BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 50, Number 3, July 2013, Pages 373–390
S 0273-0979(2013)01411-3
Article electronically published on March 28, 2013

TERCENTENNIAL ANNIVERSARY OF BERNOULLI'S LAW OF LARGE NUMBERS

MANFRED DENKER

0

The importance and value of Jacob Bernoulli's work was eloquently stated by Andreĭ Andreyevich Markov during a speech presented to the Russian Academy of Science[1] on December 1, 1913. Marking the bicentennial anniversary of the Law of Large Numbers, Markov's words remain pertinent one hundred years later:

> In concluding this speech, I return to Jacob Bernoulli. His biographers recall that, following the example of Archimedes he requested that on his tombstone the logarithmic spiral be inscribed with the epitaph *Eadem mutata resurgo*. This inscription refers, of course, to properties of the curve that he had found. But it also has a second meaning. It also expresses Bernoulli's hope for resurrection and eternal life.
>
> We can say that this hope is being realized. More than two hundred years have passed since Bernoulli's death but he lives and will live in his theorem.

Indeed, the ideas contained in Bernoulli's *Ars Conjectandi* have impacted many mathematicians since its posthumous publication in 1713. The twentieth century, in particular, has seen numerous advances in probability that can in some way be traced back to Bernoulli. It is impossible to survey the scope of Bernoulli's influence in the last one hundred years, let alone the preceding two hundred. It is perhaps more instructive to highlight a few beautiful results and avenues of research that demonstrate the lasting effect of his work.

1

Late seventeenth and early eighteenth century mathematics had not seen sufficient development for understanding and expressing laws of chance. The treatise of Huygens,[2] which considers probability as a quotient of favorable cases by all possible cases, is a vivid testimony of such early attempts to clarify the notion of probability. In those days, and it even persists nowadays, there was a dispute among different disciplines on the meaning of probability, likelihood, and chance.

2010 *Mathematics Subject Classification.* Primary 60-03; Secondary 01-01, 60F05, 62-03.
Key words and phrases. History of probability, law of large numbers, Bernoulli's theorem.
The author would like to thank Brian Nowakowski and an anonymous referee for valuable comments which led to several improvements of the text.

[1]See [38], p. 163.

[2]This is Part One of the *Ars Conjectandi* with comments and improvements by Bernoulli.

©2013 American Mathematical Society
Reverts to public domain 28 years from publication

For instance, it was discussed how legal probabilities are connected to the outcomes of rolling dice, the latter resulting in equally likely, elementary events. "There was an old legal saying, known to Leibniz, *reasons are not to be counted, but weighted*."[3] It took 200 years to get this straight, axiomatically as a number between zero and one, and defined for *events*, alternatively (but mathematically not that rigorous) as a stabilizing property of relative frequencies (von Mises approach [47]). The foundation of an axiomatic treatment of probability took more than 30 years and culminated in Kolmogorov's treatise in [30], as Kolmogorov writes himself in the introduction:

> Der diesen allgemeinen Gesichtspunkten entsprechende Aufbau der Wahrscheinlichkeitsrechnung war in den betreffenden mathematischen Kreisen seit einiger Zeit geläufig; es fehlte jedoch eine vollständige von überflüssigen Komplikationen freie Darstellung des ganzen Systems (es befindet sich allerdings ein Buch von Fréchet in Vorbereitung).[4]

When Bernoulli's[5] *Ars Conjectandi* was first published in 1713, eight years after his death, "the message of *Ars Conjectandi* was not fully absorbed at the time of its publication."[6] Today we have a fairly good understanding of the material. So it is merely a question of awareness when connecting present knowledge with its origins in Bernoulli's work. The book by Hall[7] gives an excellent historical account of the *Ars Conjectandi* with a detailed list of related literature and early impact of the work.

The weak law of large numbers, as it is called today, is a central part of the *Ars Conjectandi* and a basic theorem in contemporary probability. It follows (once the step from a philosophical to an analytical definition of probability is made) easily from Chebychev's inequality. An easy analysis of variance then leads directly to the strong law of large numbers (see Etemadi [21] for a complete workout of this old idea). At a first glance, one may think that weak laws are predecessors of strong laws. However, the motivations and approaches to prove these laws are completely different. A strong law avoids probability calculations as understood by Bernoulli to a large extent, as can be seen from the ergodic theorem or other strong laws. It merely requires the notion of null sets from measure theory and some maximal inequality; both ideas are not present in the *Ars Conjectandi*. There is no direct connection to probabilistic reasoning for the strong law as it is for the weak law. The connection to Bernoulli's original ideas may be seen in modern developments of almost sure theorems in spaces of measures, results that originated in the work of Brosamler, Schatte, and Fisher in the mid-1980s.[8] The idea of proof goes back to the ergodicity of the Brownian flow (cf. Brosamler [12], p. 566): The scaling

[3]Quoted from Franklin [23], p. 365.

[4]Kolmogorov in [30], Vorwort. The book by Fréchet is cited as *Recherches théoriques modernes,* fasc. 3 du tome I du Traité des probabilités par E. Borel et divers auteurs. Paris: Gauthiers-Villars. The text says that the main ideas of probability were common knowledge among specialists, only a complete, simplified version had been missing.

[5]Jakob (Jacques) Bernoulli *12/27/1654 (Basel) to †08/16/1705 (Basel).

[6]Shafer [43] writes in the preface "Unfortunately, the message of *Ars Conjectandi* was not fully absorbed at the time of its publication, and it has been obscured by various intellectual fashions during the past 300 years."

[7][27], Chapter 15.

[8]See [33] for the final result for second moment variables, which we follow here.

flow $(f,t) \mapsto T_t f$, $T_t f(s) = \mathrm{e}^{-t/2} f(s\mathrm{e}^t)$ $(t \in \mathbb{R}, s \geq 0)$ on the space $C(\mathbb{R}_+, \mathbb{R})$ of continuous real valued functions on $\mathbb{R}_+$ is ergodic with respect to the distribution of the Brownian motion (on $C(\mathbb{R}_+, \mathbb{R})$). An application of the ergodic theorem leads to the almost sure central limit theorem for normally distributed random variables.

The almost sure theorem (in its final form of Lacey and Philipp) states that for a sequence of independent, identically distributed, and square integrable random variables X_n $(n \in \mathbb{N})$ there is a set Ω of full probability such that for each $\omega \in \Omega$ and $t \in \mathbb{R}$

$$\lim_{N\to\infty} \frac{1}{\ln N} \sum_{n=1}^{N} \frac{1}{n} \mathbb{I}_{\{\eta\in\Omega : S_n(\eta) - na \leq t\sqrt{n}\sigma\}}(\omega) = \Phi(t), \tag{1.1}$$

where Φ is the distribution function $\Phi(t) = \int_{-\infty}^{t} \varphi(u)du$ of the standard normal probability measure with density $\varphi(u) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{u^2/2}$. Here, and in the sequel, $\mathbb{I}_A$ denotes the indicator function of the set A and S_n stands for the partial sum of the variables $X_1, \ldots, X_n$: $S_n(\omega) := \sum_{i=1}^{n} X_i(\omega)$ with expectation $a = E(X_1)$ and variance $\sigma^2 = E(X_1^2) - a^2$.

The following excerpt, the original Latin text, from Bernoulli's *Ars Conjectandi* contains the law of large numbers. We shall see below how the above theorem is related to Bernoulli's ideas. In Part Four of the *Ars Conjectandi* the main proposition reads as follows:[9]

> *Propos. Princip. Sequitur tandem Propositio ipsa, cujus gratia hæc omnia dicta sunt, sed cujus nunc demonstrationem sola Lemmatum præmissorum applicatio ad præsens institutum absolvet. Ut circumlocutionis tædium vitem, vocabo casus illos, quibus eventus quidam contingere potest, fæcundos seu fertiles; & steriles illos, quibus idem eventus potest non contingere: nec non experimenta fæcunda sive fertilia illa, quibus aliquis casuum fertilium evenire deprehenditur; & infæcunda sive sterilia, quibus sterilium aliquis contingere observatur. Sit igitur numerus casuum fertilium ad numerum sterilium vel præcisè vel proximè in ratione* $\frac{r}{s}$*, adeoque ad numerum omnium in ratione* $\frac{r}{r+s}$ *seu* $\frac{r}{t}$*, quam rationem terminent limites* $\frac{r+1}{t}$ *&* $\frac{r-1}{t}$*. Ostendendum est, tot posse capi experimenta, ut datis quotlibet (puta c) vicibus verisimilius evadat, numerum fertilium observationum intra hos limites quàm extra casurum esse, h.e. numerum fertilium ad numerum omnium observationum rationem habiturum nec majorem quàm* $\frac{r+1}{t}$*, nec minorem quàm* $\frac{r-1}{t}$*.*

The translation of the last part is taken from [7]:[10]

> *Let the number of fertile cases be to the number of sterile cases precisely or approximately as r to s; or to the number of all the cases as r to r+s, or as r to t so that this ratio is contained between the limits* $(r+1)/t$ *and* $(r-1)/t$*. It is required to show that it is possible to take such a number of experiments that it will be in any*

[9] [4], p. 236.

[10] Page 28. Other translations are cited in the bibliography.

> *number of times (for example, in c times) more likely*[11] *that the number of fertile observations will occur between these limits rather than beyond them, that is, that the ratio of the number of fertile observations to the number of all of them will be not greater than* $(r+1)/t$ *and not less than* $(r-1)/t$.[12]

The message told is fairly simple: Calculate the probability of the event that a partial sum deviates from its mean by *at most* some fixed amount. Bernoulli solved this problem in a simple case, at his time, however, it was not as simple as it looks today.

Of course, Bernoulli's weak law of large numbers has its trace in the development of probability theory. In the pre-Kolmogorov era, the nineteenth and beginning of the twentieth century, the law of large numbers was considered as a core result of probability theory, and an account of the law as of 1913 was given by Chuprov,[13] the paper being translated and reprinted in [38]. Kolmogorov's result in 1928–29 ([31]) gives a necessary and sufficient condition that a sequence of independent, identically distributed random variables with law μ obeys the weak law with some (non-random) centering constants (replacing na) if and only if $\lim_{n\to\infty} n\mu(\{x : |x| > n\}) = 0$.[14]

At this point, and as a sort of application to other areas in mathematics, let us mention that the weak law characterizes B-convex Banach spaces, which are defined by Beck in [2]. A Banach space E is called B-convex if for some $k \in \mathbb{N}$ and some $\epsilon > 0$ the inequality

$$\inf_{\xi_i=\pm 1} \|\xi_1 x_1 + \cdots + \xi_k x_k\| \le k(1-\epsilon)$$

holds for any choice of elements $x_1, \ldots, x_k \in E$ of norm 1.[15] Beck showed that this convexity property of a Banach space is equivalent to the strong law of large numbers to hold. Later, Marcus and Woyczynski ([37]) added a third equivalent condition, Bernoulli's weak law of large numbers for a sequence X_n $(n \in \mathbb{N})$ of independent, identically distributed, and E-valued symmetric random vectors. This is Theorem 0.1 in [37] observing that 1-stable Banach spaces are exactly those which are B-convex[16] and that weak convergence to the point mass in 0 is the same as the weak law of large numbers.

Although the work of Bernoulli was largely neglected at its time of appearance, it was picked up by de Moivre who in [17] proved a refinement, now known as the celebrated central limit theorem.[17] In fact, de Moivre proved a local limit theorem

---

[11] It has been pointed out by Sylla ([6], p. 121) that the correct translation of *verisimilius* is *more likely*. Some translations use the term "more probable"; see page X in [6]. We followed the translation by Sheynin since it is closer to the original, though not in the best English.

[12] See also [27], p. 259.

[13] *Statisticheskiĭ Vestnik* (1914), 1–21.

[14] The literature on weak laws is quite huge, and spans the area when the conditions in the last theorems are relaxed: non-identical distribution, arbitrary norming constants (e.g. Feller [22]), Banach space valued random variables, ...

[15] The infimum is taken over all choices of positive or negative signs for the x_i.

[16] G. Pisier, *Sur les espaces qui ne contiennent pas de* ℓ_n^1 *uniformément*, Comptes Rendus Acad. Sci. Paris **277** (1973), 991–994.

[17] The terminology is due to G. Pólya, *Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem*, Math. Z. **8** (1920), 171–181. This is quoted by H. Cramér in *Mathematical Probability and Statistical Inference. Some personal recollection from an important phase of scientific development*, Intern. Statist. Review **49** (1981), p. 311.

which says that the probability of the deviation from the mean by *exactly* some fixed amount $d \in \mathbb{Z} - np$ is asymptotic to

$$\frac{1}{\sqrt{2\pi np(1-p)}} \mathrm{e}^{-\frac{d^2}{2np(1-p)}}.$$

In modern terminology it means that this quantity[18] is

$$P(S_n = np + d) = \binom{n}{np+d} p^{np+d}(1-p)^{n(1-p)-d} \quad (0 \le d + np \le n).$$

This theorem appears the first time in the 1738 edition of the Doctrine of Chances, but de Moivre worked on the problem much earlier[19] and found Stirling's formula at the same time as Stirling himself (published in *Miscellanea Analytica de Seriebus et Quadraturis*, Tonson & Watts, London, 1730). De Moivre also remarked that upon integrating one obtains the central limit theorem (in modern language).

Returning to the Lacey–Philipp result and taking the expectation in (1.1), we see that the logarithmic average of the probabilities of the deviation from the mean converges to the corresponding probability for independent and identically standard normally distributed random variables. Moreover, equation (1.1) also shows that the quantity in which Bernoulli was interested, $\sum_{d \le x\sqrt{n}} P(S_n = np + d)$, can be approximately calculated using the data provided by the sequence of observed variables. This means for any $x > 0$ and n sufficiently large,

$$P(|S_n - np| \le n\epsilon) \ge P(|S_n - np| \le x\sqrt{n}) \approx \frac{1}{\ln n} \sum_{k=1}^{n} \frac{1}{k} \mathbb{I}_{\{|S_k - kp| < x\sqrt{k}\}}.$$

In particular, letting $x \to \infty$ recovers trivially the weak law of large numbers. This application is in the same line of reasoning as Efron's bootstrap method in statistics, but it avoids the resampling procedure. Simulations have shown that the quantile estimation based on (1.1) is a reasonable competitor to bootstrap.

2

At the time of Bernoulli (and even today) one should distinguish between the meaning of phrases such as *without reasonable doubt* in law and probabilistic terms describing uncertainty. Franklin writes on page 365 in [23]: "By 1700 law had served its purpose for the mathematical theory of probability. The service was never returned. Legal probability has continued to exist, and it is accepted in legal theory that such notions as *proof beyond reasonable doubt* involve probability." Disregarding such difficulties with the notion, the probability of the deviation from the mean is well defined in Bernoulli's work, and his viewpoint became of much wider importance more than a century later. Statistics (from the Italian *statistica*, meaning statesman) had been introduced as a new branch of science in the middle of the eighteenth century by Gottfried Achenwall,[20] and the Gauss–Markov theorem in the early 1820s ([24]) is one of the first decision theoretic results in statistics. Both are milestones in the development of statistics. The probability of the deviation from the mean became a decision theoretic tool to estimate parameters of unknown distributions and to differentiate between distributions. Its calculation and estimation are therefore crucial for decision making under uncertainty. The emergence of

[18] In the sequel, $P(A)$ denotes the probability of an event A.

[19] See Hall [27], p. 469.

[20] See his book *Noticia politica vulgo statistica.*

statistical decision theory provided a much deeper insight into the relevance and usefulness of Bernoulli's original ideas. He certainly had such applications in mind, though not well articulated.

Having said this, the appreciation of Bernoulli's contribution to the development of probability and statistics becomes evident by the following illustration. Suppose, a mathematical model contains an unknown expectation of a distribution. A simple statistical task is to estimate the unknown up to a certain precision. First of all one has to quantify precision. There are two issues here: the upper bound ϵ for the deviation from the mean, and the probability q of the event that the deviation is bounded by ϵ. Usually, ϵ is small and q close to one. For example, if the unknown distributions P are given by a Bernoulli distribution with parameter $p \in [0,1]$,[21] then the parameter p is the expectation of the unknown probability distribution P and can be estimated by repeating an experiment n times returning 0 or 1 according to the probability P. By Bernoulli's result, for a given deviation error ϵ of n repetitions we have that[22]

$$q = P(\{\omega : \left|\overline{X}_n(\omega) - p\right| \leq \epsilon\}).$$

Now, this expression tends to 1 as n increases by the weak law of large numbers. It follows that p is estimated up to any accuracy with high probability (if only the observed sample is large enough). It also follows from this that a differentiation of two means can eventually be made, since $\overline{X}_n$ is close to the true mean with large probability. Of course, this is a rather imprecise statement for practical purposes, but it points to the direction where statistical decision theory becomes relevant. To date, the above reasoning is the heart of statistical decision theory. It is a major problem in statistics to calculate such probabilities as precisely as possible, either by mathematical reasoning or by simulation of many sequences of independent samples and calculating the probabilities of deviations by the Bernoulli method: the ratio of fertile to all cases. As explained before, the *Ars Conjectandi* had a significant influence on de Moivre's work on the central limit theorem, which in turn had a tremendous impact on the British probability school in the eighteenth century. This is nicely described in the article by Schneider [42] and, moreover, demonstrates the value of Bernoulli's work for statistics.

Statistical ideas are not found directly in the *Ars Conjectandi*, although there is a direct influence on the work of Nikolaus Bernoulli in [8].[23] The law of large numbers had little or no influence of Nikolaus's work, since he was interested in calculating the expected lifetime of the last living among b species over the time span 0 to t, and showed that this equals $\frac{tb}{b+1}$. The title of Nikolaus's dissertation has to be seen in the light of the difficulty "to apply the theory of games of chance outside its own domain".[24] At this point it should be mentioned that Montmort's *Jeux de Hazard* was a much more appreciated source for a theory of games of chance than Bernoulli's work. More importantly, notice that Chapter Four of the *Ars Conjectandi* gains its value as a first step to use "mathematics of probabilities in civil, moral and economic affairs".[25]

[21] That is $P(\{1\}) = 1 - P(\{0\}) = p \in [0,1]$.

[22] We use here the statistical notation $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ for a sequence $X_1, \ldots, X_n$ of observations.

[23] See page 113 in [27].

[24] This is attributed to Montmort in [43], prepublication version, p. 10.

[25] Sylla [6], p. viii.

The formulation of the *Principal Proposition*, as it is called in Bernoulli's work, differs from later formulations of the weak law of large numbers; for example, by Poisson in 1835.[26] An interesting discussion of Bernoulli's theorem can be found in Pearson's article [39]. He claims that all French and German publications on Bernoulli's theorem (after Cournot in 1840[27]) state the theorem incorrectly by claiming that accuracy increases with the square root of the number of observations. Of course, such statements are due to de Moivre.[28] Pearson then discusses at length the number of observations needed to obtain a certain coverage probability and found that by de Moivre's result only half or a third of the observations are needed, in contrast to an application of Bernoulli's law of large numbers. At the end of his article he concludes, "Bernoulli saw the importance of a certain problem; so did Ptolemy, but it would be rather absurd to call Kepler's or Newton's solution of planetary motion by Ptolemy's name! Yet an error of like magnitude seems to be made when De Moivre's method is discussed without reference to its author, under the heading of 'Bernoulli's Theorem' ", and a sentence later "The *Pars Quarta* of the *Ars Conjectandi* has not the importance which has often been attributed to it". This was written before the axiomatic treatment of probability, and entirely from a statistician's viewpoint. Evaluations of historical facts can be quite different, as can be seen from the quote of Markov's bicentennial speech.

3

We saw that the art of determining the distribution of a sum of independent, identically distributed random variables is the core of Jacob Bernoulli's *Ars Conjectandi* and still is a central part of probability today. From a purely analytic viewpoint, the distribution of such a sum of n random variables is the n-fold convolution of their common distribution. Thus there is no harm in replacing the real line by a topological group G and considering G-valued random elements.[29] This turns the problem immediately to Fourier analysis on groups. In the real case, let μ_n denote the probability distribution of $\frac{1}{\sqrt{n}}X_1$, then the central limit theorem states that the n-fold convolutions of μ_n converge to some normal distribution on the real line in the weak topology of measures if X_1 has zero expectation and finite second moment. Moreover, for fixed m, μ_{nm} converges as well to the m-fold convolution power of the distribution of $\frac{1}{\sqrt{m}}N$, where N has the distribution given by the limiting normal distribution. Probability measures μ being the m-fold convolution power of a rescaled μ for each $m \in \mathbb{N}$ are called stable and characterize all possible limits of sums of independent, identically distributed random variables (see [41]). More precisely, let X be a random variable with distribution μ. If for each m there are numbers $b_m > 0$ and $a_m \in \mathbb{R}$ such that the m-fold convolution of the distribution of $b_m^{-1}(X - a_m)$ equals μ, then μ is called stable. All probability measures for which the above sequence of convolutions converges to a stable distribution μ form the domain of attraction of μ.

[26] In fact he says that the convergence holds towards the mean of the probabilities.

[27] He refers to Antoine Augustine Cournot, *Exposition de la théorie des chances et des probabilités*, Paris, 1843.

[28] The proofs go back to Laplace (*Théorie analytique des probabilités* (1812)) and de Moivre ([17], 2nd edition).

[29] The measurability here is defined through the Borel measurable subsets of G.

More generally, a probability measure μ of $\mathbb{R}$ is called infinitely divisible if for each $m \in \mathbb{N}$ there is a probability measure of which its m-fold convolution is μ. These infinitely divisible measures are characterized by the famous Khinchin formula: Their characteristic function $\phi_\mu(t) = E\left(e^{itX}\right)$ is given by

$$\phi_\mu(t) = \exp\left\{i\gamma t + \int_{-\infty}^{\infty}\left(\mathrm{e}^{itx} - 1 - \frac{itx}{1+x^2}\right)\frac{1+x^2}{x^2}\Gamma(dx)\right\},$$

where $\gamma \in \mathbb{R}$ is a real constant and Γ is a non-decreasing bounded function (so induces a finite measure).[30] For a stable distribution, the characteristic function is

$$\phi_\mu(t) = \exp\left\{i\gamma t - c|t|^\alpha\left(1 + i\beta\,\mathrm{sign}(t)\,\omega(t,\alpha)\right)\right\}$$

with constants $c \geq 0$, $0 < \alpha \leq 2$, $-1 \leq \beta \leq 1$, and $\gamma \in \mathbb{R}$, where $\omega(t,\alpha) = \tan\frac{\pi\alpha}{2}$ for $\alpha \neq 1$ and $\omega(t,1) = \frac{2}{\pi}\log|t|$.

Only little can be said about a complete and exact analytic formula of the distribution of partial sums, excluding of course special examples, such as stable distributions through their characteristic function $\phi_\mu(t) = \int e^{itx}\mu(dx)$ or Bernoulli distribution where the distribution is explicitly given. The general Fourier theory ensures that the distribution function F of a probability measure μ on $\mathbb{R}$ is given by

$$F(b) - F(a) = \frac{1}{2\pi}\lim_{T\to\infty}\int_{-T}^{T}\frac{e^{-ita} - e^{-itb}}{it}\phi_\mu(t)dt.$$

Of course, this development in the direction of Fourier analysis bears ideas that are not contained in the *Ars Conjectandi*. However, it forms the backbone of modern theory extending Bernoulli's ideas. Going a step further, we end the section discussing a beautiful result on infinitely divisible distributions on Lie groups. It is clear that the property of being infinitely divisible can be formulated for any probability measure on a group. Let G be a locally compact group. A probability measure μ on G is called infinitely divisible if it has roots of any integer order.[31] It is called embeddable if there is a continuous convolution semigroup $\{\mu_t : t > 0\}$ such that $\mu = \mu_1$. There is a long-standing conjecture that in all connected Lie groups any infinitely divisible probability is embeddable. It is clear that any measure μ in a continuous convolution semigroup is infinitely divisible. A striking result, obtained by Dani and McCrudden in 1992 ([14], Theorem 4.7), states that any connected Lie group that is representable through a homomorphism $\rho : G \to GL(d,\mathbb{R})$ for some $d \geq 1$ with a discrete kernel has this property. The result has been extended recently by Dani, Guivarc'h, and Shah.[32] The general case is still open.[33]

4

In order to discuss the modes of approximations of probabilities involving partial sums of independent, identically distributed random variables, we begin with convergence of densities. This is a local theory and provides approximations of

[30] The integrand is $-t^2/2$ at $x = 0$.

[31] For every $n \geq 1$, μ is the n-fold convolution of a probability measure μ_n so that the latter probability is then called a convolution root of order n.

[32] Math. Z. **272** (2012), 361–379.

[33] Dani and McCrudden [15] is an excellent recent survey on this topic. As H. Heyer writes in a review (MathReviews MR2213481) for McCrudden's 2006 survey, "This is an impressive account of recent progress on the embedding problem for probability measures on locally compact groups, a problem of considerable depth, still not solved in full generality."

integrals if the convergence is uniform. De Moivre's theorem is a good example for such a procedure to approximate the probability of deviations from the mean. According to the dominating measure for the density, one distinguishes two cases: the absolutely continuous case when the dominating measure is Lebesgue and the lattice case when the dominating measure is the counting measure on some lattice.

In the discrete case, de Moivre's result has a natural formulation in full generality: Consider a sequence of independent, identically distributed random variables X, X_n $(n \geq 1)$ with finite variance $\sigma^2 > 0$ and a lattice distribution; that is, there exist $c \in \mathbb{R}$ and $h > 0$ such that $P(X \in c + h\mathbb{Z}) = 1$. Gnedenko's extension ([25]) of de Moivre's result is that

$$\lim_{n\to\infty} \sup_{N\in\mathbb{Z}} \left| \frac{\sigma\sqrt{n}}{h} P(S_n = nc + Nh) - \varphi\left(\frac{nc + Nh - nE(X)}{\sigma\sqrt{n}}\right)\right| = 0$$

holds if and only if the span h is maximal. The rate of convergence in this result was also investigated first by Esseen [20] and later by Petrov [40], the latter showing that

$$(1 + |x|^k)\left(\sqrt{\sigma^2 n} P(S_n = nc + Nh) - \varphi(x) - \sum_{v=1}^{k-2} \frac{q_\nu(x)}{n^{\nu/2}}\right) = o\left(n^{-(k-2)/2}\right)$$

uniformly in x, where $x = \frac{N-n\mu}{\sigma\sqrt{n}}$ and

$$q_\nu(x) = \varphi(x) \sum_{(k_1,\ldots,k_\nu)} H_{\nu+2s}(x) \prod_{m=1}^{\nu} \frac{1}{k_m!}\left(\frac{\gamma_{m+2}}{(m+2)!\sigma^{m+2}}\right)^{k_m},$$

where the sum extends over all $(k_1, \ldots, k_\nu) \in \mathbb{Z}_+^\nu$ with $k_1 + 2k_2 + \cdots + \nu k_\nu = \nu$, $s = k_1 + k_2 + \cdots + k_\nu$, H_l the lth Chebychev–Hermite polynomial $(l \geq 0)$ and γ_l the cumulant of order $l \geq 1$ of the distribution.[34]

In 1964, Shepp ([44]) proposed a slightly different type of local limit theorems for integer valued random variables. In this framework, Bernoulli's approach and de Moivre's theorem read as

$$\lim_{n\to\infty} \sqrt{n}\sigma P(S_n = k_n + I) = |I|\varphi(x),$$

as $(k_n - nE(X_1))/\sqrt{n\sigma^2}$ tends to x, where I is a bounded interval and $|I|$ denotes the counting measure of I. In fact, Shepp's theorem is stated only in the case $k_n = nE(X_1) = 0$. In general, a sequence X_n $(n \geq 1)$ of independent, identically distributed random variables satisfies a local limit law in Shepp's sense if there are constants $a_n \in \mathbb{R}$ and $b_n \to \infty$ such that for any bounded interval

$$\lim_n b_n P(S_n = k_n + I) = |I|g(x),$$

as $n \to \infty$ and $(k_n - a_n)/b_n$ converges to $x \in \mathbb{R}$, where $|I|$ denotes the Haar measure on $\mathbb{Z}$ if X_1 is lattice distributed on $\mathbb{Z}$, and the Haar measure on $\mathbb{R}$ otherwise. The function g is necessarily the density of a stable distribution.

In order to describe the nature of local theorems for the absolutely continuous case one has to assume that the partial sums have absolutely continuous distributions. Then, Gnedenko's theorem ([26]) holds for a sequence X_n of independent,

[34]The book [41] states these theorems on pp. 187 and 207.

identically distributed random variables with zero mean and finite, positive variance σ^2: Under the assumption that $\frac{1}{\sqrt{n}\sigma}S_n$ has a density f_n for each $n \geq 1$,

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |f_n(x) - \varphi(x)| = 0$$

if and only if some density f_n is bounded.

The rate of convergence in this local limit theorem has also been studied by many authors, Petrov's result is easily grasped as a generalization of the above rate theorem for the lattice case.

Two fairly recent developments seem to be in order at this point. A local limit theorem for Cauchy distributed random variables is easily established for independent and identically distributed random variables in the domain of attraction of a Cauchy law. Such results can also be proved in cases when independence is relaxed, such as in [1] for Gibbs–Markov dynamical systems aiming for convergence properties of the Poincaré series. For a Fuchsian group G of the second kind the Poincaré exponent is defined as that power γ for which

$$\sum_{g\in G} (1 - |g(x)|)^s$$

diverges for $s < \gamma$ and converges for $s > \gamma$. The $\mathbb{Z}$-extension of the three punctured sphere is then represented as $\mathbb{H}\backslash G$, where G is a group of divergence type which means that the above series diverges for $s \downarrow 1$. It is shown in [1] that the Poincaré series for G is proportional to $-\ln(s-1)$ as $s \downarrow 1$. This is so since the associated geodesic flow on the Riemann surface $\mathbb{H}\backslash G$ has Poincaré sections for which the return map is Gibbs–Markov and the return time satisfies the Cauchy local limit theorem. Again we see that Bernoulli's postulate to calculate deviations from the mean is an essential tool. This example also shows that local limit theorems are much stronger than central limit theorems and lead to new results as well in other mathematical disciplines.

The almost sure central limit theorem discussed in Section 1 immediately puts the question for an almost sure version in terms of a local limit theorem. This is called an almost sure local limit theorem and is formulated as the convergence of the logarithmic averages:

$$\lim_{N\to\infty} \frac{1}{\ln N} \sum_{n=1}^{N} \frac{\sigma}{\sqrt{n}} \mathbb{I}_{\{S_n = k_n + I\}} = \varphi(x)|I| \quad \text{a.s. as } \frac{k_n - na}{\sqrt{n\sigma^2}} \to x$$

for any bounded interval I. Such theorems are known: [18] contains the almost sure version of de Moivre's theorem; extensions and clarifications by Burmeister[35] and Weber[36] are more recent.

5

By far the most attention for estimating the unknown distribution of partial sums has been given to the medium range, which means to the approximation of probabilities of the form

$$P\left(\frac{1}{n}S_n - a \leq \frac{t}{\sqrt{n}}\right), \qquad t \in \mathbb{R}, \tag{5.1}$$

[35] J. Electr. Eng. **55** No. 12 (2004), 68–71.
[36] Stoch. Anal. Appl. **29** (2011), 779–798.

under appropriate assumptions on the distribution on X_1. This is just the extension of Bernoulli's idea by de Moivre and was solved in terms of the central limit theorem, currently the core of probability theory. The form of the theorem was extended over the years by Laplace, Lyapunov, Lindeberg, Lévy, and Feller to its present form. For independent and identically distributed random variables with finite variance $\sigma^2 > 0$, it states that the above probabilities converge, as $n \to \infty$, to $\Phi(t/\sigma)$. In fact, the result holds for distributions of X_1 which are in the domain of attraction of a normal distribution, which is a slightly larger class, the variance being replaced by a truncated variance. The most useful versions of a central limit theorem are those of Lindeberg–Lévy for arrays of independent random variables[37] and the martingale central limit theorem of Billingsley and Ibragimov.[38] The latter theorem states that for a stationary sequence $(X_k)_{k\geq 1}$ of martingale differences (i.e. $E(X_k|\sigma(X_1,\ldots,X_{k-1})) = 0$) with positive and finite variance σ^2, the distributions of the sequence $S_n/(\sqrt{n}\sigma)$ follow asymptotically the standard normal law.

Current research entails relaxing the independence condition in the classical central limit theorem. One example of doing this is to consider a measure preserving transformation T on some finite measure space (Ω, Σ, m) and the stationary process $X_n = F \circ T^n$ $(n \in \mathbb{N})$ for a measurable function $F : \Omega \to \mathbb{R}$. There are many central limit theorems for such processes, the good choices for F are dense in the $L_2(m)$-space for aperiodic transformations. This is mainly a topic in dynamical systems theory. On the other hand, there are purely probabilistic open questions like the Ibragimov conjecture: "*If a stationary process (strictly stationary in the sense of time series) X_n $(n \in \mathbb{Z})$ is uniformly mixing, then the central limit theorem holds under the assumption that the variance of $S_n = X_1 + \cdots + X_n$ diverges*". This problem is related to the coboundary representation of X_1 in the $L_2(P)$-space, for if it is such a coboundary, the variance stays bounded.[39]

Since the central limit theorem estimates $P(|S_n - na| \leq \epsilon n)$ even if ϵ tends to 0 at a rate $O(n^{-1/2})$, it is of much greater use in statistics than Bernoulli's original estimate (see Pearson's remark quoted earlier). In fact, the central limit theorem plays a central role in non-parametric statistics where only for large sample sizes type one errors are controllable. For Bernoulli random variables (as a special case of a more general approach), the central limit theorem also serves to improve error bounds for statistics by transforming data. This is known as the variance stabilizing method.

A natural question is to estimate the speed of convergence in the central limit theorem. A famous result in this vein is the Berry–Esseen theorem ([9], [20]), which states that for X_1 with vanishing first and finite third moment the approximation by the normal distribution is

$$\sup_{x\in\mathbb{R}} \left|P(S_n \leq x\sqrt{n}\sigma) - \Phi(x)\right| \leq (.7975)\frac{E(|X_1|^3)}{\sigma^3} n^{-\frac{1}{2}}.$$

[37]See the references in [10].

[38]This was proved independently in [11] and [28].

[39]The process is called uniformly mixing if $P(B|A) - P(A) = O(\eta_n)$ uniformly in A and B for some sequence $\eta_n \to 0$, where A is measurable with respect to the σ-algebra generated by all X_k with $k \leq 0$ and B is measurable with respect to the σ-algebra generated by all X_k with $k > n$.

The least lower bound of the constant is still unknown.[40] The first result on such bounds were obtained by Lyapunov ([35]) who needed an additional logarithmic factor in the estimate. This was removed by Cramér under some additional assumptions which then were shown to reduce to the third moment. It is clear that the rate in the Berry-Esseen theorem is not improvable since the sum of Bernoulli random variables $(X_n)_{n\geq 1}$ with $P(X_1 = 1) = P(X_1 = -1) = 0.5$ satisfies $P(S_n = 0) = O(\frac{1}{\sqrt{n}})$, again by Bernoulli's calculations in the *Ars Conjectandi*. As a side remark, notice that Bhattacharya obtained such approximation results for convolutions of probability measures on compact groups which converge to Haar measure.[41]

Once the rate of convergence in the approximation of the probabilities of deviation from the mean is established, the natural question is for the analytic expansion of the probability in (5.1). This is a mathematically highly non-trivial problem. The books by Bhattacharya and Ranga Rao ([10]) and Petrov ([41]) are good references for it. Such results are somewhat similar to the expansion formulas mentioned in the section on local limit theorems. The ingredients for proofs are Fourier analysis, the expansion of the characteristic function, and some intriguing truncation techniques.

There is a long history of the extension of the foregoing results to multidimensional random vectors which are assumed to be independent and identically distributed. In this setting, intervals are replaced by convex sets. A highlight in this regard is the work [3] of Bentkus and Götze on the rate of convergence of quadratic forms and its application to lattice point problems of quadratic forms in ellipsoids. We quote only the reference [3] here and review some of the remarkable results. Let Q be an irrational[42] positive definite quadratic form in d dimensions, and let $E_s = \{x \in \mathbb{R}^d : Q(x) \leq s\}$. Then, in dimension ≥ 9, for a positive irrational quadratic form Q the number of lattice points in $\mathbb{Z}^d$ which lie in a shifted ellipsoid $E_s + c$ ($s \in \mathbb{R}_+,\ c \in \mathbb{R}^d$) deviates from the mean (the volume V_s of the ellipsoid E_s) by $o(V_s/s)$. Results of this type go back as far as Landau ([34]) who derived the order $V_s s^{-\frac{d}{1+d}}$ for $d \geq 1$, with refinements by Krätzel and Nowak ([32]) to the order $V_s s^{-1+\lambda}$.[43] For special results one has to name Walfisz, Landau, and Jarnik. As a consequence, for dimension $d \geq 9$, Bentkus and Götze settle affirmatively a conjecture by Davenport and Lewis that in dimension ≥ 5 the gaps of successive values in $Q(\mathbb{Z}^d)$ of an irrational quadratic form converge to zero as s tends to infinity. This is obviously not so if Q is rational. As a corollary, they also reprove Oppenheimer's conjecture in case $d \geq 9$ by their expansion of probabilities; the general case $d \geq 5$ had been settled earlier by Margulis ([36]) using methods from ergodic theory for group actions.

6

Large deviation theorems were obtained in 1929 by Khinchin ([29]) for binomial distributions, and later by Smirnov, Lévy, and Fréchet. Khinchin was interested in

[40] This may be no longer of such interest in the computer age. The constant appearing here is due to van Beek (see [10], p. 186).

[41] *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **23** (1972), 1–10.

[42] Irrational means that there is no $M \neq 0$ such that the matrix MQ has integer entries.

[43] $\lambda = 5/(6d+2)$ for $d \geq 8$ and $\lambda = 12/(14d+8)$ for $3 \leq d \leq 7$.

the deviation from the mean in the form of

$$P\left(\frac{1}{\sqrt{np(1-p)}}(S_n - p) \le x(n)\right),$$

when $\lim_{n\to\infty} n^{-1/2}x(n) = 0$. Until the beginning of the 1970s (see [41]) this was called a large deviation problem; today the terminology has changed and it is considered as moderate deviation. Khinchin derived the estimate

$$P(S_n \ge x_n\sqrt{n}) = \exp[-\frac{1}{2}x_n^2(1+o(1))],$$

where $x_n \uparrow \infty$ with a certain rate of growth. The first celebrated Cramér theorem (in its stronger form of Petrov [41], p. 218.) says that for $x = o(\sqrt{n})$

$$P(S_n > x\sigma\sqrt{n}) = (1-\Phi(x))\exp\left\{x^3 n^{-1/2}\lambda(xn^{-1/2})\right\}\left(1+O\left(\frac{x+1}{\sqrt{n}}\right)\right),$$

where λ is a power series (involving the cumulants of X_1) converging in a neighborhood of 0.[44] The assumption here is that Cramér's condition holds; that is, X_1 has a finite Laplace transform in a neighborhood of 0. These results sketch the scope for moderate deviations. Under the same assumption, Cramér's second celebrated theorem in 1938 ([13], p. 19) proves a result which has the form of a large deviation result as it is accepted today. He showed that

$$\lim_{n\to\infty}\frac{1}{n}\log P(S_n \ge \alpha n) = -I(\alpha)$$

for $\alpha > E(X_1)$, where I is the information function (see the discussion below).

Large deviation is directly concerned with the deviation probabilities in Bernoulli's sense. It aims to estimate probabilities of the form

$$P(S_n \ge \alpha n)$$

when α is larger than the mean (and the analogous expressions when α is smaller than the mean). Large deviation theory needs much stronger assumptions than moderate deviation. The standard assumption here is Cramér's condition.

The widely adopted terminology for a large deviation result is based on Varadhan's seminal work in 1966 ([45]) which has all the essential definitions and properties needed to formulate the large deviation principle. Let S be a complete metric space equipped with the Borel field. A sequence of Borel probability measures $\{\mu_n : n \ge 1\}$ on S is said to have the large deviation property if there exists a sequence $a_n \in \mathbb{R}_+$ tending to infinity and a lower semicontinuous function $I : S \to [0,\infty]$ with compact level sets such that for each closed set $C \subset S$ and open set $O \subset S$

$$\limsup_{n\to\infty}\frac{1}{a_n}\log\mu_n(C) \le -\inf_{s\in C} I(s),$$
$$\liminf_{n\to\infty}\frac{1}{a_n}\log\mu_n(O) \ge -\inf_{s\in O} I(s).$$

The function on the right-hand side is called the information function I. It is a convex function.

[44] Cramér's theorem is in [13], p. 12.

For the distributions μ_n of partial sums S_n the free energy function is defined as

$$c(t) = \lim_{n\to\infty} \frac{1}{n} \log \int e^{tx} \mu_n(dx)$$

for all $t \in U$, the domain of the Laplace transform. It is equal to $\log Ee^{tX_1}$ for independent, identically distributed random variables satisfying Cramér's condition. It plays a crucial role in deriving the large deviation principles for these random variables, since by Chebychev's inequality

$$\limsup_{n\to\infty} \frac{1}{n} \log P(S_n \geq \alpha) \leq \inf_{t\in U} c(t) - \alpha t,$$

the converse inequality being a bit more involved.

Varadhan's result from 1966 is the backbone of the theory today. It calculates certain integrals under the assumption that the large deviation property holds. The theory has been further developed by Donsker, Varadhan, and others. Details are to be found in [19] (among other excellent monographs).

Large deviation result are not so far apart from Bernoulli's original goals. Consider the case of fair coin tossing.[45] Define the function $I(z) = z\log(2z) + (1-z)\log(2(1-z))$ for $0 \leq z \leq 1$ and $I(z) = \infty$ for all other $z \in \mathbb{R}$. A simple calculation using binomial coefficients (as Bernoulli did) leads to

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\left|S_n - \frac{n}{2}\right| \geq n\epsilon\right) = -\inf_{z\notin(\frac{1}{2}-\epsilon,\frac{1}{2}+\epsilon)} I(z). \tag{6.1}$$

One observes that $1 - I(z)$ is the entropy of the Bernoulli distribution with parameter z as long as $0 \leq z \leq 1$. This is not a coincidence, it holds in general and is connected to the thermodynamical formalism for shift spaces, as developed by Ruelle and others. Let us consider the space of all one sided infinite sequences $(x_n)_{n\in\mathbb{N}}$ of zeros and ones. This space has a natural product topology and a natural map which is the left-shift (the first coordinate is deleted under the shift map). The pressure $P(\psi)$ of a continuous function ψ defined on the space of infinite sequences is the supremum over all expressions $H(\mu) + \int \psi d\mu$, where μ runs through all shift-invariant probability measures, and where $H(\mu)$ denotes the Kolmogorov–Sinai entropy of the measure. For the function $\psi((x_n)_{n\in\mathbb{N}}) = -\log(p)\mathbb{I}_{\{1\}}(x_1) - \log(1-p)\mathbb{I}_{\{0\}}(x_1)$, the supremum is attained by the Bernoulli measure on the shift space with parameter p and hence the projection maps X_n (mapping a point in the shift space to its nth coordinate) form a Bernoulli process with parameter p. Now let $f((x_n)_{n\in\mathbb{N}}) = \mathbb{I}_{\{1\}}(x_1)$, and define $E(t) = P(tf + \psi) - P(\psi)$. Then E is differentiable and

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n \geq nE'(t)) = -tE'(t) + E(t)$$

for $t > 0$. For $p = \frac{1}{2}$ one easily reduces this to (6.1).

What is discussed up to this point is often called the level 1 large deviation property. There are three levels, level 2 being the same type of problem replacing the topology of the real numbers by the weak topology on the space of probability measures.[46] Instead of looking at S_n, consider the empirical measures defined by

[45] We follow here R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*, Springer, 1985, pp. 11–13.

[46] Level 3 will not be discussed here.

$X_1, \ldots, X_n$:

$$\mu_n(\omega) = \frac{1}{n} \sum_{k=1}^{n} \delta_{X_k(\omega)},$$

where δ_x denotes the probability giving probability one to an event if and only if x belongs to this event. Thus the large deviation property for Bernoulli random variables has a formulation as random measures and their large deviation property.

Going a step further, there are quite a number of papers dealing with some finer asymptotic analysis. As an example, Deheuvels, Devroye, and Lynch ([16]) used Petrov's theorem[47] on the rate of convergence in the large deviation principle to obtain a rate theorem for the Erdős–Rényi law: Let $l_n = l_n(\alpha)$ denote the integer part of $\frac{\log(n)}{I(\alpha)}$. Then

$$\limsup_{n\to\infty} \max_{0\le m\le n-l_n} \frac{S_{m+l_n} - S_m - l_n\alpha}{\log(l_n)} = \frac{1}{2I(\alpha)} \qquad \text{a.s.}$$

and similarly the limit inferior is $= -\frac{1}{2I(\alpha)}$ a.s. Connecting to Bernoulli's result, we see that the maximal number of successes in l_n consecutive subtrials in an overall sequence of n Bernoulli trials is asymptotically proportional to $\log(l_n)$. This counting was considered to be essential by Bernoulli.

7

The terms probability and expectation can be traced back to Cardano, Fermat, and Pascal; apparently not to the science in the ancient world. The logical treatment of these notions was first developed by C. Huygens[48] in 1656, who connected his theory to games of chance. Bernoulli goes one step further in the *Ars Conjectandi*. In *Pars Quarta* he clearly connects probability to civil, moral, and economic matters, expressed in its title, *Application of the preceding theory to civil, moral and economic relations*. This is expressed in a few words on page 213 in [4] "Conjecturing some matter means measuring its probability".[49] He then discusses at length the application to various daily life problems, and, as a conclusion, states that it is necessary to either determine the probabilities by counting (a priori determination) or to observe similar cases and deduce the unknown probability (a posteriori determination) from this. The former leads to the law of large numbers, by asking the question whether this is in principle possible or not: on page 225 in [4] Jacob writes that still another problem has to be considered of which probably no one has even thought about, whether an increase of observations would result in a better approach of the unknown probability, or whether it does not converge at all or converges to a wrong limit.[50]

In a letter to Leibniz on October 3, 1703, Jacob describes the content of the *Ars Conjectandi* mentioning that most of it has been finished, the connection to the applications (in the sense of the *Pars Quarta*) still being missing. He considered this missing part to be essential. Jacob Bernoulli had a vision, he laid the foundation

[47] Theory Probab. App. **10** (1986). This is in fact a sharpening of Khinchin's result for Bernoulli random variables.

[48] According to van der Waerden [46], p. IX.

[49] Conjicere rem aliquam est metiri illius probabilitatem. This also explains why he chose the title *Ars Conjectandi*.

[50] The Latin sentence begins "Ulterius aliquid hic contemplandum. . . ."

for its realization with the law of large numbers, and 300 years later we can see that his vision became one of the important applications of mathematics.

About the author

Manfred Denker is professor of mathematics at The Pennsylvania State University and adjunct professor at Case Western Reserve University. Before he taught at the Georg-August-Universität Göttingen in Germany for 34 years.

References

[1] Jon Aaronson and Manfred Denker, *The Poincaré series of* $\mathbf{C}\backslash\mathbf{Z}$, Ergodic Theory Dynam. Systems **19** (1999), no. 1, 1–20, DOI 10.1017/S0143385799126592. MR1676950 (2001b:37042)

[2] Anatole Beck, *A convexity condition in Banach spaces and the strong law of large numbers*, Proc. Amer. Math. Soc. **13** (1962), 329–334. MR0133857 (24 #A3681)

[3] V. Bentkus and F. Götze, *Lattice point problems and distribution of values of quadratic forms*, Ann. of Math. (2) **150** (1999), no. 3, 977–1027, DOI 10.2307/121060. MR1740988 (2001b:11087)

[4] J. Bernoulli, *Ars conjectandi: Opus posthumum: accedit tractatus de seriebus infinitis, et Epistola Gallicè Scripta de ludo pilae reticularis.* Published 1713 by Impensis Thurnisiorum, fratrum in Basileae. (Latin.)

[5] J. Bernoulli: Wahrscheinlichkeitsrechnung: Ars conjectandi. 1.,2., und 4. Theil (1913). Edited and translated by R. Haussner. Oswalds Klassiker der exakten Wissenschaften Band 107/108. Reprint. Verlag Harri Deutsch, 1999. (German.)

[6] Jacob Bernoulli, *The art of conjecturing*, Johns Hopkins University Press, Baltimore, MD, 2006. Together with "Letter to a friend on sets in court tennis"; Translated from the Latin and with an introduction and notes by Edith Dudley Sylla. MR2195221 (2006j:01006)

[7] J. Bernoulli, *On the law of large numbers.* Translated to English by Oscar Sheynin, Berlin 2005. http://www.sheynin.de/download/bernoulli.pdf.

[8] N. Bernoulli, *De usu artis conjectandi in jure.* Basel 1709. Reprint in: *Die Werke von Jakob Bernoulli,* Vol. 3, 287–326, Birkhäuser Basel 1975. (Latin.)

[9] Andrew C. Berry, *The accuracy of the Gaussian approximation to the sum of independent variates*, Trans. Amer. Math. Soc. **49** (1941), 122–136. MR0003498 (2,228i)

[10] R. N. Bhattacharya and R. Ranga Rao, *Normal approximation and asymptotic expansions*, Robert E. Krieger Publishing Co. Inc., Melbourne, FL, 1986. Reprint of the 1976 original. MR855460 (87k:60062)

[11] Patrick Billingsley, *The Lindeberg-Lévy theorem for martingales*, Proc. Amer. Math. Soc. **12** (1961), 788–792. MR0126871 (23 #A4165)

[12] Gunnar A. Brosamler, *An almost everywhere central limit theorem*, Math. Proc. Cambridge Philos. Soc. **104** (1988), no. 3, 561–574, DOI 10.1017/S0305004100065750. MR957261 (89i:60045)

[13] H. Cramér, *Sur un nouveau théorème-limite de la théorie des probabilités.* Actualités Scientifiques et Industrielles **736** (1938), 5–23. Colloque consacré à la théorie de probabilités. Vol. 3, Hermann, Paris. (French.)

[14] S. G. Dani and M. McCrudden, *Embeddability of infinitely divisible distributions on linear Lie groups*, Invent. Math. **110** (1992), no. 2, 237–261, DOI 10.1007/BF01231332. MR1185583 (94d:60010)

[15] S. G. Dani and M. McCrudden, *Convolution roots and embeddings of probability measures on Lie groups*, Adv. Math. **209** (2007), no. 1, 198–211, DOI 10.1016/j.aim.2006.05.002. MR2294221 (2008g:60016)

[16] Paul Deheuvels, Luc Devroye, and James Lynch, *Exact convergence rate in the limit theorems of Erdős–Rényi and Shepp*, Ann. Probab. **14** (1986), no. 1, 209–223. MR815966 (87d:60032)

[17] A. de Moivre, *The Doctrine of Chances: or, A Method of Calculating the Probability of Events in Play.* London 1718. Second edition, London 1738, third edition, London 1756.

[18] Manfred Denker and Susanne Koch, *Almost sure local limit theorems*, Statist. Neerlandica **56** (2002), no. 2, 143–151, DOI 10.1111/1467-9574.00189. Special issue: Frontier research in theoretical statistics, 2000 (Eindhoven). MR1916315 (2003g:60049)

[19] J.-D. Deuschel, D. W. Strook, *Large Deviations.* AMS Chelsea Publ., Amer. Math. Soc., Providence 1989.

[20] Carl-Gustav Esseen, *Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law*, Acta Math. **77** (1945), 1–125. MR0014626 (7,312a)

[21] N. Etemadi, *An elementary proof of the strong law of large numbers*, Z. Wahrsch. Verw. Gebiete **55** (1981), no. 1, 119–122, DOI 10.1007/BF01013465. MR606010 (82b:60027)

[22] W. Feller, *Über das Gesetz der großen Zahlen.* Acta Litt. Scient. Szeged **8** (1937), 191–201. (German.)

[23] James Franklin, *The science of conjecture*, Johns Hopkins University Press, Baltimore, MD, 2001. Evidence and probability before Pascal. MR1893301 (2003c:01002)

[24] C. F. Gauß, *Theoria combinationis observationum erroribus minimis obnoxiae.* Commentationes Societatis Regiae Scientiarum Gottingensis recentiores 5 (classis mathematicae) 1823. Part 1: February 15, 1821, Part 2: February 2, 1823. (Theory of combinations of observations which are subject to small errors.) Publ. H. Dieterich 1825. (Latin.)

[25] B. V. Gnedenko, *O lokal'noĭ predel'noĭ theoreme teorii veroyatnosteĭ.* Uspehi matemat. nauk **3** (1948), 187–194. (On the local limit theorem in the theory of probability.) (Russian)

[26] B. V. Gnedenko, *O lokal'noĭ predel'noĭ teoreme dlya odinakovo raspredelennyh nezavisimyh slagaemyh.* Wiss. Z. Humboldt–Univ. Berlin. Math.-Naturwiss. Reihe **3** (1953/4), 287–293. (On the local limit theorem for identically distributed independent terms.)

[27] Anders Hald, *A history of probability and statistics and their applications before 1750*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1990. A Wiley-Interscience Publication. MR1029276 (91c:01003)

[28] I. A. Ibragimov, *A central limit theorem for a class of dependent random variables*, Teor. Verojatnost. i Primenen. **8** (1963), 89–94 (Russian, with English summary). MR0151997 (27 #1978)

[29] A. Khintchine, *Über einen neuen Grenzwertsatz der Wahrscheinlichkeitsrechnung*, Math. Ann. **101** (1929), no. 1, 745–752, DOI 10.1007/BF01454873 (German). MR1512565

[30] A. Kolmogoroff, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer-Verlag, Berlin, 1977 (German). Reprint of the 1933 original. MR0494348 (58 #13242)

[31] A. Kolmogoroff, *Über die Summen durch den Zufall bestimmter unabhängiger Größen.* Math. Ann. **99** (1928), 309–319; Math. Ann. **102** (1929), 484–488.

[32] Ekkehard Krätzel and Werner Georg Nowak, *Lattice points in large convex bodies*, Monatsh. Math. **112** (1991), no. 1, 61–72, DOI 10.1007/BF01321717. MR1122105 (92i:11112)

[33] Michael T. Lacey and Walter Philipp, *A note on the almost sure central limit theorem*, Statist. Probab. Lett. **9** (1990), no. 3, 201–205, DOI 10.1016/0167-7152(90)90056-D. MR1045184 (91e:60100)

[34] E. Landau, *Zur analytischen Zahlentheorie der definiten quadratischen Formen (Über die Gitterpunkte in einem mehrdimensionalen Ellipsoid).* Sitzungsberichte Preuss. Akad. Wiss. **31** (1915), 458–476. (German.)

[35] A. Liapounoff, *Nouvelle forme du théorème sur la limite de probabilité.* Mémoires de l'Académie Impériale des Sciences de St.-Pétersbourg VIII° série, **12**(5) (1901), 1–24. (French.)

[36] G. A. Margulis, *Discrete subgroups and ergodic theory*, Number theory, trace formulas and discrete groups (Oslo, 1987), Academic Press, Boston, MA, 1989, pp. 377–398. MR993328 (90k:22013a)

[37] Michael B. Marcus and Wojbor A. Woyczyński, *Stable measures and central limit theorems in spaces of stable type*, Trans. Amer. Math. Soc. **251** (1979), 71–102, DOI 10.2307/1998684. MR531970 (81i:60010)

[38] Kh. O. Ondar, ed., *The Correspondence Between A.A. Markov and A.A. Chuprov on the Theory of Probability and Mathematical Statistics.* Translated from the Russian *O teorii veroiatnosteĭ i matematicheskoĭ statistike* by Charles and Margaret Stein. Springer Verlag, New York, Heidelberg, Berlin 1981.

[39] K. Pearson, *James Bernoulli's theorem.* Biometrika **17** (1925), 201–210.

[40] V. V. Petrov, *O lokal'nyh predel'nyh teoremah dlya summ nezavisimyh slučaĭnyh veličin.* Teoriya veroyatn. i ee primen. **9** No. 2 (1964), 343–352. On local limit theorems for sums of independent random variables. Theor. Probab. Appl. **9** No. 2 (1964), 312–320.

[41] V. V. Petrov, *Sums of independent random variables*, Springer-Verlag, New York, 1975. Translated from the Russian by A. A. Brown; Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 82. MR0388499 (52 #9335)
[42] Ivo Schneider, *Direct and indirect influences of Jakob Bernoulli's Ars Conjectandi in 18th century Great Britain*, J. Électron. Hist. Probab. Stat. **2** (2006), no. 1, 17 (English, with English and French summaries). MR2393219
[43] G. Shafer, *The significance of Jacob Bernouilli's Ars Conjectandi for the philosophy of probability today.* J. of Econometrics **75** (1996), 15–32.
[44] L. A. Shepp, *A local limit theorem*, Ann. Math. Statist. **35** (1964), 419–423. MR0166817 (29 #4090)
[45] S. R. S. Varadhan, *Asymptotic probabilities and differential equations*, Comm. Pure Appl. Math. **19** (1966), 261–286. MR0203230 (34 #3083)
[46] B. L. van der Waerden, *Kommentar zu den Meditationes und der Ars Conjectandi.* In: *Die Werke von Jakob Bernoulli,* Vol. 3, 353–383, Birkhäuser Basel 1975. (German.)
[47] R. von Mises, *Grundlagen der Wahrscheinlichkeitsrechnung*, Math. Z. **5** (1919), no. 1-2, 52–99, DOI 10.1007/BF01203155 (German). MR1544374

Department of Mathematics, The Pennsylvania State University, State College, Pennsylvania 16802
E-mail address: denker@math.psu.edu