[49] J. L. Doob. Application of the theory of martingales. In Le Calcul des Probabilit´es et ses Applications., pages 23–27. Centre National de la Recherche Scientifique, Paris, 1949. Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13,.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQU

XIII

LE CALCUL DES PROBABILITÉS ET SES APPLICATIONS

(Lyon - 28 Juin au 3 Juillet 1948)

RENSEIGNEMENTS ET VENTE
AU SERVICE DES PUBLICATIONS DU C. N. R. S.
45, rue d'Ulm — PARIS (5e)

1949

APPLICATION OF THE THEORY OF MARTINGALES

by J. L. DOOB.

(Urbana, Ill., U. S. A.)

I. Introduction and definitions.

Let $\ldots, x_{-1}, x_0, x_1, \ldots$ be a sequence of random variables with $E\{|x_n|\} < \infty$ [1] for all n. The sequence is called a martingale if for every n

$$E\{\ldots, x_{n-2}, x_{n-1} \backslash x_n\} = x_{n-1}, \tag{1.1}$$

with probability 1. The sequence x_n may be infinite in only one direction.

Although other authors had derived many martingale properties, in various forms, Ville [2] was the first to study them systematically, and to show their wide range of applicability. In the following sections new applications of martingale theory, using more recent results, will be made, to the strong law of large numbers and to statistical theory. The following theorems are stated for later reference [3]. In all cases it is supposed that the variables involved satisfy the martingale property (1.1).

(i) $E(x_n)$ does not depend on n, but

$$\ldots \leqslant E\{|x_n|\} \leqslant E\{|x_{n+1}|\} \leqslant \ldots.$$

(ii) $\lim\limits_{n \to -\infty} x_n = x_{-\infty}$ exists with probability 1; $E\{x_{-\infty}\}$ exists and
$E\{x_{-\infty}\} = E\{x_n\}$. The sequence $\{x_n\}$, $n \leqslant 0$, is uniformly integrable.

(iii) If $\lim\limits_{n \to \infty} E\{|x_n|\} = l < \infty$, it follows that $\lim\limits_{n \to \infty} x_n = x_\infty$ exists with probability 1; $E\{x_\infty\}$ exists and $E\{|x_\infty|\} \leqslant l$. In particular if the x_n' s are all $\geqslant 0$, $l = E\{x_n\} < \infty$ and the stated conclusions hold.

(iv) If $\ldots, y_0, y_1, \ldots$ y are any random variables with $E\{|y|\} < \infty$, and if x_n is defined by

$$x_n = E\{\ldots, y_{n-1}, y_n \backslash y\}$$

then the sequence $\{x_n\}$ is a martingale. Moreover sinc $E\{|x_n|\} \leqslant E\{|y|\}$, (iii) is applicable. In this case

$$x_\infty = E\{\ldots, y_0, y_1, \ldots \backslash y\}.$$

2. Application to the strong law of large numbers.

Let $u_1, u_2, \ldots$ be mutually independent randor variables with a common distribution function, and sup pose that $E\{u_n\}$ exists. Then

$$\lim_{n \to \infty} \frac{u_1 + \cdots + u_n}{n} = E\{u_1\} \tag{2.1}$$

with probability 1. This particular case of the stron law of large numbers for stationary processes can b derived from martingale theory as follows.

Define $\ldots, y_{-2}, y_{-1}$ by

$$y_{-n} = u_1 + \ldots + u_n$$

and x_n by

$$\begin{aligned} x_{-n} &= E\{\ldots, y_{-n-1}, y_{-n} \backslash y_{-1}\} \\ &= E\{y_{-n} \backslash y_{-1}\} = \frac{u_1 + \cdots + u_n}{n} \end{aligned}$$

Then by (iv) the sequence $\{x_n\}$ is a martingale, an hence $\lim\limits_{n \to -\infty} x_n$ exists with probability 1, giving the existenc of the limit in (2.1). The limit is easily identified a $E\{u_1\}$.

3. Application to inverse probabilities.

Suppose that for each value of parameter θ, whic varies in some Borel linear set $f(\theta, y)$ is a probabilit

[1] In the following $E\{x\}$ will denote the expectation of x and $E\{y \backslash x\}$ will denote the (conditional) expectation of x for given y. In most cases below y will be replaced by infinitely many conditioning variables.

[2] Etude critique de la notion de collectif, Paris 1939.

[3] Doob, transactions of the American Mathematical Society 47 (1940) 455-486.

density in y, corresponding to the distribution function F (θ, y),

$$(3.1) \qquad \mathrm{F}(\theta, y) = \int_{-\infty}^{y} f(\theta, z)\, dz.$$

Hypothesis (A): *F (θ, y) is a Baire function of (θ, y)*. For this to be true it is necessary and sufficient that F (θ, y) be a Baire function of θ for each value of y. This is the only regularity hypothesis to be imposed on F (θ, y); it implies that $f(\theta, y)$ can also be assumed to be a Baire function of (θ, y).

Suppose that θ is chosen in accordance with some probability distribution having density $f(\theta)$ and that n sample values of y, $y_1, \ldots, y_n$ are then drawn independently using the $f(\theta, y)$ distribution. We shall apply martingale theory to the classical problem of estimating the value of θ chosen, from a knowlecdge of the sample values $y_1, \ldots, y_n$.

Hypothesis (B). *To different values of θ correspond different distributions,* that is if $\theta_1 \neq \theta_2$, then F (θ_1, y) and F (θ_2, y) are not identically equal.

Hypotheses (A) and (B) are the only hypotheses to be imposed on F (θ, y). The estimation problem stated above can only be solved if for large n the value of θ drawn is very nearly a function of $y_1, \ldots, y_n$. This relationship can be formalized as follows. For every n the quantities $\theta, y_1, \ldots, y_g$ are random variables with joint density $f(\theta) \prod_{j=1}^{n} f(\theta, y_j)$. If $n = \infty$ we have a sequence of random variables $\theta, y_1, y_2, \ldots$, and we now show that θ is a function of $y_1, y_2, \ldots$. To show how to calculate F (θ, y) (and therefore θ, by Hypothesis (B) this we show in terms of the sequence $y_1, y_2, \ldots$. This is simply done using the form of the strong law of large numbers proved in Section 2. In accordance with this law, if $\nu_n(y)$ is the number of the first n y_j s which are $\leq y$

$$(3.2) \qquad \mathrm{F}(\theta, y) = \lim_{n \to \infty} \frac{\nu_n(y)}{n}$$

with probability 1, for each y. Hence (3.2) holds with probability 1 simultaneously for all rational values of y. Excluding the exceptional y_j sequences, the *nth* term on the right, for a given y_j sequence, is a distribution function in y which converges for all rational y to F (θ, y). Thus (3.2y) as now interpreted states the distribution function of the first n y_j's converges to F (θ, y) with probability 1. It thus follows from Hypothesis (B) that θ is a function of y_j sequences (neglecting zero probabilities as usual), and in the following we shall write $\hat{\theta}$ or $\theta(y_1, y_2, \ldots)$ for the variable considered in this connection.

Preliminary Hypothesis (C), (implied by (A) and (B), see below).

$\hat{\theta}$ is a measurable function of y_j sequences, that is a random variable on $y_1, y_2, \ldots$ sample space. The significance of this measurability hypothesis will be discussed below. It is essentially a restriction on the regularity of F (θ, y) in θ.

Assuming Hypothesis (A) the conditional distribution of θ for given $y_1, \ldots, y_n$ has density (roughly the probility that $\theta = \alpha$)

$$(3.3) \qquad \frac{f(\alpha) \prod_{1}^{n} f(\alpha, y_j)}{\int f(\varphi) \prod_{1}^{n} f(\varphi, y_j)\, d\varphi}$$

where the integral is taken over the whole range of the parameter. Now when $n \longrightarrow \infty$, $y_1, \ldots, y_n$ becomes $y_1, y_2, \ldots$; any sequence $y_1, y_2, \ldots$ (excluding a class of sequences of total probability θ determines a value of $\hat{\theta}$ and the problem is to see how the distribution determined by (3.3) becomes concentrated at $\hat{\theta}$ when $n \longrightarrow \infty$. This is the full formalization of the estimation problem described in intuitive language at the beginning of this section. Hypothesis (B) is obviously a necessary condition if $\hat{\theta}$ is to be estimated from $y_1, \ldots, y_n$. In fact what is estimated is not θ but the distribution function F $(\hat{\theta}, y)$ and without Hypothesis (B) even if F $(\hat{\theta}, y)$ were known precisely $\hat{\theta}$ would not be uniquely determined. It is clear that some sort of regularity condition on F $(\hat{\theta}, y)$ in θ is also necessary since we can at best hope to approximate the true distribution function F $(\hat{\theta}, y)$ more and more closely as $n \longrightarrow \infty$ and must make enough hypotheses to insure that approximating F $(\hat{\theta}, y)$ means approximating $\hat{\theta}$. It is somewhat surprising that Hypotheses (A) and (B) are sufficient.

There are several statements it would be desirable to prove. For example it would be desirable to prove that the conditional distribution has as $n \longrightarrow \infty$ the « true value » $\hat{\theta} = \theta(y_1, y_2, \ldots)$ as its expectation that is

$$(3.4) \qquad \lim_{n \to \infty} \mathrm{E}\{y_1, \ldots, y_n \backslash \hat{\theta}\} = \lim_{n \to \infty} \frac{\int \alpha f(\alpha) \prod_{1}^{n} f(\alpha, y_j)\, d\alpha}{\int f(\varphi) \prod_{1}^{n} f(\varphi, y_j)\, d\varphi} = \hat{\theta}$$

with probability 1. Secondly it would be desirable to prove that the distribution becomes concentrated around the true value, as $n \longrightarrow \infty$, in the sense say that variance goes to O, that is

$$(3.5) \qquad \lim_{n \to \infty} \mathrm{E}\{y_1, \ldots, y_n \backslash \hat{\theta}^2\} - \mathrm{E}^2\{y_1 \ldots, y_n \backslash \hat{\theta}\} = 0$$

with probability 1. These two statements imply that the conditional probability of differing from the true value $\hat{\theta}$ by at least ε goes to o, as $n \longrightarrow \infty$ for every $\varepsilon > 0$. This fact can itself be formulated as follows : let $z_{\lambda,\mu} = 1$ when $\lambda \leqslant \hat{\theta} \leqslant \mu$ and $z_{\lambda,\mu} = 0$ otherwise. Then

$$\lim_{n\to\infty} Pr\{y_1, \ldots, y_n \backslash \lambda \leqslant \hat{\theta} \leqslant \mu\} \tag{3.6}$$

$$= \lim_{n\to\infty} \frac{\int_\lambda^\mu f(\alpha) \prod_1^n f(\alpha, y_j)\, d\alpha}{\int f(\varphi) \prod_1^n f(\varphi, y_j)\, d\varphi} = z_{\lambda\mu}$$

that is

$$\lim_{n\to\infty} E\{y_1, \ldots, y_n \backslash z_{\lambda,\mu}\} = z_{\lambda,\mu} \tag{3.6'}$$

with probability 1. Now equations (3.4), (3.5), and (3.6) are all trivial applications of the martingale theorems stated in Section 1. Equation (3.4) requires the existence of

$$E\{\hat{\theta}\} = \int_{-\infty}^{\infty} \theta f(\theta)\, d\theta;$$

equation (3.5) requires that of

$$E\{\hat{\theta}^2\} = \int_{-\infty}^{\infty} \theta^2 f(\theta)\, d\theta;$$

but equation (3.6) requires nothing beyond Hypotheses (A) and (B) (and (C) which will be shown below to be implied by them). We can even go further than (3.6) which states that there is convergence of the distributions, by examining the densities. In fact the conditional densities also converge in the sense that excluding a set of $\hat{\theta}$ values of probability, 0,

$$\lim_{n\to\infty} Pr\{y_1, \ldots, y_n \backslash \hat{\theta} = \alpha\} \tag{3.7}$$

$$= \lim_{n\to\infty} \frac{f(\alpha) \prod_1^n f(\alpha, y_j)}{\int f(\varphi) \prod_1^n f(\varphi, y_i)\, d\varphi} = 0$$

with probability 1. The first term in (3.7) is only suggestive, and explains why the sequence of random variables involved is a martingale, a fact which is easily checked directly (although it may be necessary to exclude a set of $\hat{\theta}$ values of probability 0). The fact that there is convergence then follows from (iii), Section 1, since the random variables are non-negative. The fact that the limit is 0 follows easily from the integrated version (3.6). A refinement of the argument can be used to prove that there is a limit ($\geqslant 0$) in (3.7) for all α.

We must still discuss the significance of Hypothesis (C), a condition which must be translated into a condition on the function F (θ, y). Without Hypothesis (C) the limits in (3.4), (3.5) and (3.6) would still exist, but would be different. For example that in (3.4) would become $E\{y_1, y_2, \ldots \backslash \hat{\theta}\}$.

It is quite possible mathematically to have a pathological situation in which $\hat{\theta}$, although it is a function of the *yj's*, is not measurable on $y_1, y_2, \ldots$ space, and hence in which $E\{y_1, y_2, \ldots \backslash \hat{\theta}\}$ is not $\hat{\theta}$. If this occurred here the limit in (3.5) would be

$$E\{y_1, y_2, \ldots \backslash \hat{\theta}^2\} - E^2\{y_1, \ldots, y_n \backslash \hat{\theta}\}$$

which would be positive, with positive probability. In order to analyze Hypothesis (C) further we shall consider distribution functions as points of a space Δ defining the distance between two distribution functions as usual as the maximum distance between their graphs (filled in with vertical lines at the jumps) measured along lines with slope -1.

The space Δ is then separable, since the distribution functions which are rational valued and constant except for finitely many jumps, which occur only at rational values, are everywhere dense in Δ. The family of distribution functions F (θ, y) is a curve Γ in Δ, and the regularity of F (θ, y) in θ can be described in terms of the regularity properties of this curve. For each θ there is a single point of this curve, and (Hypothesis (B)) distinct values of θ correspond to distinct points of the curve. The transformation $\theta \longleftrightarrow$ *point of* Γ can be described by the function F (θ) which for each θ is the point of Γ corresponding to θ, so that F (θ) is simply the function F (θ, y) from a different point of view. We now prove that (C) is implied by (A) and (B); the latter hypotheses are assumed true throughout the following discussion.

a. F (θ) is a Baire function of θ. In fact since Δ is separable, with the step functions described above dense in ΔF (θ) is a Baire function if for each point G of Δ which is a step function increasing only at a finite number of points, and for each positive ρ, the values of θ for which the Δ distance from F (θ) to G *is* $\leqslant \rho$ form a Borel set. This fact is an immediate consequence of Hypothesis (A).

(*b*) *F $(\hat{\theta})$ is a measurable function of the yj's.* In fact if $\nu_n(y)/n$ is defined as in (3.2), it is a distribution function, and therefore is a point of Δ depending on the *y' s;* in other words $\nu_n(y)/n$ defines a function $G_n(y_1, \ldots, y_n)$ taking on values in Δ. In terms of Δ distance

$$\lim_{n\to\infty} G_n(y_1, \ldots, y_n) = F(\hat{\theta}) \tag{3.8}$$

with probability 1; this is merely (3.2) in a different form. Hence it is sufficient to prove that $G_n(y_1, \ldots, y_n$ is measurable, and for this, since Δ is separable, it is sufficient to prove that the $y_1, \ldots, y'_n s$ for which the distance from G_n to any given point G of Δ is $\leqslant \rho$ is measurable, for every ρ. It is even sufficient to prove this when G is a step function with a finite number of jumps. In this case the statement is obvious.

(*c*) *To a Borel θ set corresponds a Borel set on* Γ. This follows from the fact that a Baire function [F (θ) in the present case] which has a single valued inverse takes Borel sets into Borel sets.

(*d*) *(A) and (B) together imply* (θ). In fact the inequality $\hat{\theta} < k$ defines a Borel set on Γ, by (*c*), which in turn defines a measurable $y_1 y_2, \ldots$ set by (*b*). Hence θ is measurable as a function of the $y_j' s$, as was to be proved.

Thus the only hypotheses necessary to ensure the truth of the inverse probability theorems discussed in this section are (A) and (B) (except for the existence of moments of the θ distribution needed in (3.4) and (3.5) as already noted. The condition (B) is necessary and (A) is not far from being necessary. Note that these theorems are «probability 1 theorems». The estimate of the value of θ drawn may not be good for a θ set of probability 0. This distinguishes the present discussion from that of von Mises who solved the problem in the Bernoulli case for individual θ values, with necessarily stronger hypothèse [4]. Finally we note that although the problem was stated in terms of density functions, this formulation only served to simplify the notation; the restriction is entirely unnecessary.

It is interesting to consider the special case when θ varies through a finite or denumerable set, assigning positive probability to each θ value. In this case Hypothesis (A) is automatically satisfied, and we find that [assuming only Hypothesis (B)] θ can always be estimated accurately no matter how F (θ, y) varies with θ. As a striking special case, suppose that θ varies through all positive integers, and that F $(1, y) = \lim\limits_{n\to\infty}$ F (n, y) for all y. In this case one might expect trouble in estimating $\hat{\theta}$ when $\theta = 1$, but the discussion shows that this is not true. This is of course to be taken as a theoretical statement. Pratically the situation would give considerable trouble.

4. Discussion,

In answer to questions asked by Dr. Rao, the following additional points are noted.

(i) The estimation problem was stated in § 3 as a problem in inverse probabilities. The results can be interpreted however in a way which makes unnecessary the hypothesis of an *a priori* distribution. We apply this only to one case, suggested by (3.4). Let $f(\theta, y)$ be as in § 3, and for some $\theta = \theta_0$ let $y_1, y_2, \ldots$ be mutually independent, with the distribution determined by $f(\theta_0, y)$. The problem is to determine the unknown θ_0 from the sample $y_1, y_2, \ldots$ Here θ_0 is *not* a random variable. We assume hypotheses (A) and (B) on $f(\theta, y)$. Let $f(\theta)$ be a density function in θ and suppose that $\int |\theta| f(\theta)\, d\theta < \infty$. We do not use $f(\theta)$ to make θ a random variable, but merely use it as a weighting function to define [cf. (3.6)]

$$\theta_n(y_1, \cdots, y_n) = \frac{\int \alpha f(\alpha) \prod\limits_1^n f(\alpha, y_j)\, d\alpha}{\int f(\varphi) \prod\limits_1^n f(\varphi, y_j)\, d\varphi}. \tag{4.1}$$

This function of $y_1, \ldots, y_n$ is to be taken as an estimate of θ_0, the «true value». Now, according to § 3 if $f(\theta)$ were used to define θ probabilities θ_n and θ would be random variables and $\lim\limits_{n\to\infty} \theta_n = \theta$ with probability 1 in $\theta, y_1, y_1, \ldots$ sample space. Hence for fixed θ (neglecting possibly a θ set of θ probability 0) $\lim\limits_{n\to\infty} \theta_n = \theta$ with probability 1 in $y_1, y_2, \ldots$ sample space. Dropping the probability interpretation of $f(\theta)$ we can now finally say that if any weight function $f(\theta)$ is chosen (4.1) defines an estimate of θ_0, the true value, with the following consistency property. Defining $y_1, y_2, \ldots$ probabilities by $f(\theta_0, y)$, $\lim\limits_{n\to\infty} \theta_n = \theta_0$ with probability 1 for each value of θ_0 except possibly for a set of values over which the integral of f vanishes. In other words $\{\theta_n\}$ is a consistent statistic except possibly for a θ_0 set over which the integral of $f(\theta)$ vanishes. This exceptional set has Lebesgue measure 0 if $f(\theta)$ is always positive, but may depend on the choice of f.

(ii) The estimation problem was stated in § 3 in parametric form; θ varied in a Borel linear set. The parametric form is unnecessary. It can be supposed that all possible distributions are considered, that is θ is now supposed to vary over all points of Δ, the space of distribution functions. An *a priori* distribution of θ is then a probability distribution over the Borel sets of Δ. Hypotheses (A) and (B) are now satisfied automatically, and the proof that (C) is also true θ isalmost immediate, since the only delicate point was the proof of (*c*) which is now true by definition; the fact that a Borel θ set corresponds to a Borel Δ set is now true because θ space is Δ. The arguments around this point in § 3 were really aimed at proving that $f(\theta)$ defined a probability measure of Borel Δ sets, which we are now making a hypothesis in the non-parametric form. The results of § 3 all go over, with the obvious modifications due to the fact

[4] Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik, Leipzig 1931, p. 188-192.

that θ is now not numerically valued. In conclusion we remark that the arguments of § 3 are valid whenever θ varies on a Borel set of a complete metric space. The difference between parametric and nonparametric estimation is rather subtle at this level. If θ space is a line, or an n-dimensional Euclidean space, the problem would be called parametric. If θ space is all of Δ the problem would presumably be called nonparametric, at least if the θ probability measure did not confine θ with probability 1 to the homeomorphic image of a line or of n space, in which case we should be again in the parametric problem just described.