Edited by: **Abe Shenitzer**
Mathematics, York University, North York, Ontario M3J 1P3, Canada

The Development of Rigor in Mathematical Probability (1900–1950)

Joseph L. Doob

1 Introduction. This paper is a brief informal outline of the history of the introduction of rigour into mathematical probability in the first half of this century. Specific results are mentioned only in so far as they are important in the history of the logical development of mathematical probability.

The development of science is not a simple progression from one advance to the next. Judged by hindsight, the development is slow, proceeds in a zigzag course, with many wrong turns and blind alleys, and frequently moves in directions condemned by leading scientists. In the 1930's Banach spaces were sneered at as absurdly abstract, later it was the turn of locally convex spaces, and now it is the turn of nonstandard analysis. Mathematicians are no more eager than other humans to embrace new ideas, and full acceptance of mathematical probability was not realized until the second half of the century. In particular, many statisticians and probabilists resented the mathematization of probability by measure theory, and some still place mathematical probability outside analysis. The following quotations (in translation) are relevant.

Planck: *A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up with it.*

Poincaré: *Formerly, when one invented a new function, it was to further some practical purpose; today one invents them in order to make incorrect the reasoning of our fathers, and nothing more will ever be accomplished by these inventions.*

Hermite: (in a letter to Stieltjes) *I recoil with dismay and horror at this lamentable plague of functions which do not have derivatives.*

Probability theory began, and remained for a long time, an idealization and analysis of certain real life phenomena outside mathematics, but gradually, in the first half of this century, mathematical probability became a normal part of mathematics. The mathematization of probability required new ideas, and in particular required a new approach to the idea of acceptability of a function. In

Reprinted with kind permission of Birkhäuser Verlag AG, from *Development of Mathematics 1900–1950*, edited by Jean-Paul Pier, Basel, 1994, ISBN 0-8176-2821-5

view of the above quotations it is not surprising that acceptance of this mathematization was slow and faced resistance. In fact even now some probabilists fear that mathematization has removed the intrinsic charm from their subject. And they are right in the sense that the charm of the old, vague probability-mathematics, based on nonmathematical definitions, has split into two quite different charms: those of real world probability and of mathematical precision. But it must be stressed that many of the most essential results of mathematical probability have been suggested by the nonmathematical context of real world probability, which has never even had a universally acceptable definition. In fact the relation between real world probability and mathematical probability has been simultaneously the bane of and inspiration for the development of mathematical probability.

2 What is the real world (nonmathematical) problem? What is usually called (real world) probability arises in many contexts. Besides the obvious contexts of gambling games, of insurance, of statistical physics, there are such simple contexts as the following. Suppose an individual rides his bicycle to work. The rider would be surprised if, when the bicycle is parked, the valve on the front tire appeared in the upper half of the tire circle 10 successive days, just as surprised as if 10 successive tosses of a coin all gave heads. However, it is clear that (tire context) if the ride is very short, or (coin context) if the coin starts close to the coin landing place and the initial rotational velocity of the coin is low, the surprise would decrease and the probability context would become suspect. The moral is that the specific context must be examined closely before any probabilistic statement is made. If philosophy is relevant, an arguable question, it must be augmented by an examination of the physical context.

3 The law of large numbers. In a repetitive scheme of independent trials, such as coin tossing, what strikes one at once is what has been christened the *law of large numbers*. In the simple context of coin tossing it states that in some sense the number of heads in n tosses divided by n has limit $1/2$ as the number of tosses increases. The key words here are *in some sense*. If the law of large numbers is a mathematical theorem, that is, if there is a mathematical model for coin tossing, in which the law of large numbers is formulated as a mathematical theorem, either the theorem is true in one of the various mathematical limit concepts or it is not. On the other hand, if the law of large numbers is to be stated in a real world nonmathematical context, it is not at all clear that the limit concept can be formulated in a reasonable way. The most obvious difficulty is that in the real world only finitely many experiments can be performed in finite time. Anyone who tries to explain to students what happens when a coin is tossed mumbles words like *in the long run*, *tends*, *seems to cluster near*, and so on, in a desperate attempt to give form to a cloudy concept. Yet the fact is that anyone tossing a coin observes that for a modest number of coin tosses the number of heads in n tosses divided by n seems to be getting closer to $1/2$ as n increases. The simplest solution, adopted by a prominent Bayesian statistician, is the vacuous one: never discuss what happens when a coin is tossed. A more common equally satisfactory solution is to leave fuzzy the question of whether the context under discussion is or is not mathematics. Perhaps the fact that the assertion is called a *law* is an example of this fuzziness. The following statements have been made about this law (my emphasis):

Laplace: (1814) *This <u>theorem</u>, implied by common sense, was difficult to prove by analysis.*

Ville: (1939) *One sees no reason for this proposition to be true; but as it is impossible to prove experimentally that it is false, one can at least safely state it.*

Bauer: (translated from the context of dice to that of coins) *It is an experimentally established fact that the quotient... exhibits a deviation from* $1/2$ *which approaches* 0 *for large* n.

These statements illustrate the enduring charm of discussions of real world probability. Mathematicians, unfortunately, have felt forced to think about the following question, or at least to write about it.

4 What is probability? Here are some attempts to answer this question and to discuss the teaching of the subject.

Poincaré: (1912) *One can scarcely give a satisfactory definition of probability.*

Mazurkiewicz: (1915) *The theory of probability is not an independent element of mathematical instruction; nevertheless it is very desirable that a mathematician knows its general principles. Its fundamental concepts are incompletely determined. They contain many unsolved difficulties.*

v. Mises: (1919) *In fact, one can scarcely characterize the present state other than that probability is not a mathematical discipline.* (He proceeded to make it into a mathematical discipline by basing mathematical probability on a sequence of observations («Beobachtungen») with properties that cannot be satisfied by a mathematically well defined sequence. In a lighter mood he is said to have defined probability as a number between 0 and 1 about which nothing else is known.)

E. Pearson: (1935) (Oral communication) *Probability is so linked with statistics that, although it is possible to teach the two separately, such a project would be just a tour de force.*

Uspensky: (1937) In a useful textbook, he gave the following once common textbook definition. *If, consistent with condition* S, *there are* n *mutually exclusive, and equally likely cases, and* m *are favorable to the event* A, *then the mathematical probability of* A *is defined as* m/n.

The foregoing should make obvious the advisability of separating mathematical probability theory from its real world applications. Note however that no one doubts the real world applicability of mathematical probability. Gambling, genetics, insurance and statistical physics are here to stay.

Only *mathematical* probability will be discussed below, except for the following remark on coin tossing. Newtonian mechanics provides a partial mathematical model for coin tossing. In coin tossing, a solid body falls under the influence of gravity. Its motion is determined in Newton's model by his laws, and any discussion of what the coin does cannot be complete unless these laws are applied. Only these laws, rather than philosophical remarks, can explain the quantitative influence and importance of the initial and final conditions of the coin motion in order to justify allusions to equal likelihood of heads and tails. Of course these laws can at best reduce the analysis to considerations of the initial and final conditions of the toss, but these conditions can show what the «equal likelihoods» depends on and thereby give it a plausible interpretation.

5 Mathematical probability before the era of precise definitions. There were many important advances in mathematical probability before 1900, but the subject was not yet mathematics. Although nonmathematical probabilistic contexts suggested problems in combinatorics, difference equations and differential equations, there was a minimum of attention paid to the mathematical basis of the contexts, a maximum of attention to the pure mathematics problems they suggested. This unequal treatment was inevitable, because measure theory, needed for mathematical modeling of real world probabilistic contexts, had not yet been invented.

It was always clear that, however classical mathematical probability was to be developed, the concept of additivity of probability as applied to incompatible real world events was fundamental. Additive functions of sets were of course familiar to mathematicians from concepts of volume, mass and so on, long before 1900. It was realized that contexts involving averages led to probability. It was frequently clear how to use the contexts to suggest problems in analysis, but it was not clear how to formulate an overall mathematical context, that is, how to define a mathematical structure in which the various contexts could be placed.

A weaker condition than additivity was less familiar but turned out to be essential later. The standard loose language will be used here. If $x_1, x_2, \ldots$ are numbers obtained by chance, and if A is a set of numbers, consider the probability that at least one of the members of this sequence lies in A, or, in more colorful language, consider the probability that an orbit of this motion through points of a line hits A. The usual calculation (ignoring here all notions of rigour), defines a function $A \to \phi(A)$ which in general is not additive. In fact ϕ satisfies the inequality

$$\phi(A) + \phi(B) - \phi(A \cup B) \geq \phi(A \cap B), \tag{5.1}$$

whereas additivity of ϕ would imply equality in (5.1). The point is that the left side of (5.1) is the probability that the sequence $x\bullet$ hits both A and B, a probability at least equal to, and in general greater than, $\phi(A \cap B)$, the probability that the sequence hits $A \cap B$. The inequality (5.1), the *strong subadditivity* inequality, is satisfied also by the electrostatic capacity of a body in $\mathbf{R}^3$, and this fact hints at the close connection between potential theory and probability, developed in great detail in the second half of the century with the help of Choquet's theory of mathematical capacity.

6 The development of measure theory. Recall that a *Borel field* (= σ algebra) of subsets of a space is a collection of subsets which is closed under the operations of complementation and the formation of countable unions and intersections. The class of *Borel sets* of a metric space is the smallest set σ algebra containing the open sets of the space. A *measurable space* is a pair, $(S, \mathbb{S})$, where S is a space and $\mathbb{S}$ is a σ algebra of subsets of S. The sets of $\mathbb{S}$ are the *measurable sets* of the space. In the following, if S is metric, the coupled σ algebra making it into a measurable space will always be the σ algebra of its Borel sets. In particular $(\mathbf{R}^N, \mathbb{R}^N)$ denotes N dimensional Euclidean space coupled with its Borel sets. The superscript will be omitted when $N = 1$. A measurable function from a measurable space $(S_1, \mathbb{S}_1)$ into a measurable space $(S_2, \mathbb{S}_2)$ is a function from S_1 into S_2 with the property that the inverse image of a set in $\mathbb{S}_2$ is a set in $\mathbb{S}_1$.

Measure theory started with Lebesgue's thesis (1902), which extended the definition of volume in $\mathbf{R}^N$ to the Borel sets. Radon (1913) made the further step to general measures of Borel sets of $\mathbf{R}^N$ (finite on compact sets). These measures are usually extended to slightly larger classes than the class of Borel sets, by *completion*. Finally Fréchet (1915), 13 years after Lebesgue's thesis, pointed out that all that the usual definitions and operations of measure theory require is a σ algebra of subsets of an abstract space on which a measure, that is, a positive countably additive set function, is defined. At each step of this progression not necessarily positive countably additive set functions—signed measures—were incorporated into the theory. As noted below, the Radon-Nikodym theorem (1930), which gives conditions necessary and sufficient that a countably additive function of sets can be expressed as an integral over the sets, turned out to be the final essential result needed to formulate the basic mathematical probability definitions. Thus it was 28 years before Lebesgue's theory was extended far enough to be adequate for the mathematical basis of probability. This extension was not developed in order to provide a basis for probability, however. Measure theory was developed as a part of classical analysis, and applications in analysis were immediate, for example to the (Lebesgue measure) almost everywhere derivability of a monotone function.

There has been criticism of the fact that mathematical probability is usually prescribed not only to be additive but even to be countably additive. The question whether real world probability is countably additive, if the question is to be meaningful, asks whether a mathematical model of real world probabilistic phenomena *necessarily* always involves *countably* additive set functions. In fact there may well be real world contexts for which the appropriate mathematical model is based on finitely but not countably additive set functions. But there have been very few applications of such set functions in either mathematical or nonmathematical contexts, and such set functions will not be discussed further here.

7 Early applications of explicit measure theory to probability. Some probabilistic slang will be needed, enduring relics of the historical background of probability theory. A probability space is a triple $(S, \mathbb{S}, \mathrm{P})$, where $(S, \mathbb{S})$ is a measurable space and P is a measure on $\mathbb{S}$ with $\mathrm{P}(S) = 1$. A measure with this normalization is a *probability measure*. A *random variable* is a measurable function from a probability space $(S, \mathbb{S}, \mathrm{P})$, into a measurable space $(S', \mathbb{S}')$. The space S', or, when one writes carefully, $(S', \mathbb{S}')$, is the *state space* of the random variable. Mutual independence of random variables is defined in the classical way. The *distribution* of a random variable x is the measure P_x on $\mathbb{S}'$ defined by setting

$$P_x(A') = P\{s \in S : x(s) \in A'\}.$$

The joint distribution of finitely many random variables defined on the same probability space is obtained by making x into a vector and specifying $\mathbb{S}'$ and S' correspondingly. A stochastic process is a family of random variables $\{x(t,\bullet), t \in \mathcal{I}\}$ from some probability space $(S, \mathbb{S}, \mathrm{P})$, into a state space $(S', \mathbb{S}')$. The set $\mathcal{I}$ is the *index set* of the process. Thus a stochastic process defines a function of two variables, $(t, s) \to x(t, s)$, from $\mathcal{I} \times S$ into the state space. The function $x(t,\bullet)$ from S into S' is the tth random variable of the process; the function $x(\bullet, s)$ from $\mathcal{I}$ into S' is the sth sample function, or sample path, or sample sequence if $\mathcal{I}$ is a sequence.

Borel (1909) pointed out that in the dyadic representation $x = x_1 x_2 \ldots$ of a number x between 0 and 1, in which each digit x_j is either 0 or 1, these digits are functions of x, and if the interval $[0, 1]$ is provided with Lebesgue measure, a probability measure on this interval, these functions miraculously become random variables which have exactly the distributions used in calculating coin tossing probabilities. That is, 2^{-n} is the probability assigned to the event that, in a tossing experiment, the first n tosses yield a specified sequence of heads and tails, and 2^{-n} is also the total length (= Lebesgue measure) of the finite set of intervals whose points x have dyadic representations with a specified sequence of 0's and 1's in the first n places. Thus a mathematical version of the law of large numbers in the coin tossing context is the existence in some sense of a limit of the sequence of function averages $\{(x_1 + \cdots + x_n)/n,\ n \geq 1\}$. Classical elementary probability calculations imply that this sequence of averages converges in measure to $1/2$, but a stronger mathematical version of the law of large numbers was the fact deduced by Borel—in an unmendably faulty proof—that this sequence of averages converges to $1/2$ for (Lebesgue measure) almost every value of x. A correct proof was given a year later by Faber, and much simpler proofs have been given since. [Fréchet remarked tactfully: «Borel's proof is excessively short. It omits several intermediate arguments and assumes certain results without proof.»] This theorem was an important step, an example of a new kind of convergence theorem in probability. Observe that (fortunately) pure mathematicians need not interpret this theorem in the real world of real people tossing real coins. Some of the quotations given above indicate that they not only need not but should not.

Daniell (1918) used a deep approach to measure theory in which integrals are defined before measures to get a (rather clumsy) approach to infinite sequences of random variables by way of measures in infinite dimensional Euclidean space.

The Brownian motion stochastic process in $\mathbf{R}^3$ is the mathematical model of Brownian motion, the motion of a microscopic particle in a fluid as the particle is hit by the molecules of the fluid. The process is normalized by supposing it starts at the origin of a cartesian coordinate system in $\mathbf{R}^3$, and a (normalized) Brownian motion process in $\mathbf{R}$ is the process of a coordinate function of a normalized process in $\mathbf{R}^3$, vanishing initially. A (normalized) Brownian motion process in $\mathbf{R}^N$ is a process defined by N mutually independent Brownian motion processes in $\mathbf{R}$. It was well known what the joint distributions of the random variables of a Brownian motion process should be, and it had been taken for granted that in a proper mathematical model the class of continuous paths would have probability 1. By 1900, Bachelier had even derived various important distributions related to the Brownian motion process in $\mathbf{R}$, such as that of the maximum change during a time interval, by finding corresponding distributions for a certain discrete random walk and then going to the limit as the walk steps tended to 0. More precisely, what Bachelier derived were distributions valid for a Brownian motion process if in fact there was such a thing as a Brownian motion process, and if it was approximable by his random walks. Observe that there was no question about the existence of Brownian motion; Brownian motion is observable under a microscope. But there was as yet no proof of the existence of a stochastic process, a mathematical construct, with the desired properties. Wiener (1923) constructed the desired Brownian motion process, now sometimes called the *Wiener process*, by applying the Daniell approach to measure theory to obtain a measure with the desired properties on a space S of continuous functions: if $x(t,\bullet)$ is the random variable defined by the value at time t of a function in S, the stochastic process of these random variables is a stochastic process with sample functions the members of S, and with the joint random variable distributions those prescribed for the Brownian motion process.

Bachelier's results remained unnoticed for years, and in fact were rediscovered several times. Wiener's work, like his fundamental work in potential theory, had little immediate influence because it was published in a journal which was not widely distributed. It was an aspect of his genius that he carried out his Brownian motion research then and later without knowledge of the slang and some of the useful elementary mathematical techniques of probability theory.

Steinhaus (1930) demonstrated that classical arguments to derive standard probability theorems could be placed in a rigorous context by taking Lebesgue measure on a linear interval of length 1 as the basic probability measure, interpreting random variables as Lebesgue measurable functions on this interval, and expectations of random variables as their integrals. No new proofs were required; all that was required was a proper translation of the classical terminology into his context. If this were all mathematization of probability by measure theory had to offer, the scorn of rigorous mathematics expressed by some nonmathematicians would be justified.

8 Kolmogorov's 1933 monograph. Kolmogorov (1933) constructed the following mathematical basis for probability theory.

(a) The context of mathematical probability is a probability space $(S, \mathbb{S}, \mathrm{P})$. The sets in $\mathbb{S}$ are the mathematical counterparts of real world events; the points of S are counterparts of elementary events, that is of individual (possible) real world observations.

(b) Random variables on $(S, \mathbb{S}, \mathrm{P})$, are the counterparts of functions of real world observations. Suppose $\{x(t,\bullet), t \in \mathscr{I}\}$ is a stochastic process on a probability space $(S, \mathbb{S}, \mathrm{P})$, with state space S'. A set of n of the process random variables has a probability distribution on S'^n. Such finite dimensional distributions are mutually compatible in the sense that if $1 \leq m < n$, the joint distribution of $x(t_1,\bullet), \ldots, x(t_m,\bullet)$ on S'^m is the m-dimensional distribution induced by the n-dimensional distribution of $x(t_1,\bullet), \ldots,$ $x(t_n,\bullet)$ on S'^n.

(c) Conversely, Kolmogorov proved that given an arbitrary index set $\mathscr{I}$, and a suitably restricted measurable space $(S'', \mathbb{S}')$ (for example, the measurable space can be a complete separable metric space together with the σ algebra of its Borel sets) and a mutually compatible set of distributions on S'^n, for integers $n \geq 1$, indexed by the finite subsets of $\mathscr{I}$, there is a probability space and a stochastic process $\{x(t,\bullet), t \in \mathscr{I}\}$ defined on it, with state space S', with the assigned joint random variable distributions. To prove this result he constructed a probability measure on a σ algebra of subsets of the product space $S'^{\mathscr{I}}$, the space of all functions from $\mathscr{I}$ into S', and obtained the required random variables as the coordinate functions of $S'^{\mathscr{I}}$.

(e) The expectation of a numerically valued integrable random variable is its integral with respect to the given probability measure.

(f) The classical definition of the conditional probability of an event (measurable set) A, given an event B of strictly positive probability, is $\mathrm{P}(A \cap B)/P(B)$. In this way, for fixed B, new probabilities are obtained, and expectations of random variables for given B are computed in terms of these new *conditional* probabilities. More generally, given an arbitrary collection of random variables, conditional probabilities and expectations relative to given values of those random variables are needed, functions of the values assigned to the conditioning random variables. If $(S, \mathbb{S}, \mathrm{P})$ is a probability space, and if a collection of random variables is given, let $\mathbb{F}$ be the smallest sub σ algebra of $\mathbb{S}$ relative to which all the given random variables are measurable. This σ algebra is the *σ algebra generated by conditions imposed on the given random variables*. A reasonable interpretation of a measurable real valued function of the given collection of random variables is a measurable function from $(S, \mathbb{F})$ into $\mathbf{R}$. The Kolmogorov *conditional expectation* of a real valued integrable random variable x on $(S, \mathbb{S}, \mathrm{P})$, relative to a σ algebra $\mathbb{C}$ of measurable sets, is a random variable which is measurable relative to $\mathbb{C}$ and has the same integral as x on every set in $\mathbb{C}$. The existence of such a random variable, and its uniqueness up to P-null sets, is assured by the Radon-Nikodym theorem. The conditional expectation of x relative to a collection of random variables is defined as the conditional expectation of x relative to the σ algebra generated by conditions on the random variables. A conditional probability of a measurable set A is defined as the conditional expectation of the random variable which is 1 on A and 0 elsewhere.

Kolmogorov's 1933 exposition paints a discouraging picture of mathematical progress. In the first pages of his monograph he states explicitly that real valued random variables are measurable functions and expectations are their integrals. Even as late as 1933, however, he must have thought that mathematicians were not familiar with measure theory. In fact in the body of his monograph, when he comes to the definition of a real valued random variable, he does not simply refer back to the first pages of the monograph and say that a random variable is a measurable function. Instead he actually defines measurability of a real valued function, and similarly when he defines the expected value of a random variable he does not simply state that it is the integral of the random variable with respect to the given probability measure, but he actually defines the integral. Later in the monograph, when he needs Lebesgue's theorem allowing taking limits of convergent function sequences under the sign of integration, he does not simply refer to Lebesgue but gives a detailed proof of what he needs. As confirmation of Kolmogorov's caution in invoking measure theory, the author recalls his student experience in 1932 when there were professorial disapproving remarks on the extreme generality of a seminar lecture given by Saks on what is now called the Vitali-Hahn-Saks theorem, a theorem which has since become an important tool in probability theory. [He also recalls that he did not understand the point of Kolmogorov's measure on a function space until long after he had read the monograph.]

It was some time before Kolmogorov's basis was accepted by probabilists. The idea that a (mathematical) random variable is simply a function, with no romantic connotation, seemed rather humiliating to some probabilists. A prominent statistician in 1935 wondered whether two orthogonal real valued random variables with zero means (integrals) are necessarily independent, as they are under the added hypothesis that they have a bivariate Gaussian distribution. He was rather surprised by the example of the sine and cosine functions on the interval $[0, 2\pi]$, with probability measure defined as Lebesgue measure divided by 2π. These two functions, orthogonal and with zero means but not independent, are not the kind of random variables probabilists were used to. Some analysts may be gratified, some humiliated, to learn that in discussing Fourier series they can be accused of discussing probabilities and expectations.

9 Expansion backwards of the Kolmogorov basis. Kolmogorov's basis for mathematical probability can be expanded, and should be expanded in the view of some probabilists, who want to start with some not necessarily numerical mathematical version of the confidence of observers that certain events will occur, and to proceed postulationally to numerical evaluations of this confidence, and finally to additivity. Such an analysis may be enlightened in discussing the appropriateness of mathematical probability as a model for real world phenomena, but any approach to the subject which ends with a justification of the classical calculations and is mathematically usable, will end with Kolmogorov's basis, however phrased, because all the measure basis to probability does is to give a formal precise mathematical framework for the classical calculations and their present refinements. This framework had made it possible to apply mathematical probability in many other mathematical fields, for example to potential theory and partial differential equations. Although such applications were made in the past before the acceptance of measure theory as the basis of probability, the probabilistic context served only to suggest mathematics and was not an integral part of the mathematics. The meaning of solutions as probabilities and expectations could not be formulated and exploited.

10 Uncountable index sets. If the index set $\mathscr{T}$ of a stochastic process $\{x(t,\bullet), t \in \mathscr{T}\}$ is an interval of the line, and if the state space of the random variables is $\mathbf{R}$, the class of continuous sample functions may not be measurable. This difficulty arises in the processes derived by the Kolmogorov construction of a measure on a function space, for example, whatever the choice of joint distributions of the process random variables. To understand the difficulty, observe that if the index set $\mathscr{T}$ of a stochastic process with state space $\mathbf{R}$ is an interval, and if $\mathscr{J}$ is a subset of $\mathscr{T}$, the function $s \to \sup_{t \in \mathscr{J}} x(t,s)$ is measurable if $\mathscr{J}$ is countable, but need not be measurable if $\mathscr{J}$ is uncountable. If boundedness and continuity of sample functions are to be discussed, some modification of the probability relations of the random variables of a stochastic process should be devised to make such suprema measurable functions. A clumsy approach was proposed by Doob (1937) but a more usable one was not devised until after 1950.

11 Reluctance to accept measure theory by probabilists. There was considerable resistance to the acceptance and exploitation of measure theory by probabilists, both in Kolmogorov's day and later. The following quotation is an example of the reluctance of some mathematicians to separate the mathematics from the context that inspired it.

Kac (1959) *How much fuss over measure theory is necessary for probability theory is a matter of taste. Personally I prefer as little fuss as possible because I firmly believe that probability theory is more closely related to analysis, physics and statistics than to measure theory as such.*

12 New relations between functions made possible by the mathematization of probability. Probability theory suggested new relations between functions. For example consider the sequence $x_1, x_2, \ldots$ of real valued integrable random variables on a probability space $(S, \mathbb{S}, \mathrm{P})$ and suppose that the conditional expectation of x_n given $x_1, \ldots, x_{n-1}$ vanishes (P) almost everywhere, for $n > 1$, that is, the integral of x_n over any set determined by conditions on the preceding random variables vanishes. If these random variables are square integrable, this condition is equivalent to the condition, much stronger than mutual orthogonality, that x_n is orthogonal to every square integrable function of $x_1, \ldots, x_n$. Bernstein (1927) seems to have been the first to treat such sequences systematically. This condition on a sequence of functions means that in a reasonable sense the sequence of partial sums of the given sequence is the counterpart of a fair game. In fact, the partial sums $y_1, y_2, \ldots$ are characterized by the property that the expectation of y_n

relative to $y_1, \ldots, y_{n-1}$ is equal to y_{n-1} almost everywhere on the probability space. Processes with this property, called *martingales*, first used explicitly by Ville (1939), have had many applications, for example to partial differential equations, to derivation, and to potential theory. Another important class of sequences of random variables is the class of sequences with the Markov property. These sequences are characterized by the fact that when $n \geq 1$ the conditional probabilities for x_n relative to $x_1, \ldots, x_{n-1}$ are equal almost everywhere to those for x_n relative to x_{n-1}. Roughly speaking, the influence of the present, given the past, depends only on the immediate past. The Markov property, introduced in a very special case by Markov in 1906, (named in his honor by others) has proved very fruitful, for example, leading in the second half of the century to a probabilistic potential theory, generalizing and including classical potential theory.

13 What is the place of probability theory in measure theory, and more generally in analysis? It is considered by some mathematicians that if one deals with analytic properties of probabilities and expectations then the subject is part of analysis, but that if one deals with sample sequences and sample functions then the subject is probability but not analysis. These authors are in the interesting situation that in considering a function of two variables, $(t, s) \to x(t, s)$—as in considering stochastic processes—they call it analysis if the family of functions $x(t,\bullet)$ as t varies is studied, but call it probability and definitely not analysis if the family of functions $x(\bullet, s)$ as s varies is studied. More precisely, they regard discussions of distributions and associated questions as analysis, but not discussions in terms of sample functions. This point of view is expressed in the following quotation.

Protter *By developing his integral in* 1944 *with stochastic processes as integrands, Itô was able to study multidimensional diffusions with purely probabilistic techniques, an improvement over the analytic methods of Feller.*

The following remark on the convergence of a sum of orthogonal functions illustrates the difficulty in separating (mathematical) probability from the rest of analysis. The measure space is a probability space, but with trivial changes the discussion is valid for any finite measure space.

If $x\bullet$ is an orthogonal sequence of functions, on a probability measure space, and if x_n^2 has integral σ_n^2, then (Riesz-Fischer) $\sum x_n$ converges in the mean if

$$\sum \sigma_n^2 < +\infty. \tag{13.1}$$

The orthogonal series converges almost everywhere if either (Menšov-Rademacher) (13.1) is strengthened to

$$\sum \sigma_n^2 \log^2 n < +\infty, \tag{13.2}$$

or (Lévy, 1937) the condition (13.1) is kept but the orthogonality condition is strengthened to the condition in Section 12.

The reader should judge which of these results is measure theoretic and which is probabilistic, whether there is any point in evicting mathematical probability from analysis, and if so whether measure theory should also be evicted.

101 West Windsor Road, Apt. 1104
Urbana, Illinois 61801-6663
doob@symcom.math.uiuc.edu