2 pages, complement to sect. 9).

Let

$$\xi = b\,\beta + x$$

where ξ, x are two $n \times 1$ matrices, b a $n \times r$ matrix ($r < n$), β a $r \times 1$ matrix (all real); the entries $\xi_1, \ldots, \xi_n$ of ξ and $x_1, \ldots, x_n$ of x are random variables, b a matrix of rank r (whose enties $b_{i\lambda}$ are known real numbers), β a vector of "unknown" parameters whose entries $\beta_1, \ldots, \beta_r$ are to be estimated on the basis of the "observations" $\xi_1, \ldots, \xi_n$. The hypotheses made on the "errors" $x_1, \ldots, x_n$ are as follows:

$$\mathbb{E}\,x_i = 0, \quad \mathbb{E}\,x_i^2 = \frac{m^2}{p_i}, \quad 1 \le i \le n,$$

where the "variance" parameter m^2 is unknown but the "weights" p_i are given ($0 < p_i < \infty,\ 0 < m < \infty$). Further, $x_1, \ldots, x_n$ are supposed to be independent (although this is not clearly spelled out); for a great deal of the calculations, only the orthogonality of the x_i's is used i. e.

$$\mathbb{E}\,(x_i\,x_j) = 0, \quad 1 \le i \ne j \le n;$$

for some calculations the existence of the 4th moments $\mathbb{E}\,x_i^4$ is needed. In matrix notation,

$$\mathbb{E}\,x = 0, \quad \mathbb{E}\,(x\,x') = m^2\,p^{-1} = \text{ variance-covariance matrix of } x$$

where

$$p = \operatorname{diag}(p_1, \ldots, p_n)$$

and $'$ indicates transposition.

The method of least squares ("méthode des moindres carrés") as given by LEGENDRE in 1805 for the estimation of β on the basis of ξ (observed) and known b consists in obtaining $\hat{\beta}$ such that

$$\inf_{\beta \in \mathbb{R}^r} (\xi - b\beta)'(\xi - b\beta) = \| \xi - b\hat{\beta} \|^2 \tag{6.1}$$

Our notation is explained by the following: b' is the transpose of b and

$$(\xi - b\beta)'(\xi - b\beta) = \| \xi - b\beta \|^2 = \sum_{i=1}^{n} \left(\xi_i - \sum_{\lambda=1}^{r} b_{i\lambda}\, \beta_\lambda \right)^2$$

The formula for $\hat{\beta}$ turns out to be

$$\hat{\beta} = (b'b)^{-1}\, b'\, \xi \tag{6.2}$$

In this solution no attention is paid to the p_i's and the random variables x_i; (6.2) is obtained by simply solving the algebraic minimization problem (6.1) which gives rise to the so-called "normal equations" (GAUSS' terminology) equivalent to the calculation of the inverse of the $r \times r$ matrix $b'b$: solve for $\hat{\beta}$ in

$$(b'b)\, \hat{\beta} = b'\, \xi \tag{6.3}$$

The solution (6.2) may be considered to be that corresponding to the case $p_i = 1,\ 1 \leq i \leq n$. For the general case, the minimization to be considered is that of

$$(\xi - b\beta)'\, p\, (\xi - b\beta) = \| \sqrt{p}\, \xi - \sqrt{p}\, b\beta \|^2 \tag{6.4}$$

which gives

$$\hat{\beta} = (b'p\, b)^{-1} b'p\, \xi \tag{6.5}$$

so that $p = e_n = (n \times n)$ identity matrix in (6.5) gives (6.2). The fact that $b'p\, b$ is invertible is a consequence of the assumption that rank $b = r$ (b being a $n \times r$ matrix and p a strictly positive diagonal matrix); this is proved in foot-note 7 of p. 711 of the Notes.

The basic theorem of the statistical theory of least squares is that $\hat{\beta}$ in (6.5) is the same as (the unique) linear unbiased, minimum variance estimator of β i. e. $\hat{\beta} = a\xi$ where a is an $r \times n$ matrix such that

$$\mathbb{E}\, \hat{\beta} = \beta \quad \text{(unbiasedness of } \hat{\beta}) \tag{6.6}$$

and the variance of each component $\hat{\beta}_\lambda$ of $\hat{\beta}$ $(1 \leq \lambda \leq r)$ is minimal where

$$\text{var } \hat{\beta}_\lambda = \mathbb{E}\, (\hat{\beta}_\lambda - \beta_\lambda)^2 = m^2 \sum_{i=1}^{n} \frac{a_{\lambda i}^2}{p_i}. \tag{6.7}$$

HAUSDORFF does not state or prove this theorem (due to GAUSS) but he simply determines the $r \times n$ matrix a by using the criteria (6.6) and (6.7). Since

$$\mathbb{E}\, (a\xi) = a\, \mathbb{E}\, \xi = a\, b\, \beta$$

(6.6) gives that

$$ab = e_r = (r \times r) \text{ identity matrix.}$$

HAUSDORFF now minimizes (6.7) under the condition $ab = e_r$ by using Lagrange multipliers; this leads to

$$a = (b'p\,b)^{-1}\,b'p \tag{6.8}$$

and to the formula (6.5) for $\hat{\beta} = a\,\xi$. Note that equations (22), (23), (24) of section 9 in HAUSDORFF's Notes correspond to

$$a = d\,b'p, \quad d\,c = e_r, \quad c = b'p\,b$$

in matrix notation whence follows (6.8) and then (6.5). HAUSDORFF writes η for $\hat{\beta}$; the notation $\hat{\beta}$ as estimate for the "unknown" parameter β is common in statistical literature where the expressions

$$\varepsilon_i = \xi_i - \sum_{\lambda=1}^{r} b_{i\lambda}\,\hat{\beta}_\lambda, \quad 1 \le i \le n \tag{6.9}$$

are known as "residuals" (HAUSDORFF's "scheinbare Beobachtungsfehler"). Let us now write down the variance-covariance matrix of $\hat{\beta}$:

$$\mathbb{E}\,(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = m^2\,(b'p\,b)^{-1} \tag{6.10}$$

This is an immediate algebraic consequence of (6.8) and the following:

$$\hat{\beta} - \beta = ax, \quad \mathbb{E}\,(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = \mathbb{E}\,(a\,x\,x'a) = m^2(a\,p^{-1}a')$$

HAUSDORFF gives only the diagonal terms of the matrix (6.10) (in the formulae (21), (25) of section 9 of his Notes). He now goes on to obtain an estimate for the variance parameter m^2; for this he uses the statistic

$$\zeta = \frac{1}{n-r}\,\sum_{i=1}^{n} p_i\,\varepsilon_i^2 \tag{6.11}$$

where the residuals ε_i are given by (6.9). He then shows by a straight-forward calculation that

$$\mathbb{E}\,\zeta = m^2 \tag{6.12}$$

and that

$$\mathbb{E}\,(\zeta - m^2)^2 = \frac{1}{(n-r)^2}\,\sum_{i=1}^{n} g_{ii}^2\,\lambda_4(x_i) + \frac{2}{n-r}\,m^4 \tag{6.13}$$

where $g = [g_{ij}]$ is the $(n \times n)$ matrix given by (cf. (31 of section 9 of the Notes)

$$g = p - p\,b\,a = p - p\,b(b'p\,b)^{-1}b'p \tag{6.14}$$

and

$$\lambda_4(x_i) = \mu_4(x_i) - 3\mu_2^2(x_i), \quad \mu_4(x_i) = \mathbb{E}\,x_i^4, \quad \mu_2(x_i) = \mathbb{E}\,x_i^2 = \frac{m^2}{p_i};$$

$\lambda_4(x_i)$ is the 4th cumulant of x_i. Thus, ζ is an *unbiased* estimator of the variance parameter m^2; further, HAUSDORFF easily establishes that under some natural conditions on the errors x_i, the right hand side in (6.13) goes to 0 as $n \to \infty$ so that ζ is a so-called *consistent* estimator of m^2 i. e. $\zeta = \zeta_n \to m^2$ in probability as $n \to \infty$. If the x_i's are all normally distributed then $\lambda_4(x_i) = 0$ and the right hand side of (6.13) simplifies to $2m^4/(n-r)$ which clearly goes to 0 as $n \to \infty$.

It is useful to specialize the above results to the case $r = 1$ which is the one with which HAUSDORFF begins his discussion; here, we write (as in HAUSDORFF)

$$\xi_i = \alpha + x_i, \quad 1 \le i \le n$$

with $\mathbb{E}\, x_i = 0,\ \mathbb{E}\, x_i^2 = m^2/p_i$. Then we obtain from the foregoing (taking $b = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$, β replaced by $\alpha,\ \xi = b\alpha + x$) that

$$\hat{\alpha} = \frac{1}{P} \sum_{i=1}^{n} p_i\, \xi_i, \quad P = \sum_{i=1}^{n} p_i$$

with

$$\mathbb{E}\, \hat{\alpha} = \alpha, \quad \mathbb{E}\, (\hat{\alpha} - \alpha)^2 = \frac{m^2}{P};$$

also,

$$\zeta = \frac{1}{n-1} \sum_{i=1}^{n} p_i\, \varepsilon_i^2, \quad \varepsilon_i = \xi_i - \hat{\alpha}, \quad \mathbb{E}\, \zeta = m^2$$

with $\mathbb{E}\, (\zeta - m^2)^2$ given by (6.13) with $r = 1$ which becomes (if $p_i = 1,\ 1 \le i \le n,\ \lambda_4(x_i) = \lambda_4$ independent of i)

$$\mathbb{E}\, (\zeta - m^2)^2 = \frac{1}{n}\, \lambda_4 + \frac{2}{n-1}\, m^4.$$

All of the results above had been given by GAUSS in his *Theoria Combinationis* (1823–28); the exact references to each of the above is given precisely in HALD [Ha 1998] (pp. 465–489; chapter 21 in general). GAUSS' earlier work *Theoria Motus* (1809) is also analized in detail in [Ha 1998] (chapter 19); HALD clearly outlines LEGENDRE's somewhat earlier independent publication of 1805 and gives an account of the ensuing GAUSS–LEGENDRE priority dispute. Clearly, GAUSS' analysis, statistically and numerically, was much more complete than LEGENDRE's, although LEGENDRE's prior publication of the theory as a non-statistical solution to a problem in interpolation theory had been apparently widely appreciated by many practitioners of the period.

Naturally, GAUSS' analysis did not use any matrices; however, as HALD and others have pointed out, the use of the formalism of matrices clarifies much of the complicated algebra without using any of the theorems of the theory of matrices. HAUSDORFF's Notes work out everything using elementary algebra

although as he himself had realized (as indicated above) that the results can be written down more succintly by using matrices.

Many generalizations of the above theory exist in current statistical literature; one may relax the condition of rank $b = r$ and try to estimate other linear functions of the β; the case where the variance-covariance matrix of the error vector x is some general positive definite matrix (other than m^2 diag $(p_1^{-1}, \ldots, p_n^{-1})$) has also been studied. Some of these references are given in [Ha 1998]; others can be found in standard statistical texts like KENDALL and STUART [KS 1973], chapter 19. The statistical literature around the least squares theory, both theoretical and practical, is immense and contains many of the most important statistical methods like regression analysis as well as analysis of variance and covariance.

The theory outlined in HAUSDORFF's Notes is sometimes subsumed by statisticians under the general title of the Gauss-Markov theorem; HALD [Ha 1998] p. 471 rightly (it seems to us) points out the inappropriateness of this appellation in so far as all the results were given already by GAUSS. HALD makes interesting remarks on GAUSS' unbiased estimator ζ of the variance parameter m^2 (cf. (6.11) above) and the formula for the variance of ζ (cf. (6.13) above); the formula due to GAUSS reproduced by HALD ([Ha 1998] p. 477, eqn. (1)) is the special case of our (6.13) with $p_i = 1,\ 1 \leq i \leq n$. HALD remarks that this formula for the variance of ζ seems to have "disappeared from the literature" (ibid. p. 479) except in the case $r = 1$; it was therefore interesting to see the general formula given (and derived in detail) in these Notes. The estimation of the variance parameter m^2 had obviously intrigued HAUSDORFF since he refers to this problem specially (p. 712) regretting the fact that a minimum variance estimate of m^2 must involve much calculation ("mit grösserer Rechnung verknüpft"); in this direction, he gives a reference to a 1892/93 paper by BRUNS; a study of the latter does not reveal a clear solution to the problem raised. Some remarks on this are given in HALD [Ha 1998] p. 480 with a reference to PLACKETT (1960).

Remarks on the Exercises

The exercises are mostly standard and concern the elementary probability theory discussed in sections 1 and 2 of the Notes; their solutions are given in some detail. Exercise 11 concerning the kth decimal digit of $\log_{10} x$ (and its generalization) is somewhat different; a related problem appears in PÓLYA–SZEGÖ *Aufgaben und Lehrsätze aus der Analysis* I (II. Abschnitt, 178–181) and is attributed to J. FRANEL. The reference to URBAN in exercise 11 concerns: F. M. URBAN, *Grundlagen der Wahrscheinlichkeitsrechnung und der Theorie der Beobachtungsfehler*. Teubner, Leipzig 1923.

§ 7 Conclusion

The Notes present a good course in probability theory at a high level of mathematical precision. Some of the material presented was either new or else

not easily accessible in the 1920's. The construction of the probability measures in $\mathbb{R}^n$ and the associated integration theory are given very efficiently. The treatment of the weak and strong laws is exact albeit unnecessarily restricted to simple random variables (taking only finitely many distinct values); this was already an improvement when his *Grundzüge* (1914) was published. HAUSDORFF's presentation of the central limit theorem in LYAPOUNOV's form using LINDBERG's method was a definite advance over most published accounts in the books of the 1920's. No wonder CRAMÉR writes in his autobiographical remarks (*Half a century with probability theory, some personal recollection*, Ann. Prob. 4 (1976), 509–546, cf. 512; in [C 1994], vol. II, p. 1355):

> The work of Liapounov was very little known outside Russia, but I had the good luck to be allowed to see some notes on his work made by the German mathematician Hausdorff, and these had a great influence on my subsequent work in the field.

Another novelty of the Notes is the treatment of the generalized Chebyshev–Markov limit theorem by using the new methods of M. RIESZ. Finally, the least squares theory was treated succinctly but thoroughly.

We must now underline the main weaknesses of the Notes from the point of view of modern theory; these were perhaps inevitable since the theory of probability had not yet attained the mathematical firmness that it was to acquire after the publication in 1933 of KOLMOGOROV's famous axiomatization (see [GK 1954] for exact reference). To pin-point the two major shortcomings of these Notes we must indicate the modern formulation of probability theory used universally in mathematical discourse. A probability space is a triple $(\Omega, \Sigma, \mathbb{P})$ where Ω is some set, Σ is a σ-algebra of subsets of Ω and $\mathbb{P} : \Sigma \to [0, 1]$ is a σ-additive measure (called probability) with $\mathbb{P}(\Omega) = 1$. A real-valued random variable is a measurable map $X : \Omega \to \mathbb{R}$; its law or distribution is given by the probability measure μ_X induced by X in $\mathbb{R}$ i. e. $\mu_X(A) = \mathbb{P}(X^{-1}A)$, A being a Borel subset of $\mathbb{R}$. Analogously, a measurable map $X : \Omega \to \mathbb{R}^n$ defines a random vector; other types of random elements and their laws are defined as in the real-valued case. All this must have been clear (or almost clear) to many around the 1920's but rarely spelled out; most writers (as in these Notes) simply contented themselves with the laws μ_X defined in $\mathbb{R}^n$ and even this only in terms of the so-called cumulative distribution functions. But without a clear definition of random variables, even very simple notions like convergence in probability or convergence almost surely require impossible circumlocutions. The obvious fact that two real random variables X, Y can have the same law but be always unequal does not become evident. But it is not enough to have the necessary definitions concerning random variables; these, after all, although providing an enormous terminological convenience, follow the usual well-understood theory of measurable functions clearly established since LEBESGUE's work.

Another crucial element is a proof of the existence of suitable probability spaces to accomodate interesting theories. Thus, in order to study a sequence of independent real-valued random variables $X_1, X_2, \ldots$ with prescribed laws

$\mu_1, \mu_2, \cdots$ one must establish a theorem guaranteering the existence of a probability space $(\Omega, \Sigma, \mathbb{P})$ and appropriate measurable functions $X_n : \Omega \to \mathbb{R}$. This can now be ensured by taking

$$\Omega = \mathbb{R} \times \mathbb{R} \times \cdots, \quad \mathbb{P} = \mu_1 \otimes \mu_2 \otimes \cdots, \quad X_n(\omega_1, \omega_2, \ldots) = \omega_n, \quad n = 1, 2, \ldots$$

However, the existence of the product measure $\mathbb{P}$ was not known until much after 1923 and several erroneous proofs had circulated before 1930. Indeed, a major theorem in KOLMOGOROV's monograph proved (for the first time) the existence of a suitable probability space made out of $\mathbb{R}^T$ (for any index set T) which would accomodate any stochastic process whose description was not self-contradictory. HAUSDORFF must have felt the need for such theorems when he mentions in his section 5 that "Die bisherigen Betrachtungen, insbesondere die von § 4, schweben insofern noch in der Luft ..." Here he was thinking of having a probability space which would at least accomodate a sequence of independent events $\{A_n\}$ with assigned probabilities $w(A_n) = p_n$; this he could have easily obtained in $[0, 1]$ but there is no evidence that he ever tried to do this.

Leaving aside the problem of the existence of suitable probability spaces which concerns the solution of a well-defined mathematical problem, we must add a few words on the strange avoidance of a proper definition of the very notion of random variables on the part of almost all authors of books and monographs well into the 1950's. For them a random variable (a chance variable or some such term) was simply given by some probability distribution in $\mathbb{R}$ (or $\mathbb{R}^n$) and in case of $\mathbb{R}$, this would be specified by a monotone non-decreasing function $F : \mathbb{R} \to [0, 1]$ (with $F(-\infty) = 0$, $F(\infty) = 1$), $F(x)$ being interpreted as the probability that the variable concerned is $< x$ (or $\leq x$). This is the point of view in HAUSDORFF's Notes and this was the accepted way of discussing things until 1933; even later books like LÉVY's famous *Théorie de l'addition des variables aléatoires* (1937) or CRAMÉR's *Mathematical methods of statistics* (1946) (see [GK 1954] for exact references) or FRÉCHET's 1950 volume ([Fr 1950]) do not consider it necessary to provide any mathematical definition of a random variable. Indeed, GNEDENKO and KOLMOGOROV point this out explicitly as regards CRAMÉR's book ([GK 1954], p. 13); DOOB's appendix in [GK 1954] clearly spells out the necessity of a formal definition of random variables and allied matters and his own monograph [D 1953] sets up standards of rigour in probability theory which are now universally accepted. Of course, this avoidance of exact definitions was not unique to probability theory; for long periods, fundamental notions like those of real numbers, vectors, tensor products, manifolds etc. were generally left in a state of "flou artistique".

These Lecture Notes of HAUSDORFF represent the utmost limits of understanding and clarification of mathematical probability to which HAUSDORFF had attained in 1923–1933. A study of his abundant Nachlaß on that subject and related matters (some published in this volume) does not indicate any progress on his part beyond what is perceptable here. HAUSDORFF sees clearly that mathematical probability is a branch of measure and integration theory but fails to make any decisive steps beyond this recognition. Further, his ma-

stery of the limit theorems of the theory, both in their formulation as well as in their proofs, remains bounded by what is seen here in these Notes. The decisive changes brought about in probability theory by the spate of progress in that theory in the 1930's remained unperceived by him.

Appendices

For the two appendices which follow we shall use the standard modern probabilistic notations and definitions both to describe the older results as well as to state the present day state of knowledge. Recall that underlying any probabilistic discourse, there is a probability triple $(\Omega, \Sigma, \mathbb{P})$ (as explained in the preceding section 7); if $X : \Omega \to \mathbb{R}$ is a real-valued random variable we write

$$\mathbb{E}\, X = \int_\Omega X \, d\mathbb{P}, \quad \mathbb{E}\,(X; A) = \int_A X \, d\mathbb{P} \quad \text{if } \ A \in \Sigma;$$

if $\mu = \mu_X$ is the law (or distribution) of X and $F(x) = F_X(x) = \mathbb{P}(X \leq x),\ x \in \mathbb{R}$, is the associated (cumulative) distribution function then

$$\mathbb{E}\, X = \int_{-\infty}^{\infty} x \, d\mu(x) = \int_{-\infty}^{\infty} x \, dF(x),$$

$$\mathbb{E}\,(X; X \in B) = \int_B x \, d\mu(x) = \int_B x \, dF(x)$$

where B is a Borel subset of $\mathbb{R}$; more generally, if $f : \mathbb{R} \to \mathbb{R}$ is a Borel function then

$$\mathbb{E}\, f(X) = \int_{-\infty}^{\infty} f(x) \, d\mu(x) = \int_{-\infty}^{\infty} f(x) \, dF(x)$$

where $f(X) = f \circ X$. In the sequal when we write $\mathbb{E}\, X = 0$, we shall tacitly assume that $\mathbb{E}\,|X| < \infty$; more genarally, whenever we write $\mathbb{E}\, f(X)$ it will be assumed that $\mathbb{E}\,|f(X)| < \infty$. References to older papers not explicitly given are to be found in [BN 1977] or in the standard recent books like [CT 1988], [P 1995].

Appendix A

Concerns strong laws and Borel's proof of his strong law

(i) Let $X_1, X_2, \ldots$ be a sequence of real-valued random variables; strong laws concern the almost sure (a. s.) behaviour of the partial sums

$$S_n = X_1 + \cdots + X_n.$$

The so-called weak laws study the same problem for convergence in probability (i. e. convergence in measure). Of course, the study can be (and in recent years, has been) extended to the case of vector-valued and other types of random

variables; however, we shall not say anything about this. Most of the results concern the existence of numerical sequences $\{a_n\}$, $\{b_n\}$, $0 < b_n$, such that

$$\text{(a)} \quad \lim_{n\to\infty} \left\{ \frac{S_n - a_n}{b_n} \right\} \quad \text{exists a. s.} \qquad \text{(b)} \quad 0 < \limsup_{n\to\infty} \frac{|S_n - a_n|}{b_n} < \infty \ \text{ a. s.}$$

$$\text{(c)} \quad \sum_{n=1}^{\infty} c_n \, \mathbb{P}\left(\frac{|S_n - a_n|}{b_n} > \varepsilon \right) < \infty \quad \text{for some} \quad c_n > 0, \, \varepsilon > 0.$$

Naturally, the statements (a), (b), (c) are related to each other in various ways. The cases of (a) and (c) where $b_n = n$, $a_n = \mathbb{E}\, S_n$ and the X_j's are mutually independent are the ones which have been most intensively studied; a statement like (b) leads to the classical law of the iterated logarithms, so-called because in the very important case where the X_j's are independent and identically distributed with $\mathbb{E}\, X_j = 0$, $\mathbb{E}\, X_j^2 < \infty$, (b) holds with $a_n = 0$, $b_n = \sqrt{n \log\log n}$ (theorem of Hartmann–Wintner (1941)). For the purposes of the present discussion, let us recall two classical results concerning (a) and (c) for the case of independent, identically distributed (i. i. d.) random variables X_j, $j \geq 1$. The first is the theorem of Kolmogorov (1930): if $\{X_j\}$ is a sequence of i. i. d. random variables then

$$\lim_{n\to\infty} \frac{S_n}{n} = 0 \quad \text{a. s.} \tag{A.1}$$

if and only if $\mathbb{E}\, X_j = 0$. The second is a theorem due to Hsu and Robbins (1947), Erdös (1949): if $\{X_j\}$ is a sequence of i. i. d. random variables then

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left| \frac{S_n}{n} \right| > \varepsilon \right) < \infty \tag{A.2}$$

for every $\varepsilon > 0$ if and only if $\mathbb{E}\, X_j^2 < \infty$ and $\mathbb{E}\, X_j = 0$. There are, of course, numerous sharpenings and generalizations of these two theorems but we mention these in particular in order to illustrate the limitations of the method used by Hausdorff to prove his strong law. It is an immediate corollary of the easy half of the Borel-Cantelli lemma (see section 3 above) that if $\{X_j\}$ is *any* sequence of real-valued random variables then the validity of (A.2) for every $\varepsilon > 0$ implies (A.1); this was the strategy used by Hausdorff to obtain his strong laws. The above-mentioned theorems indicate that this method by itself could not possibly yield (A.1) under general hypotheses like those of Kolmogorov. Hausdorff's technique (in section 2 of his Notes, cf. calculations after theorem IV) consists in first deriving a 4-th moment bound like

$$\mathbb{E}\, |S_n|^4 \leq C \cdot n^2 \tag{A.3}$$

(C some positive constant) from where (A.2) will follow via the simple (so-called Markov) inequality:

$$\mathbb{P}\left(\left| \frac{S_n}{n} \right| > \varepsilon \right) \leq \frac{1}{\varepsilon^4} \, \mathbb{E} \left| \frac{S_n}{n} \right|^4 \leq \frac{C}{n^2 \, \varepsilon^4}$$

(which is equation (17) in HAUSDORFF's Notes, section 2). Thus HAUSDORFF's arguments would allow him to prove the strong law (A.1) under the hypotheses

$$\mathbb{E}\, X_j = 0, \quad \sup_j \mathbb{E}\, X_j^4 < \infty$$

the sequence $\{X_j\}$ being formed of independent random variables (not necessary i. i. d.). HAUSDORFF does not however state this explicitly in the Notes but it is this argument (which we shall call the 4th-moment argument) that appeared first in his *Grundzüge* (pp. 420–421) and later in NL HAUSDORFF : Kapsel 44 : Fasz. 833 (1916) (printed in this volume, pp. 768–775). The 4th-moment argument was used, independently, by CANTELLI (1917) (cf. [BN 1977] for exact reference) and has been reutilized by different authors including HAUSDORFF himself (cf. NL HAUSDORFF : Kapsel 34 : Fasz. 372 (1928–31) reprinted in [H 2002], pp. 819–823).

The Borel strong law treated by HAUSDORFF in his *Grundzüge* concerned the special case of $X_j = x_j - \frac{1}{2}$, the sequence $\{x_j\}$ being i. i. d. with $x_j = 0$ or 1 with probability $\frac{1}{2}$. At the end of the proof, HAUSDORFF remarked (loc. cit. p. 421) that (in our notation)

$$\lim_{n\to\infty} n^\theta \left(\frac{x_1 + \cdots + x_n}{n} - \frac{1}{2} \right) = \lim_{n\to\infty} \frac{S_n}{n^\alpha} = 0 \quad \text{a. s.} \tag{A.4}$$

for any $\theta < \frac{1}{2}$ (i. e. $\alpha = 1 - \theta > \frac{1}{2}$); recall that here

$$S_n = X_1 + \cdots + X_n = (x_1 + \cdots + x_n) - \frac{n}{2}.$$

This remark of HAUSDORFF has been reported by several authors (e. g. FELLER [F 1968], p. 209); however, we have found no indication of any proof of this result in any of HAUSDORFF's papers. It is possible that HAUSDORFF had imagined a proof by using moments of the order p where one would first establish that (for some positive constant C_p)

$$\mathbb{E}\, |S_n|^p \le C_p\, M^p\, n^{p/2} \tag{A.5}$$

for X_j's bounded by M ($|X_j| \le M$), X_j's being independent with $\mathbb{E}\, X_j = 0$. Then the simple reasoning that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left| \frac{S_n}{n^\alpha} \right| > \varepsilon \right) \le \sum_{n=1}^{\infty} \varepsilon^{-p}\, \mathbb{E}\, \left| \frac{S_n}{n^\alpha} \right|^p \le \text{ const. } \sum_{n=1}^{\infty} n^{-p(\alpha - 1/2)} < \infty$$

if $p(\alpha - 1/2) > 1$ will establish (A.4) just as (A.1) was proved via the 4th-moment argument above. Inequality (A.5) (at least in the case $X_j = x_j - \frac{1}{2}$ as above) was within the reach of HAUSDORFF's techniques as shown by STEINHAUS (1923) (cf. [BN 1977] for exact reference). Indeed, MARCINKIEWICZ and ZYGMUND (1937) (cf. [M 1964], p. 257) proved that if only the X_j's are

independent, $\mathbb{E}\,X_j = 0,\ \mathbb{E}\,|X_j|^p < \infty,\ j \geq 1,\ p > 1$, then (for some positive constant $C_p < \infty$)

$$\mathbb{E}\,|S_n|^p \leq C_p\,\mathbb{E}\,\left\{(X_1^2 + \cdots + X_n^2)^{p/2}\right\}$$

whence can be derived easily that

$$\mathbb{E}\,|S_n|^p \leq C_p \cdot M \cdot n^{p/2}$$

under the sole hypothesis that the X_j's are independent and

$$\mathbb{E}\,X_j = 0, \quad \sup_j \mathbb{E}\,|X_j|^p = M < \infty,$$

p being a constant ≥ 2; it suffices to use the elementary fact that from JENSEN's convexity inequality for real numbers, one has

$$(X_1^2 + \cdots + X_n^2)^{p/2} \leq n^{p/2}\ \frac{|X_1|^p + \cdots + |X_n|^p}{n}, \quad p \geq 2.$$

Thus (A.5)-like inequalities can be obtained under various hypotheses (including some which relax even the condition of independence of the X_j's considerably) leading to (A.4) via the reasoning indicated above. No such simple argument would seem to lead to the more refined laws of the iterated logarithms mentioned before.

The literature around the problems indicated above is immense; a recent volume by PETROV [P 1995] (cf. also [P 1975]) surveys the case of sums of independent random variables; a volume by STOUT [S 1974] treats more general sums; KASHIN and SAAKYAN [KaS 1989] report on the case of sums of orthogonal random variables. Several advanced modern text-books devote much space to these problems (e. g. CHOW and TEICHER [CT 1988]). Further references can be found in these volumes.

(ii) In this paragraph, we discuss separately BOREL's proof of his strong law which left HAUSDORFF so unconvinced ("principiell ganz unklar", section 4 (p. 631) of HAUSDORFF's Notes). BOREL's original paper containing the proof (and much other material) is his often cited article *Les probabilités dénombrables et leurs applications arithmétiques*, Rendiconti del Circolo Matematico di Palermo 27 (1909), 247–271. This paper has been analysed in detail by BARONE and NOVIKOFF [BN 1977] (of which Part II apparently never appeared) who have rightly pointed out all the important flaws in BOREL's work and given some indications of later improvements, corrections and adjustments due to others. Here we shall concentrate on its one major weakness which cannot be repaired by a simple appeal to a more careful use of measure theory. In order to do thus, we must briefly but precisely recall the point at stake.

BOREL's theorem in question can be formulated as follows:

$$\mathbb{P}\left(\lim_{n\to\infty}\ \frac{S_n}{n} = \frac{1}{2}\right) = 1 \tag{A.6}$$

where $S_n = X_1 + \cdots + X_n$, $X_i = 0$ or 1 with probability $\frac{1}{2}$, $i \geq 1$, the X_i's being independent. In BOREL's discussion, X_i is realized as the ith digit of the binary expansion of a number in [0, 1] and $\mathbb{P}$ is simply the Lebesgue measure; this interpretation, however, plays no rôle in BOREL's proof which proceeds as follows (in our notation): let

$$p_n = \mathbb{P}(n - \lambda_n \sqrt{n} \leq S_{2n} \leq n + \lambda_n \sqrt{n}), \quad n \geq 1$$

where $\{\lambda_n\}$ is any sequence of real numbers such that

$$0 < \lambda_n \uparrow \infty, \quad \lim_{n\to\infty} \frac{\lambda_n}{\sqrt{n}} = 0 \tag{A.7}$$

(for example, says BOREL, $\lambda_n = \log n$). The crucial point in BOREL's proof is that if $q_n = 1 - p_n$ then

$$\sum_{n=1}^{\infty} q_n < \infty \tag{A.8}$$

Assuming (A.8), the proof of (A.7) is easy and given correctly by BOREL by arguing that (A.8) implies that with probability 1 (again via the easy half of the Borel-Cantelli lemma)

$$n - \lambda_n \sqrt{n} \leq S_{2n}(\omega) \leq n + \lambda_n \sqrt{n}, \quad n \geq n_0 = n_0(\omega)$$

whence

$$\frac{1 - \lambda_n/\sqrt{n}}{1 + \lambda_n/\sqrt{n}} \leq \frac{2n - S_{2n}}{S_{2n}} \leq \frac{1 + \lambda_n/\sqrt{n}}{1 - \lambda_n/\sqrt{n}}, \quad n \geq n_0$$

which gives, in view of (A.7),

$$\lim_{n\to\infty} \frac{S_{2n}}{2n} = \frac{1}{2} \quad \text{almost surely ;}$$

since $|X_i| \leq 1$, this implies

$$\lim_{n\to\infty} \frac{S_n}{n} = \frac{1}{2} \quad \text{almost surely}$$

which is (A.6). The most serious lacuna in BOREL's argument is in the proof of (A.8); here, BOREL assumes with unbelievable negligence that p_n is *equal* to

$$\Theta(\lambda_n) = \frac{2}{\sqrt{\pi}} \int_0^{\lambda_n} e^{-x^2}\, dx.$$

In other words, the de Moivre–Laplace limit statement (for a fixed $\lambda > 0$)

$$\lim_{n\to\infty} \mathbb{P}(n - \lambda\sqrt{n} \leq S_{2n} \leq n + \lambda\sqrt{n}) = \frac{2}{\sqrt{\pi}} \int_0^{\lambda} e^{-x^2}\, dx = \Theta(\lambda)$$

is stretched to mean that $p_n = \Theta(\lambda_n)$ even for $\lambda_n \uparrow \infty$ in certain ways! It is, of course, easy to show that

$$1 - \Theta(\lambda_n) \leq \frac{1}{\lambda_n \sqrt{\pi}}\, e^{-\lambda_n^2}$$

so that

$$\sum_{n=1}^{\infty} \{1 - \theta(\lambda_n)\} < \infty \tag{A.9}$$

whenever

$$\sum_{n=1}^{\infty} \frac{1}{\lambda_n}\, e^{-\lambda_n^2} < \infty \tag{A.10}$$

which certainly holds if $\lambda_n = \log n$ (but not if $\lambda_n = \sqrt{\log n}$). Note that since, for $\lambda \to \infty$,

$$1 - \Theta(\lambda) \sim \frac{1}{\lambda \sqrt{\pi}}\, e^{-\lambda^2}$$

the conditions (A.9) and (A.10) on $\{\lambda_n\}$ are equivalent; this remark shows that for the validity of either of them some more care in the choice of $\lambda_n \uparrow \infty$ than simply demanding (A.7) must be exercised. However, this is a minor point since there are many possible choices of λ_n verifying (A.7) and (A.10) e. g. $\lambda_n = (\log n)^{\alpha},\ \alpha > 1/2$ or $\lambda_n = n^{\alpha},\ 0 < \alpha < 1/2$. The essential difficulty in carrying out BOREL's proof is in choising $\{\lambda_n\}$ satisfying (A.7) in such a way that (A.8) holds. If we can show that $\{\lambda_n\}$ can be chosen so that (A.7) and (A.10) hold and

$$\limsup_{n\to\infty} \frac{q_n}{1 - \Theta(\lambda_n)} < \infty \tag{A.11}$$

then (A.8) will be guaranteed and BOREL's proof will have been completed. This can be done but the work is laborious. Indeed, it can be shown that if $\lambda_n \to \infty$ in such a way that

$$\frac{\lambda_n}{n^{1/6}} \to 0 \quad \text{as} \quad n \to \infty$$

then

$$\lim_{n\to\infty} \frac{q_n}{1 - \Theta(\lambda_n)} = 1 \tag{A.12}$$

For the special case of binomial probabilities involved here, this is given in FELLER [F 1968], p. 193 and in many other places; this is a very special case of a more general (so-called "large deviation") result of CRAMÉR from 1938 ([C 1994], vol. II, p. 905). Hence, a choice of $\{\lambda_n\}$ satisfying (A.7), (A.10) and (A.12) is possible, thus validating BOREL's proof.

The first attempt at justifying BOREL's proof which we have seen is in FRÉCHET's book [Fr 1950] (1st ed. 1937); here, after having stated (p. 231) that "La démonstration de M. Borel est excessivement brève", FRÉCHET goes

on to complete the proof in the way we have indicated above, his main difficult step being that of arriving at (A.11) (and eventually even to (A.12)) via some elementary but tedious estimation of binomial probabilities. FRÉCHET also points out (ibid. p. 236) a paper of CRAMÉR (1934) which gives the conclusion that for any choice of $\lambda_n \uparrow \infty$, (A.8) holds if and only if (A.10) holds; again, CRAMÉR's result is valid more generally than just in the case of Bernoulli probabilities ([C 1994], vol. I, p. 700). This last is again obtained by a "large deviation" estimate whose nature we shall explain later in another context (cf. commentary on NL HAUSDORFF : Kapsel 44 : Fasz. 834 printed in this volume, p. 776–790). FRÉCHET also writes out (ibid. p. 239) HAUSDORFF's proof as given in the *Grundzüge* whose simplicity and clarity we have already indicated. The FRÉCHET type proof of BOREL's theorem now appears as a curiosity in some books e. g. [CT 1988], exercises 5 and 6, p. 52.

BARONE and NOVIKOFF [BN 1977] rightly conclude that HAUSDORFF's 1914 proof is the first complete, correct and explicit proof of BOREL's strong law; they also aptly indicate (ibid. p. 171) a proof of FABER (1910) and the related later independent proof of RADEMACHER (1918) based on LEBESGUE's differentiation theorem; however, they correctly point out that FABER was uncertain of the relationship of his result with that of BOREL's; RADEMACHER mentions HAUSDORFF's proof as well as that of FABER in a short supplement to his original paper. [BN 1977] has the reference to FABER's paper but not that of RADEMACHER's. RADEMACHER published three influential papers related to almost everywhere convergence. The paper related to BOREL's strong law is (1) *Zu dem Borelschen Satze über die asymptotische Verteilung der Ziffern in Dezimalbrüchen*, Math. Zeitschrift 2 (1918), 306–311 (in [Ra 1974], pp. 123–128). The two other papers concern the convergence almost everywhere of orthogonal series: (2) *Über die asymptotische Verteilung gewisser konvergenzerzeugender Faktoren*, Math. Zeitschrift 11 (1921), 276–288 (in [Ra 1974], pp. 196–208); (3) *Einige Sätze über Reihen von allgemeinen Orthogonalfunktionen*, Math. Annalen 87 (1922), 112–138 (in [Ra 1974], pp. 231–257). This last introduces the so-called Rademacher functions; we shall see that these were independently introduced by HAUSDORFF earlier in an unpublished manuscript (NL HAUSDORFF : Kapsel 44 : Fasz. 861, dated 2. 3. 1915); we shall report on this later in this volume, pp. 757–760.

In conclusion, we note that BOREL's interesting but very incomplete proof replaced the problem of proving the rather easy strong law (A.6) by the much more intricate one of the equivalence of the conditions (A.8) and (A.10).

Appendix B
Lindeberg's condition; Berry-Esseen theorem

(i) The purpose of this appendix is to give exactly what was proved in LINDEBERG's 1922 paper to which HAUSDORFF refers in his Notes. Although the paper has been cited by many authors in numerous books and articles, none of them seem to have pointed out LINDEBERG's original formulations of his fa-

mous condition; besides, very few seem to have indicated exactly LINDEBERG's own method of proof. As regards the latter, HAUSDORFF's account is very close to LINDEBERG's; since HAUSDORFF does not dicuss LINDEBERG's general conditions in any form, it seemed useful to state them in their original form (in the following paragraph (ii)). In (iii) we indicate the rates of convergence to the normal law in the central limit theorem known today in order to compare them with what HAUSDORFF achieves in his Notes.

We shall use, essentially, LINDEBERG's notations except for transcribing them, using expectations ($\mathbb{E}$), as indicated before.

(ii) LINDEBERG considers a finite sequence of independent real-valued random variables $X_1, X_2, \ldots, X_n$ which are square integrable and are such that

$$\mathbb{E}\, X_j = 0, \quad \mathbb{E}\, X_j^2 = \sigma_j^2, \quad 1 \le j \le n. \tag{B.1}$$

His first theorem is stated as follows: suppose further that

$$\sum_{j=1}^{n} \sigma_j^2 = 1, \quad \mathbb{E}\,|X_j|^3 < \infty, \quad 1 \le j \le n;$$

then for any $\varepsilon > 0$ there exists a number $\eta > 0$ such that

$$\left| \mathbb{P}(X_1 + \cdots + X_n \le x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \right| < \varepsilon, \quad x \in \mathbb{R} \tag{B.2}$$

whenever

$$\sum_{j=1}^{n} \mathbb{E}\,|X_j|^3 < \eta.$$

In fact, LINDEBERG proves more precisely (by his simple direct method) that

$$\left| \mathbb{P}(X_1 + \cdots + X_n \le x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \right| < 3 \left(\sum_{j=1}^{n} \mathbb{E}\,|X_j|^3 \right)^{1/4} \tag{B.3}$$

(under the hypotheses (B.1) and $\sum_{j=1}^{n} \sigma_j^2 = 1$). The constant 3 in the estimate (B.3) is just a convenient choice, no attempt having been made to make it any smaller; of course, (B.3) is the quantitative form of LINDEBERG's first theorem and is what HAUSDORFF obtains (see formula (22) of his Notes, section 7) except for his unprescribed constant μ. HAUSDORFF's method is almost the same as LINDEBERG's. LINDEBERG was obviously motivated by LYAPOUNOV's theorem to which he refers in his paper; he then obtains a version of (B.2) without the hypothesis $\sum_j \sigma_j^2 = 1$ which is just obtained by a rescaling. From this he derives his second theorem concerning bounded X_i's which he states as follows: suppose that $|X_i| \le d_n,\ 1 \le i \le n$ and let $r_n^2 = \sigma_1^2 + \cdots + \sigma_n^2$, the X_i's still fulfilling (B.1); then, for any $\varepsilon > 0$, there exists $\eta > 0$ such that

$$\left| \mathbb{P}(X_1 + \cdots + X_n \le x) - \frac{1}{r_n \sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(\frac{-t^2}{2r_n^2} \right) dt \right| < \varepsilon, \quad x \in \mathbb{R}$$

provided that $\frac{d_n}{r_n} < \eta$.
LINDEBERG considered his second theorem as important ("für die mathematische Statistik unbedingt notwendig") although it is, of course, derived as an immediate consequence of his first theorem. Now LINDEBERG moves on to the general case of X_j's (independent) satifying just (B.1). Here he states his third theorem as follows: suppose that $\sum_{j=1}^n \sigma_j^2 = 1$ and let

$$s(x) = \begin{cases} |x|^3 & \text{if} \quad |x| < 1 \\ x^2 & \text{if} \quad |x| \geq 1 \end{cases} ;$$

then for any $\varepsilon > 0$, there exists a number $\eta > 0$ such that (B.2) holds as soon as

$$\sum_{j=1}^n \mathbb{E}\{s(X_j)\} < \eta.$$

Here again if we followed LINDEBERG's suggestions (cf. his paper, p. 221) we would get the following inequality which is a more precise form of his third theorem:

$$\left| \mathbb{P}(X_1 + \cdots + X_n \leq x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du \right| < 3 \left\{ \sum_{j=1}^n \mathbb{E}\, s(X_j) \right\}^{1/4} \tag{B.4}$$

Again, the constant 3 on the right hand side above is a mere convenience. LINDEBERG now points out that the function s used in the third theorem above can be replaced by

$$s_\rho(x) = \begin{cases} |x|^3 & \text{if} \quad |x| \leq \rho \\ \rho x^2 & \text{if} \quad |x| > \rho \end{cases}$$

where ρ is any positive number. More suggestively, he states that his third theorem can be expressed in other forms; in order to do this let us introduce the following notation (not in LINDEBERG's paper): for independent real random variables $X_1, \ldots, X_n$ satisfying (B.1), write

$$L_1 = L_1(X_1, \ldots, X_n) = \textstyle\sum_{j=1}^n \mathbb{E}\,|X_j|^3$$

$$L_2 = L_2(X_1, \ldots, X_n) = \textstyle\sum_{j=1}^n \mathbb{E}\, s(X_j)$$

$$L_3 = L_3(X_1, \ldots, X_n) = 1 - \textstyle\int_0^1 d\tau \sum_{j=1}^n \mathbb{E}\,(|X_j|^2;\ |X_j| \leq \tau)$$

Under the hypothesis that $\sigma_1^2 + \cdots + \sigma_n^2 = 1$, LINDEBERG now shows that L_3 is small if and only if L_2 is small; indeed, it is easily shown that if $0 < \varepsilon < 1$ then

$$L_2 < \varepsilon \quad \Rightarrow \quad L_3 < 2\sqrt{\varepsilon}$$

and

$$L_3 < \varepsilon \quad \Rightarrow \quad L_2 < 2\sqrt{\varepsilon}.$$

Thus, LINDEBERG can restate his third theorem as his fourth in the following form: if independent random variables $X_1, \ldots, X_n$ satisfy (B.1) and $\sigma_1^2 + \cdots + \sigma_n^2 = 1$ then for any $\varepsilon > 0$ there exists $\eta > 0$ such that (B.2) holds whenever $L_3 < \eta$.

By obvious centering and rescaling, LINDEBERG now states his general fifth and last theorem for any square integrable independent real-valued random variables $X_1, \ldots, X_n$ with

$$\mathbb{E}\, X_j = b_j, \quad B_n = b_1 + \cdots + b_n, \quad \mathbb{E}\,(X_j - b_j)^2 = \sigma_j^2, \quad \sum_{j=1}^{n} \sigma_j^2 = r_n^2.$$

His final statement for the infinite sequence $\{X_j\}$ then is equivalent to the following:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_1 + \cdots + X_n - B_n}{r_n} \le x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du$$

uniformly in $x \in \mathbb{R}$ if

$$\lim_{n\to\infty} L_3\left(\frac{X_1 - b_1}{r_n}, \ldots, \frac{X_n - b_n}{r_n}\right) = 0 \tag{B.5}$$

where we have used LINDEBERG's symbols except for using $\mathbb{P}$ for probability and defining L_3 using expectation symbol $\mathbb{E}$. If we take $B_j = 0,\ j \ge 1$, then (B.5) is the same as having

$$\lim_{n\to\infty} \frac{1}{r_n^2} \int_0^1 d\tau \sum_{j=1}^{n} \mathbb{E}\,(|X_j|^2;\, |X_j| > \tau r_n) = 0 \tag{B.6}$$

From (B.6) we obtain the usual (equivalent) Lindeberg condition written as follows: for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{r_n^2} \sum_{j=1}^{n} \mathbb{E}\,(|X_j|^2;\, |X_j| > \varepsilon r_n) = 0 \tag{B.7}$$

It has been pointed out by CHOW and TEICHER ([CT 1988], pp. 295–296) that (B.7) (for all $\varepsilon > 0$) is also equivalent to the following: for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{r_n^2} \sum_{j=1}^{n} \mathbb{E}\,(|X_j|^2;\, |X_j| > \varepsilon r_j) = 0.$$

In all this it is, of course, tacitly supposed that $r_n > 0$ for some n onwards.

It is surprising that LINDEBERG in his elegant paper containing one of the most general forms of the central limit theorem does not mention its validity for the simple special case of independent *identically distributed* real-valued random variables $X_1, X_2, \ldots$ with $\mathbb{E}\, X_i = 0,\ \mathbb{E}\, X_i^2 = \sigma^2,\ 0 < \sigma < \infty$; in

this case, his condition (in the form (B.6) or (B.7)) can be verified easily. This important special case is given in HAUSDORFF's Notes under the extra assumption that $\mathbb{E}\,|X_i|^3 < \infty$; it remained for LÉVY to state this extremely useful theorem in his 1925 book *Calcul des probabilités*. LÉVY's analysis is based on characteristic functions but he also adapted LINDEBERG's reasoning in his own original way. We shall not analyse LÉVY's technique from the point of view of LINDEBERG's method; suffice it to say that LÉVY took good notice of LINDEBERG's paper as soon as it appeared and went on to his own independent development leading to stable laws and other central limit theorems. The well-known book by GNEDENKO and KOLMOGOROV [GK 1954] gives an excellent description of much of the further work until 1949. We only remark that very few of the authors present their theorems in the finitary $\varepsilon - \eta$ format which LINDEBERG had adopted; important exceptions are LÉVY (cf. his books *Calcul des probabilités* (1925) as well as *Théorie de l'addition des variables aléatoires* (1937] both referred to in [GK 1954]) and DOOB [D 1953].

We now summarize briefly LINDEBERG's method and indicate HAUSDORFF's slight modification of it. Suppose U is the distribution function of $X_1 + \cdots + X_n$, the X_i's being independent, real-valued random variables with mean 0 and with variance of the sum 1; let Φ be the distribution function of the standard normal law (with mean 0 and variance 1); then LINDEBERG establishes the following estimate:

$$\left| \int_{-\infty}^{\infty} f(x-t)\, dU(t) - \int_{-\infty}^{\infty} f(x-t)\, d\Phi(t) \right| < c\, k\, L(X_1, \ldots, X_n) \tag{B.8}$$

where L is a suitable functional associated with the X_i's (like L_1, L_2, L_3 above), c an absolute constant and k is a number which bounds $\|f^{(3)}\|$ ($\|g\|$ being the supremum norm of $g : \mathbb{R} \to \mathbb{R}$), f, f', f'' being also bounded (eventually by $k/24$, $k/24$, $k/12$). By choosing f suitably, the inequality (B.8) is converted to an inequality

$$|U(x) - \Phi(x)| < c\, k\, L + c' k^{-1/3}$$

where c, c' are absolute constants; by choosing k to be proportional to $L^{-3/4}$ (essentially minimizing the right hand side of the preceding inequality), one obtains

$$|U(x) - \Phi(x)| < \mu\, L^{1/4} \tag{B.9}$$

where μ is a constant which can be calculated from c, c'. The choice of f in (B.8) is what now-a-days is called the choice of a mollifier. HAUSDORFF's argument is a slight variation which is perhaps pedagogically more convenient; HAUSDORFF establishes (B.8) for any $C^{(3)}$-function f with $\|f^{(3)}\| \le k$ and then takes $f(x) = \gamma(x/l)$, $l > 0$, where γ is some fixed $C^{(3)}$-function (explicitly given) with $0 \le \gamma \le 1$, $\gamma(x) = 0$ if $x \le 0$, $\gamma(x) = 1$ if $x \ge 1$ with $\|\gamma^{(3)}\| \le m$ (where m can be calculated exactly). From (B.8) one gets

$$|U(x) - \Phi(x)| \le c\, m\, \frac{L}{l^3} \;+\; \sqrt{\frac{2}{\pi}}\, l$$

by following HAUSDORFF's simple argument (the same as LINDEBERG's) which again leads to (B.9) with some absolute constant μ which can, in principle, be exactly determined. Thus both HAUSDORFF's argument as well as that of LINDEBERG are essentially arguments based on the choice of mollifiers; (B.9) seems to be the best that can be achieved with such methods even if we chose C^∞-mollifiers. LINDEBERG's method has been fruitfully used for vector-valued random variables as well cf. [PS 2000] (BENTKUS, GÖTZE et al pp. 42–50); most text-books today use LÉVY's method of characteristic functions as in [GK 1954].

Estimates like (B.9) naturally raise questions about the optimal rates of convergence in the central limit theorem. We briefly describe the current state of affairs in the next section.

(iii) Already LYAPOUNOV had given a good rate of convergence in his work of 1900–1901; this was improved by CRAMÉR in several papers between 1923–1937 (see [C 1994] or [GK 1954] for references). Definitive results were obtained by BERRY (1941) and ESSEEN (1945) (cf. [GK 1954]) which we now describe in order to compare them with estimates like (B.9) obtained by LINDEBERG and HAUSDORFF. The BERRY-ESSEEN result can be stated as folows: let $X_1, \cdots, X_n$ be independent, real-valued random variables with

$$\mathbb{E}\, X_j = 0, \quad \mathbb{E}\, X_j^2 = \sigma_j^2, \quad \mathbb{E}\, |X_j|^3 < \infty, \quad 1 \le j \le n$$

and let $b_n^2 = \sigma_1^2 + \cdots + \sigma_n^2,\ b_n > 0$, and

$$L = L(n) = \frac{1}{b_n^3} \sum_{j=1}^{n} \mathbb{E}\, |X_j|^3 ;$$

put

$$U_n(x) = \mathbb{P}\left(\frac{X_1 + \cdots + X_n}{b_n} \le x \right), \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du, \quad x \in \mathbb{R};$$

then

$$\sup_x |U_n(x) - \Phi(x)| \le c\, L(n) \tag{B.10}$$

where c is an absolute constant.

If the X_j's above are identically distributed with $\sigma_j^2 = \sigma^2$ $(\sigma > 0)$ and $\mathbb{E}\, |X_j|^3 = \beta$ then $L(n) = \rho\, n^{-1/2},\ \rho = \beta/\sigma^3$ and (B.10) becomes

$$\sup_x |U_n(x) - \Phi(x)| \le c'\, \rho\, n^{-1/2} \tag{B.11}$$

where $c' \le c$ is again an absolute constant.

Much research has been devoted to the determination of c and c'; the best result seems to be due to PAUL VAN BEEK (1972) with

$$0.40974 \le c \le 0.7975;$$

note also the following simple fact:

$$\frac{1}{\sqrt{2\pi}} = 0.39\cdots \le c' \le c$$

where $1/\sqrt{2\pi}$ comes from the consideration of the special case of the symmetric binomial distribution; exact references can be found in [P 1975]. An estimate like (B.10) has also been given under LINDEBERG's general conditions; thus, let $g : \mathbb{R} \to \mathbb{R}$ be non-negative, even and non-decreasing for $x > 0$ with $x/g(x)$ non-decreasing for $x > 0$; if $X_1, \ldots, X_n$ are real-valued, independent random variables with

$$\mathbb{E}\, X_j = 0, \quad \mathbb{E}\, X_j^2 = \sigma_j^2 < \infty, \quad \mathbb{E}\,\{X_j^2\, g(X_j)\} < \infty, \quad 1 \le j \le n$$

$$b_n^2 = \sigma_1^2 + \cdots + \sigma_n^2 > 0$$

then (with notations as before)

$$\sup_x |U_n(x) - \Phi(x)| \le \frac{A}{b_n^2\, g(b_n)} \sum_{j=1}^{n} \mathbb{E}\,\{X_j^2\, g(X_j)\} \tag{B.12}$$

where A is an absolute constant (depending only on the choice of g); cf. PETROV in [PS 2000], p. 5. If

$$g(x) = \begin{cases} |x| & \text{if} \quad |x| < 1 \\ 1 & \text{if} \quad |x| \ge 1 \end{cases}$$

then $x \mapsto x^2\, g(x)$, $x \in \mathbb{R}$, gives precisely LINDEBERG's function s (defined above); if $b_n = 1$ then (B.12) becomes

$$\sup_x |U_n(x) - \Phi(x)| \le A\, L_2(X_1, \ldots, X_n)$$

in our notation (given above) which is better then (B.4); the best value of A does not seem to have been studied.

Note that in HAUSDORFF's Notes, instead of (B.10) one obtains the weaker estimate $\text{const} \cdot L^{1/4}(n)$ and instead of (B.11) (the identically distributed case) one gets the weaker estimate $\text{const} \cdot n^{-1/8}$.

References

[BN 1977] BARONE, J.; NOVIKOFF, A.: *A history of the axiomatic formulation of probability from Borel to Kolmogorov*. Part I. Archive for History of Exact Sciences **18** (1977/78), 123–190.

[C 1994] CRAMÉR, H.: *Collected works*, Vols. I, II. Springer-Verlag, Berlin 1994.

[**CT 1988**] Chow, Y. S.; Teicher, H.: *Probability theory* (2nd ed.). Springer-Verlag, Berlin 1988.

[**D 1953**] Doob, J. L.: *Stochastic processes*. Wiley, London 1953.

[**F 1968**] Feller, W.: *An introduction to probability theory and its applications*, Vol. I (3rd ed.). Wiley, New York 1968.

[**Fr 1950**] Fréchet, M.: *Recherches théoretiques modernes sur le calcul des probabilités, premier livre* (2nd ed.). Gauthier-Villars, Paris 1950.

[**GK 1954**] Gnedenko, G. K.; Kolmogorov, A. N.: *Limit distributions for sums of independent random variables* (translated from the Russian by K. L. Chung). Addison-Wesley, Cambridge (Mass.) 1954.

[**H 2001**] Hausdorff, F.: *Gesammelte Werke*, Band IV. Springer-Verlag, Berlin 2001.

[**H 2002**] Hausdorff, F.: *Gesammelte Werke*, Band II. Springer-Verlag, Berlin 2002.

[**Ha 1998**] Hald, A.: *A history of mathematical statistics from 1750 to 1930*. Wiley, New York 1998.

[**KS 1973**] Kendall, M. G.; Stuart, A.: *The advanced theory of statistics*, Vol. 2 (3rd ed.). Hafner, New York 1973.

[**KaS 1989**] Kashin, B. S.; Saakyan, A. A.: *Orthogonal series*. American Math. Soc., Providence 1989.

[**M 1964**] Marcinkiewicz, J.: *Collected papers*. PWN, Warsaw 1964.

[**P 1975**] Petrov, V. V.: *Sums of independent random variables*. Springer-Verlag, Berlin 1975.

[**P 1995**] Petrov, V. V.: *Limit theorems of probability theory*. Oxford University Press, Oxford 1995.

[**PS 2000**] Prokhorov, Yu. V.; Statulevičius (editors): *Limit theorems of probability theory*. Springer-Verlag, Berlin 2000.

[**R 1988**] Riesz, M.: *Collected papers*. Springer-Verlag, Berlin 1988.

[**Ra 1974**] Rademacher, H.: *Collected papers*, Vol. I. MIT Press, Cambridge (Mass.) 1974.

[**S 1974**] Stout, W. F.: *Almost sure convergence*. Academic Press, New York 1974.

[**U 1937**] Uspensky, J. V.: *Introduction to mathematical probability*. Mc Graw-Hill, New York 1937.