PROBABILITY AND STATISTICS[*]

BY

J. L. DOOB[†]

The theory of probability has made much progress recently in the direction of completely mathematical formulations of its methods and results.[‡] The purpose of this paper is to make a further contribution in this direction. In order to analyze the results of repeated trials of an experiment, a certain space of infinitely many dimensions is the proper tool. This space is discussed in the first section of the paper. In the second section, the results of the first are applied to obtain for the first time a complete proof of the validity of the method of maximum likelihood of R. A. Fisher, which is used in statistics to estimate the true probability distribution when the results of a repeated experiment are known.

1. The space $\Omega(F)$

It will be seen that the space $\Omega(F)$ described below provides the natural basis for the analysis of experiments with repeated trials. The preliminary facts, which are not new, will be stated in the form of a theorem.

Theorem 1. *Let $F(x)$ be a monotone non-decreasing function, defined for $-\infty < x < \infty$, and satisfying*

$$F(x-0) = F(x), \quad \lim_{x\to\infty} F(x) = 1, \quad \lim_{x\to-\infty} F(x) = 0. \tag{1}$$

There is a σ-field[§] of point sets on the x-axis, including all Borel measurable sets, and a completely additive non-negative set function $p_F(A)$ defined on this σ-field, such that if I is any interval $a \leqq x < b$, $p_F(I) = F(b) - F(a)$.[||]

[*] Presented to the Society, March 31, 1934; received by the editors April 11, 1934.

[†] National Research Fellow.

[‡] Cf. the treatment of A. Kolmogoroff, Ergebnisse der Mathematik, vol. 2, No. 3: *Grundbegriffe der Wahrscheinlichkeitsrechnung.*

[§] A field is a collection of point sets with the property that if A and B are sets in the collection, $A+B$, $A-A\cdot B$, $A\cdot B$ are also. A field is a σ-field if whenever $A_1, A_2, \cdots$ is a sequence of sets in the field, $\sum_{j=1}^{\infty} A_j$ is also in the field. It will then follow that $\prod_{j=1}^{\infty} A_j$ is in the field. A set function $p(A)$ defined on the sets of a σ-field is completely additive if when $A_1, A_2, \cdots$ is a sequence of disjunct sets in the field, $p\,(\sum_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} p(A_j)$.

[||] The sets in the field of definition of p_F will be called measurable with respect to $F(x)$. If A is measurable with respect to $F(x)$, $p_F(A)$ is the variation of $F(x)$ over A. The definitions of functions measurable with respect to $F(x)$ and of their integration are formulated in the usual way, giving the Lebesgue-Stieltjes integral.

Let $\Omega(F)$ be the space whose points are the sequences $(\cdots, x_{-1}, x_0, x_1, \cdots)$, where x_j is any real number. There is a σ-field of point sets of $\Omega(F)$, including all sets determined by conditions of the form

$$x_j \epsilon E_j \qquad (j = 0, \pm 1, \cdots), \tag{2}$$

where the point sets $E_1, E_2, \cdots$ are measurable with respect to $F(x)$ and a completely additive non-negative set function $P_F(\Lambda)$ defined on this field, such that if Λ is of the type (2),

$$P_F(\Lambda) = \prod_{j=-\infty}^{j=\infty} p_F(E_j). \tag{3}$$

The sets in the field of definition of P_F will be called measurable with respect to $F(x)$; the measurability (with respect to $F(x)$) and integration of functions defined on $\Omega(F)$ are then defined in the usual way. This space was first discussed by Daniell.*

It should be noted that if $\phi(\omega)$ is a measurable function on $\Omega(F)$, and if $\phi(\omega)$ depends only on x_1: $\phi(\omega) = f(x_1)$, $f(x)$ is measurable with respect to $F(x)$, and

$$\int_{\Omega(F)} \phi(\omega) d\omega = \int_{-\infty}^{\infty} f(x) dF(x) \tag{4}$$

where the existence of either integral implies that of the other.

This space $\Omega(F)$ is introduced as a tool in the rigorous analysis of certain ideas in the theory of probability. Let $F(x)$ determine a probability distribution, i.e. we suppose that there is a chance variable $\boldsymbol{x}$ such that the probability that $\boldsymbol{x} < x$ is $F(x)$. Then (1) is satisfied. If a single trial is made, $p_F(A)$ is the probability that the value of x obtained will be in the set A. If a finite succession of trials is made, obtaining values $\xi_1, \cdots, \xi_n$, and if Λ is a point set of $\Omega(F)$ on which P_F is defined, $P_F(\Lambda)$ is the probability that there is a point ω: $(\cdots, x_0, \cdots)$ in Λ such that $x_j = \xi_j$, $j = 1, \cdots, n$. The usual interpretation if Λ is a set of the form (2) is obvious. The advantage of this point of view† is that the set-up is independent of the number of trials. Chance varia-

* Annals of Mathematics, (2), vol. 20 (1919), pp. 281–288. Daniell actually only considered the space whose points are sequences of the form $(x_1, x_2, \cdots)$, but the treatment of $\Omega(F)$ could be carried through in the same way. These considerations concerning the space $\Omega(F)$ can be considered as a particular case of a general treatment given by Kolmogoroff, loc. cit., pp. 24–30.

† A similar point of view was taken by A. Khintchine, Zeitschrift für angewandte Mathematik und Mechanik, vol. 13 (1933), pp. 101–103, who treated the case of a chance variable which only takes on the values 1 or 0 (making less restrictions on $P_F(\Lambda)$ however). This space was used for the same purpose by E. Hopf, Journal of Mathematics and Physics of the Massachusetts Institute of Technology, vol. 13 (1934), pp. 51–102. The place of these methods in the theory of stochastic processes was discussed by the writer in the Proceedings of the National Academy of Sciences, vol. 20 (1934), pp. 376–379.

bles become measurable functions on $\Omega(F)$, and their integrals on $\Omega(F)$ are their expectations. The law of large numbers will be seen to correspond to the ergodic theorem of Birkhoff.* The convergence of a sequence of chance variables in probability† is simply convergence in measure on $\Omega(F)$.‡

THEOREM 2. *The transformation T of $\Omega(F)$ into itself,*

$$T: \qquad x_j' = x_{j+1} \qquad (j = 0, \pm 1, \cdots),$$

is a one-to-one measure-preserving transformation. If Λ is a measurable set invariant under T, $P_F(\Lambda) = 0$, or $P_F(\Lambda) = 1$.§ If $\phi(\omega)$ is any measurable function on $\Omega(F)$ such that $\int_{\Omega(F)} |\phi(\omega)|^2 d\omega$ exists and such that $\phi(T\omega) = e^{i\lambda}\phi(\omega)$ for some real number λ,

$$\phi(\omega) = \int_{\Omega(F)} \phi(\omega) d\omega$$

almost everywhere on $\Omega(F)$.

The second part of the theorem includes the first part if $\lambda = 0$, and if $\phi(\omega)$ is considered as the characteristic function of a point set, so only the second part of the theorem need be considered. The proof will be given in several steps.

(i) Let $F(x)$ be 0 for $x<0$, x for $0 \leqq x \leqq 1$ and 1 for $x>1$, and let $p_F(A)$ and $P_F(\Lambda)$ for this $F(x)$ be denoted by $p_0(A)$, $P_0(\Lambda)$, respectively. Let Ω_0 be the subset of $\Omega(F)$ consisting of the points $(\cdots, x_{-1}, x_0, x_1, \cdots)$ whose coordinates satisfy the inequalities $0 \leqq x_j \leqq 1$, $j = 0, \pm 1, \cdots$. It will be shown that the general set functions $p_F(A)$ and $P_F(\Lambda)$ can be derived from $p_0(A)$ and $P_0(\Lambda)$. In fact, let $y = F(x)$ transform the points of the x-axis into points of the interval $0 \leqq y \leqq 1$, where if $F(x)$ has a jump at x_0, the point x_0 will be made to correspond to the interval $F(x_0) \leqq y \leqq F(x+0)$. Then $p_F(A)$ is defined for those and only those sets whose images on the y-axis are Lebesgue measurable, and for such sets $p_F(A)$ is defined as the Lebesgue measure of the image of A. In the same way the set Λ_x on $\Omega(F)$ measurable with respect to $F(x)$ goes over into a set Λ_y on Ω_0 on which $P_0(\Lambda)$ is defined, and $P_F(\Lambda_x)$

* Cf. A. Khintchine, loc. cit., and E. Hopf, loc. cit., p. 95.

† For the definition of convergence in probability, see for instance Kolmogoroff, loc. cit., p. 31.

‡ Convergence in measure was defined and discussed by F. Riesz, Paris Comptes Rendus, vol. 148 (1909), pp. 1303–1305.

§ If $F(x)$ does not increase, except for equal jumps at $x = 0, \cdots, 9$, the set function $P_F(\Lambda)$ has a simple interpretation as ordinary two-dimensional Lebesgue measure, and this property (metrical transitivity) was proved by W. Seidel, Proceedings of the National Academy of Sciences, vol. 19 (1933), pp. 453–456. Hopf obtained this result from the second part of the corollary to this theorem (see below) by a different method.

$=P_0(\Lambda_\nu)$. Then it is sufficient to prove Theorem 2 for the space Ω_0 and the set function $P_0(\Lambda)$.

(ii) The set of all complex-valued functions $\phi(\omega)$ on Ω_0 whose real and imaginary parts are measurable on Ω_0 and such that

$$\int_{\Omega_0} |\phi(\omega)|^2 d\omega$$

exists can be considered as the set of elements of a Hilbert space* $\mathfrak{H}$ if the inner product of $\phi_1(\omega)$, $\phi_2(\omega)$ is defined in the usual way as

$$\int_{\Omega_0} \phi_1(\omega)\overline{\phi_2(\omega)}d\omega.†$$

It is easily seen that the set of functions of the form

$$\exp\left\{2\pi i \sum_{j=1}^{n} n_j x_j(\omega)\right\}$$

where $x_j(\omega)$ is the value of x_j for the point $\omega:(\cdots, x_{-1}, x_0, x_1, \cdots)$ and where n_j, n are arbitrary integers, form a complete orthonormal set of functions in $\mathfrak{H}$.‡ If these functions, arranged in some order, are $\phi_0(\omega)$, $\phi_1(\omega)$, $\cdots$ where $\phi_0(\omega)\equiv 1$, to every function $\phi(\omega)$ in $\mathfrak{H}$ corresponds a series $\sum_{j=0}^{\infty} a_j\phi_j(\omega)$, where the coefficient a_j is determined by

$$a_j = \int_{\Omega} \phi(\omega)\overline{\phi_j(\omega)}d\omega, \tag{5}$$

such that

$$\int_{\Omega_0} |\phi(\omega)|^2 d\omega = \sum_{j=0}^{\infty} |a_j|^2. \tag{6}$$

(iii) Now suppose that $\phi(T\omega)=e^{i\lambda}\phi(\omega)$. Then if b_0, b_1, $\cdots$ are the coefficients corresponding to $\phi(T\omega)$,

$$b_j = e^{i\lambda}a_j, \tag{7}$$

and, from the simple form of the transformation T, if $j>0$,

$$a_j = b_{\tau(j)} = e^{i\lambda}a_{\tau(j)} \tag{8}$$

* For a general reference to Hilbert space see, for instance, M. H. Stone, *Linear Transformations in Hilbert Space*, American Mathematical Society Colloquium Publications, vol. 15 (especially chapter I). The properties of Ω_0 which are needed here (separability, etc., if distance is properly defined), are given by Daniell, loc. cit., p. 281. Using these properties the proof that the functions $\{\phi(\omega)\}$ form a Hilbert space follows the lines of a similar theorem in Stone, pp. 23–29.

† If ξ is a complex number, $\bar{\xi}$ will denote its conjugate.

‡ This concept is discussed by Stone, loc. cit., pp. 7–14, where the facts stated below are proved.

where $\tau(j)\neq j$. Repeating this we find a sequence of coefficients $a_{m_1}, a_{m_2}, \cdots$, where $m_1=j$, $m_j=\tau(m_{j-1})$ if $j>1$, whose absolute values are all equal. Evidently $m_i\neq m_j$ if $i\neq j$. This contradicts (6) unless $a_j=0$. Then $a_j=0$ if $j>0$, and $\phi(\omega)=a_0$, as was to be proved.[*]

COROLLARY. (i) *If $\phi(\omega)$ is any integrable function on $\Omega(F)$,*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\phi(T^j\omega) = \int_{\Omega(F)}\phi(\omega)d\omega \tag{9}$$

almost everywhere on $\Omega(F)$.

(ii) *If $\phi_1(\omega)$, $\phi_2(\omega)$ are measurable functions the squares of whose absolute values are integrable on $\Delta(F)$,*

$$\lim_{n\to\infty}\int_{\Omega(F)}\phi_1(T^n\omega)\phi_2(\omega)d\omega = \left\{\int_{\Omega(F)}\phi_1(\omega)d\omega\right\}\left\{\int_{\Omega(F)}\phi_2(\omega)d\omega\right\}. \tag{10}$$

(i) This part of the corollary is simply the ergodic theorem in this case.[†]

(ii) This part of the corollary corresponds to the extension of the ergodic theorem given by E. Hopf, B. O. Koopman and J. von Neumann, to the particular case where there are no "angle variables."[‡] It is obvious when $\phi_1(\omega)$ and $\phi_2(\omega)$ each depend only on a finite number of the coordinates of ω: $(\cdots, x_0, \cdots)$ since in that case the terms in (10) are equal to the limit prescribed for sufficiently large values of n. Since any measurable function can be approximated by functions depending only on a finite number of coordinates,[§] the general theorem can be reduced to this case.

The following lemma is needed for the proof of the next theorem.

LEMMA. *Let $F(x)$ be defined as in Theorem 1. Define measure on the x-axis by the set function p_F. Let $f(x)$ be a function defined for almost all values of x and measurable (with respect to $F(x)$). Then if*

$$\limsup_{n\to\infty}|f(x_n)|/n<\infty \tag{11}$$

on a set of points ω: $(\cdots, x_0, \cdots)$ of $\Omega(F)$ of positive measure, $\int_{-\infty}^{\infty}f(x)\,dF(x)$ exists (as a Stieltjes-Lebesgue integral).[||]

[*] Stone, loc. cit., p. 10.

[†] For a simple proof of the ergodic theorem, following the lines of the first proof, given by Birkhoff, cf. A. Khintchine, Mathematische Annalen, vol. 107 (1933), pp. 485–488. In this proof the function $\phi(x, r)$ corresponds to the function $\sum_{j=1}^{r}\phi(\tau^j\omega)$ used here.

[‡] E. Hopf, Proceedings of the National Academy of Sciences, vol. 18 (1932), pp. 204–209; B. O. Koopman and J. von Neumann, ibid., pp. 255–263. In these treatments a continuous set of transformations is considered, instead of the set of iterates of a single transformation as here, but the treatment needs no essential change to make it applicable to this case.

[§] Cf. Daniell, loc. cit., p. 283.

[||] The Stieltjes-Lebesgue integral is defined in the same was as the ordinary Lebesgue integral except that p_F-measure is used instead of ordinary Lebesgue measure.

By hypothesis there is a positive number M such that

$$\limsup_{n\to\infty} \frac{|f(x_n)|}{n} \leqq M \tag{12}$$

on a set of points Λ of $\Omega(F)$, $P_F(\Lambda)>0$. Let Λ_N be the point set on $\Omega(F)$ at which

$$\underset{n\geqq N}{\text{L.U.B.}} \left\{\frac{f(x_n)}{n}\right\} > M.^*$$

Then $\Lambda_N \supset \Lambda_{N+1}$ and

$$\lim_{N\to\infty} P_F(\Lambda_N) = 1 - P_F(\Lambda) < 1.$$

Let E_n be the set of values of x at which $f(x)>nM$. Then a point $\omega:(\cdots, x_0, \cdots)$ belongs to the complement of Λ_N if and only if x_n is in the complement of E_n for $n\geqq N$. Then the complement of Λ_N is of the form (2), so from (3),

$$P_F(\Lambda_N) = 1 - \sum_{n=N}^{\infty}[1 - p_F(E_n)]. \tag{13}$$

Since $\lim_{N\to\infty} P_F(\Lambda_N)<1$, the infinite product is convergent. Then $\sum_{n=0}^{\infty} p_F(E_n)$ must be convergent,† and it is easily shown from the definition of the Lebesgue-Stieltjes integral that this implies that $f(x)$ is integrable (with respect to $F(x)$) over the set E_0. Substituting $-f(x)$ for $f(x)$, the proof shows that $f(x)$ is also integrable (with respect to $F(x)$) over the set where it is negative. Then $\int_{-\infty}^{\infty} f(x)dF(x)$ exists, as was to be proved.

The following theorem will be put in the phraseology of the theory of probability. Like the lemma, it is simply a theorem on integration on $\Omega(F)$.

THEOREM 3. *Let $x_1, x_2, \cdots$ be a sequence of independent chance variables with the same distributions.*

(i) *If the expectation E of x_j exists, then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} x_j = E \tag{14}$$

with probability 1.

(ii) *If there is a sequence of real numbers $c_1, c_2, \cdots$ such that the probability is positive that*

* Throughout this paper, if $a_1, a_2, \cdots$ is a sequence of real numbers, L.U.B. $\{a_n\}$ will denote its least upper bound.

† W. F. Osgood, *Lehrbuch der Funktionentheorie*, vol. 1, 4th edition, p. 528.

$$(15)\qquad \limsup_{n\to\infty}\left|\frac{1}{n}\sum_{j=1}^{n}x_j - c_n\right| < \infty,$$

it follows that the expectation of x_j exists, and we have Case (i) *again.**

Let $F(x)$ be the probability that $x_j<x$. If the expectation of x_j exists, it is, by (4),

$$\int_{\Omega(F)} x_j(\omega)d\omega = \int_{-\infty}^{\infty} xdF(x)$$

where ω is the point $(\cdots, x_0, \cdots)$.

(i) The first part of the theorem is simply the Corollary of Theorem 2 applied to the function $\phi(\omega)=x_0(\omega)$.

(ii) We can suppose in (ii) that there is a point set Λ on $\Omega(F)$ of positive P_F-measure, a positive number M and an integer N such that

$$(16)\qquad \left|\frac{1}{n}\sum_{j=1}^{n}x_j(\omega) - c_n\right| < M$$

on Λ if $n\geqq N$. On replacing n by $n-1$ and multiplying by $(n-1)/n$,

$$(17)\qquad \left|\frac{1}{n}\sum_{j=1}^{n-1}x_j(\omega) - \frac{n-1}{n}c_{n-1}\right| < M$$

on Λ if $n\geqq N+1$. Subtracting (17) from (16),

$$(18)\qquad \left|\frac{x_n(\omega)}{n} - \left(c_n - \frac{n-1}{n}c_{n-1}\right)\right| < 2M$$

on Λ if $n\geqq N+1$. By (1), $x_n(\omega)/n$ approaches 0 in measure as n becomes infinite. Then there is an integer $N_1\geqq N+1$ such that on a subset Λ_n of Λ of positive P_F-measure

$$|x_n(\omega)/n| < M \text{ if } n \geqq N_1.$$

Hence

$$(19)\qquad |c_n - c_{n-1}(n-1)/n| < 3M.$$

From (18) and (19),

$$|x_n(\omega)/n| < 5M$$

on Λ. The lemma can now be applied, and it shows that $\int_{-\infty}^{\infty}x\,dF(x)$ exists as a Stieltjes-Lebesgue integral. This integral is the expectation of the chance variable x_j.

* A. Kolmogoroff, Ergebnisse der Mathematik, vol. 2, No. 3: *Grundbegriffe der Wahrscheinlichkeitsrechnung*, p. 59, announced the first part of this theorem, and also the second part, under the assumption that the probability is 1 that the upper limit in (15) is 0.

The following theorem will be needed in the application of the results of this section. Its proof is simple and will be omitted.

THEOREM 4. *If $F(x)$ is defined as in Theorem* 1, *and if $F(x)$ has an integrable derivative $f(x)$:*

$$F(x) = \int_{-\infty}^{x} f(x)dx,$$

there is a point set $\Lambda(F)$ on $\Omega(F)$, $P_F[\Lambda(F)]=1$, with the following property. If $g(x)$ is any function defined and continuous almost everywhere (*in the sense of p_F-measure*) *on the infinite interval $-\infty<x<\infty$, and such that $\int_{-\infty}^{\infty}g(x)f(x)dx$ exists, then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} g(x_j) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

at every point $\omega:(\cdots, x_0, \cdots)$ of $\Lambda(F)$.[*]

2. THE METHOD OF MAXIMUM LIKELIHOOD

For each value of p in some point set E let $f(x, p)$ be a probability density over the interval $-\infty<x<\infty$.[†] Assume that the chance variable $\boldsymbol{x}$ has a probability distribution whose density is $f(x, p)$ for some (unknown) value of p in E. Then an important problem in statistics is that of estimating the true value of p by means of large samples of values of $\boldsymbol{x}$, obtained independently. This is done by the method of maximum likelihood of R. A. Fisher[‡], which has supplanted the use of Bayes' theorem. If $x_1, \cdots, x_n$ is a sample of values of $\boldsymbol{x}$, and if $f(x, p)$ is the probability density of the distribution of values of $\boldsymbol{x}$, the probability of obtaining a sample of values $x_1', \cdots, x_n'$ where x_j' is in a small interval with midpoint x_j, is, in the limit, proportional to $\prod_{j=1}^{n} f(x_j, p)$. The method of maximum likelihood takes as an approximation to p_0, the true value of p, the value p_n of p (or one of them if there are several) which makes this product a maximum. If p_n approaches p_0 in probability as the samples become larger, p_n is called a consistent estimate of p. A

[*] The theorem will be needed as here stated. It can be stated in terms of Riemann-Stieltjes integration, making unnecessary any restrictions on $F(x)$.

[†] This means that $f(x, p)\geqq 0$, that $f(x, p)$ is defined for almost all values of x, is measurable and integrable over the x-axis, and that $\int_{-\infty}^{\infty}f(x)dx=1$. It is supposed that there is a chance variable $\boldsymbol{x}(p)$ whose values are distributed in such a way that the probability of $\boldsymbol{x}(p)$ being in any measurable point set A is $\int_A f(x)dx$.

[‡] Philosophical Transactions of the Royal Society of London, (A), vol. 222, pp. 309–368, especially pp. 309–330. The proofs given by Fisher and by H. Hotelling, these Transactions, vol. 32 (1930), pp. 847–859, of the validity of the method of maximum likelihood (in the sense that theorems similar to the ones to be proved in this section hold) are not rigorous.

rigorous proof will be given in this section that, under certain hypotheses, the method of maximum likelihood furnishes consistent estimates.

THEOREM 5. *For each value of p in a point set E let $f(x, p)$ be a probability density on the infinite interval $-\infty < x < \infty$. Let x be a chance variable whose distribution is determined by the probability density $f(x)$, and suppose that for each set of numbers $x_1, \cdots, x_n$, $n=1, 2, \cdots$, it is possible to find a value of p in E: $p=p_n(x_1, \cdots, x_n)$ such that*

$$\prod_{j=1}^{n} f(x_j, p_n) \geqq \prod_{j=1}^{n} f(x_j). \tag{20}$$

Then if

$$F(x) = \int_{-\infty}^{x} f(x)dx,$$

there is a set of points Λ of $\Omega(F)$ of total probability 1: $P_F(\Lambda)=1$, *with the following properties. Let ω: $(\cdots, x_0, \cdots,)$ be a point of Λ and let $\{p_{\alpha_n}(x_1, \cdots, x_{\alpha_n})\}$ be any subsequence of $\{p_n(x_1, \cdots, x_n)\}$ for $x_j=x_j(\omega)$, $j=1, 2, \cdots$. Set*

$$f_n(x) = \underset{m \geqq n}{\text{L.U.B.}} \; \{f(x, p_{\alpha_m})\}. \tag{21}$$

Suppose that $f_n(x)/f(x)$ is continuous, except possibly for a set of values of x of zero probability[*], *and that*

$$\int_{-\infty}^{\infty} f(x) \log^{+} \left[\frac{f_n(x)}{f(x)} \right] dx \tag{22}$$ [†]

exists. It follows

(i) *that the integral*

$$\int_{-\infty}^{\infty} f(x) \log \left[\frac{\limsup_{n \to \infty} f(x, \; p_{\alpha_n})}{f(x)} \right] dx \tag{23}$$

exists and is not negative;

(ii) *that if* $\limsup_{n\to\infty} f(x, p_{\alpha_n})$ *is integrable, and if*

$$\int_{-\infty}^{\infty} \limsup_{n \to \infty} f(x, p_{\alpha_n}) dx \leqq 1, \tag{24}$$

then $\limsup_{n\to\infty} f(x, p_{\alpha_n}) = f(x)$ *except possibly on a set of zero probability*;

(iii) *that if the sequence $\{f(x, p_{\alpha_n})\}$ converges (except possibly on a set of* 0 *probability), the limit function is $f(x)$ (except possibly on a set of* 0 *probability).*

[*] This means that the integral of $f(x)$ over the exceptional set is 0, i.e., that $f(x)=0$ almost everywhere (in the sense of Lebesgue measure) on the set. In the following integrals, in which ratios with $f(x)$ in the denominator appear, we define the ratios as 1 when $f(x)=0$.

[†] If $\xi \geqq 0$, $\log^{+} \xi$ is defined as $\log \xi$ when $\xi > 1$, and 0 otherwise.

In the application to statistical problems, it is part (iii) which would be customarily used. Thus, consider the problem of estimating the mean of a normal distribution, where the density is

$$f(x,\,p) = \frac{1}{(2\pi)^{1/2}}e^{-(x-p)^2/2}, \tag{25}$$

the true value of p being p_0. In this case if p_{a_n} approaches any finite value, it is seen at once that (22) exists. Since $f(x,\,p)$ is continuous in p, (iii) shows that $f(x,\,p_{a_n})$ approaches $f(x,\,p_0)$, so that p_{a_n} converges to p_0, the true value. On the other hand, suppose that p_{a_n} converges to either $+\infty$ or $-\infty$. Then the integral (22) exists. By (iii), $f(x,\,p_{a_n})$, which converges to 0 (since $|p_{a_n}|\rightarrow\infty$), approaches $f(x,\,p_0)$. This is impossible, so $\lim_{n\rightarrow\infty}p_n=p_0$, with probability 1. It is usual to take for the approximation p_n the average $(1/n)\sum_{j=1}^{n}x_j$.

It is evident that if Λ_0 is the set of points $\omega:(\cdots,\,x_0,\,\cdots)$ of $\Omega(F)$ such that at least one coordinate x_j is in the set of values of x at which $f(x)=0$, $P_F(\Lambda_0)=0$. It will be shown that the set Λ of this theorem can be taken as the set $\Lambda(F)-\Lambda_0\cdot\Lambda(F)$, where $\Lambda(F)$ was described in Theorem 4. Suppose then that $\omega:(\cdots,\,x_0,\,\cdots)$ is in this set.

(i) From (20) and (21), if $L_t(y)$ is defined for every positive number t as $\log y$ if $y\geqq t$, and as $\log t$ if $y<t$,

$$\frac{1}{a_n}\sum_{j=1}^{a_n}L_t\left[\frac{f_N(x_j)}{f(x_j)}\right] \geqq \frac{1}{a_n}\sum_{j=1}^{a_n}\log\left[\frac{f_N(x_j)}{f(x_j)}\right] \geqq \frac{1}{a_n}\sum_{j=1}^{a_n}\log\left[\frac{f(x_j,\,p_{a_n})}{f(x_j)}\right] \geqq 0, \tag{26}$$

if $n\geqq N$. Now since $f\log^+(f_N/f)$ is integrable, $fL_t(f_N/f)$ is integrable (over the entire x-axis). Then letting n become infinite in (26), we have, from Theorem 4,

$$\int_{-\infty}^{\infty}f(x)L_t\left[\frac{f_N(x)}{f(x)}\right]dx \geqq 0. \tag{27}$$

As N increases, $L_t(f_N/f)$ does not increase, and

$$\lim_{n\rightarrow\infty}L_t(f_N/f) = L_t(\hat{f}/f),$$

where

$$\hat{f}(x) = \limsup_{n\rightarrow\infty}f(x,\,p_{a_n}).$$

Then we can go to the limit under the integral sign in (27)[*], obtaining

$$\int_{-\infty}^{\infty} f(x)L_t\left[\frac{\hat{f}(x)}{f(x)}\right]dx \geqq 0. \tag{28}$$

Let E_t be the set of values of x at which $f(x) \leqq t$. The integral (28) can be separated into integrals over E_t and its complement, CE_t. Doing this, we find that

$$0 \leqq p_F(E_t) \log \frac{1}{t} \leqq \int_{CE_t} f(x)L_t\left[\frac{\hat{f}(x)}{f(x)}\right]dx \text{ if } t \leqq 1. \tag{29}$$

Letting t approach 0, (29) shows that $p_F(E_0)=0$ and that furthermore the integral (23) exists and is not negative.

(ii) From (i),

$$0 \leqq \int_{-\infty}^{\infty} f(x) \log\left[\frac{\hat{f}(x)}{f(x)}\right]dx. \tag{30}$$

Now by a well known inequality[†], and using (24),

$$\int_{-\infty}^{\infty} f(x) \log\left[\frac{\hat{f}(x)}{f(x)}\right]dx \leqq \log \int_{-\infty}^{\infty} \hat{f}(x)dx \leqq 0. \tag{31}$$

There is equality in (31) only when $\hat{f}(x)=f(x)$ for almost all x (in the sense of p_F-measure), and there is necessarily equality, by (30), so (ii) is proved.

(iii) To prove (iii) it is only necessary to reduce it to (ii), by showing that, if

$$\hat{f}(x) = \lim_{n\to\infty} f(x, p_{a_n}),$$

$\int_{-\infty}^{\infty}\hat{f}(x)dx$ exists and is not greater than 1. We have

$$\int_{-\infty}^{\infty} f(x, p_{a_n})dx = 1,$$

so by Fatou's lemma[‡], $\hat{f}(x)$ is integrable over $-\infty < x < \infty$ and $\int_{-\infty}^{\infty}\hat{f}(x)dx \leqq 1$.

[*] The situation is visualized more readily when the integral is written as

$$\int_{-\infty}^{\infty} L_t\left[\frac{f_N(x)}{f(x)}\right]dF(x).$$

The integrand is bounded uniformly above by the integrable function $L_t(f_1/f)$ and below by $\log t$, so we can integrate term by term.

[†] Making the substitution $y=F(x)$, the inequality needed becomes

$$\int_0^1 \log g(y)dy \leqq \log \int_0^1 g(y)dy,$$

where $g(y)=\hat{f}(x)/f(x)$.

[‡] P. Fatou, Acta Mathematica, vol. 30 (1906), pp. 375–376.

The treatment of the principle of maximum likelihood given above was for continuous distributions. The most general statement of the other extreme is as follows. To each integer $n \geqq 1$ is assigned a probability $a(n, p)$ depending on p which varies on some point set. The intrinsic conditions are

$$a(n, p) \geqq 0, \quad \sum_{n=1}^{\infty} a(n, p) = 1.$$

For each sample of integers $r_1, \cdots, r_n$ there is a value p_n of p such that

$$\prod_{j=1}^{n} a(r_j, p_n) \geqq \prod_{j=1}^{n} a(r_j, p_0),$$

where p_0 is the true value of p. The problem is to show (under suitable restrictions on $a(n, p)$), that p_n approaches p_0 in probability. This problem can be treated in a similar manner to the one just treated.

The method of maximum likelihood, when analyzed more carefully, yields further information. Reverting to continuous distributions, suppose that for each value of p in a neighborhood of p_0, $f(x, p)$ is the density of a probability distribution. The function $p_n(x_1, \cdots, x_n)$ will be called an nth approximation of maximum likelihood to p_0 if it is defined on $\Omega(F)$ (where $F(x) = \int_{-\infty}^{x} f(x, p_0)dx$) on a set of P_F-measure 1, if

$$\prod_{j=1}^{n} f(x_j, p_n) \geqq \prod_{j=1}^{n} f(x_j, p_0)$$

and if $\Pi_{j=1}^{n} f(x_j, p)$ for fixed $x_1, \cdots, x_n$ has a relative maximum at $p = p_n$. It is is no restriction to assume that $p_0 = 0$.

THEOREM 6. *For each value of p in some neighborhood $|p| \leqq a_1$, $a_1 > 0$, of $p = 0$, let $f(x, p)$ be a probability density in the infinite interval $-\infty < x < \infty$. Let the true distribution of x be determined by the probability density $f(x, 0)$. Suppose*

(i) *that* $\log f(x, p)$ *can be expressed in the form*

$$\log f(x, p) = \log f(x, 0) + p\alpha(x) + \frac{p^2}{2}\beta(x) + \gamma(x, p),^* \tag{32}$$

where $\alpha(x)f(x, 0)$, $\alpha(x)^2 f(x, 0)$, $\beta(x)f(x, 0)$ are Lebesgue measurable and integrable over $-\infty < x < \infty$ and where

* We shall assume in the discussion of this theorem that x does not take on any value at which $f(x, 0) = 0$. This means leaving out sets of total probability 0 on the x-axis and on $\Omega(F)$, where $F(x) = \int_{-\infty}^{x} f(x, 0)\,dx$.

$$\frac{\partial}{\partial p}\gamma(x, p) = \gamma_p(x, p)$$

exists for $|p| \leqq a_2 \leqq a_1$, $a_2 > 0$, *and is continuous at* $p = 0$;

(ii) *that if*

$$\phi(x) = \underset{0<|p|\leqq a_2}{\text{L.U.B.}} \left\{ \frac{|\gamma_p(x, p)|}{p^2} \right\} \tag{33}$$

then $\phi(x)f(x, 0)$ *is integrable over* $-\infty < x < \infty$[*];

(iii) *that if* $\delta(x, p)$ *is defined by*

$$f(x, p) = f(x, 0)\left\{1 + p\alpha(x) + \frac{p^2}{2}[\beta(x) + \alpha(x)^2] + \delta(x, p)\right\}, \tag{34}$$

$$\lim_{p\to 0} \frac{1}{p^2}\int_{-\infty}^{\infty} \delta(x, p)f(x, 0)dx = 0.\text{[†]} \tag{35}$$

Then

$$\int_{-\infty}^{\infty} \alpha(x)^2 f(x, 0)dx + \int_{-\infty}^{\infty} \beta(x)f(x, 0)dx = 0. \tag{36}$$

Suppose that

$$\sigma^2 = \int_{-\infty}^{\infty} \alpha(x)^2 f(x, 0)dx > 0.$$

Then if $p_n(x_1, \cdots, x_n)$ *is an nth approximation of maximum likelihood to* $p = 0$, *and if* p_n *approaches* 0 *in probability*:

$$\lim_{n\to\infty} \overline{P}_F(|p_n| > \epsilon) = 0\text{[‡]} \tag{37}$$

for every $\epsilon > 0$,

$$\lim_{n\to\infty} \overline{P}_F(\sigma n^{1/2} p_n < \lambda) = \lim_{n\to\infty} \underline{P}_F(\sigma n^{1/2} p_n < \lambda) = \frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\lambda} e^{-x^2/2}dx, \tag{38}$$

for every constant λ, *uniformly in* λ.

[*] We take this to mean that $\int_{-\infty}^{\infty}\phi(x)dF(x)$ exists so that $\phi(x)$ can be $+\infty$ on a set of zero p_F-measure.

[†] Since $f(x, p)$ is integrable over $-\infty < x < \infty$, it follows from (i) that $\delta(x, p)f(x, 0)$ is also.

[‡] Such expressions will be taken to mean the probability that $|p_n| > \epsilon$ (i.e. the P_F-measure of the set of those points on $\Omega(F)$ where $|p_n| > \epsilon$), etc. In (37) we use $\overline{P}_F$, the outer measure on $\Omega(F)$, instead of P_F, since we have not assumed that $p_n(x_1, \cdots, x_n)$ is measurable with respect to $F(x)$. Similarly, $\underline{P}_F$ will denote the inner measure on $\Omega(F)$.

The theorem states simply that, under suitable restrictions on the character of $f(x, p)$ in p, p_n will be normal for large n, with variance $1/(\sigma^2 n)$.*

Since

$$\int_{-\infty}^{\infty} f(x, p)dx = 1 \tag{39}$$

for all p in the neighborhood considered,

$$\begin{aligned} p\int_{-\infty}^{\infty} \alpha(x)f(x, 0)dx + \frac{p^2}{2}\int_{-\infty}^{\infty} [\beta(x) + \alpha(x)^2]f(x, 0)dx \\ + \int_{-\infty}^{\infty} \delta(x, p)f(x, 0)dx = 0. \end{aligned} \tag{40}$$

Dividing through by p and letting p approach 0, we find that, in view of (35),

$$\int_{-\infty}^{\infty} \alpha(x)f(x, 0)dx = 0. \tag{41}$$

Dividing (40) through by p^2 and letting p approach 0, we find in view of (35), that (36) is true.

The logarithm of the likelihood of a value of p, obtained from n trials, is defined as

$$\begin{aligned} L_n(p) = \sum_{j=1}^{n} \log f(x_j, p) = \sum_{j=1}^{n} \log f(x_j, 0) \\ + p\sum_{j=1}^{n}\alpha(x_j) + \frac{p^2}{2}\sum_{j=1}^{n}\beta(x_j) + \sum_{j=1}^{n}\gamma(x_j). \end{aligned} \tag{42}$$

Since $L_n(p)$ has a relative maximum at p_n,

$$L_n'(p_n) = \sum_{j=1}^{n}\alpha(x_j) + p_n\sum_{j=1}^{n}\beta(x_j) + \sum_{j=1}^{n}\gamma_p(x_j, p_n) = 0, \tag{43}$$

if we suppose that $|p_n| < a_2$.

(A) If $p_n = 0$, $\sum_{j=1}^{n}\alpha(x_j) = 0$ also, excluding possibly a set of zero probability on $\Omega(F)$. For if $p_n = 0$, (43) becomes $\sum_{j=1}^{n}\alpha(x_j) = 0$ (if a set of zero probability on $\Omega(F)$ is ignored), since the hypotheses of the theorem imply that $\gamma_p(x, 0) = 0$ on a set of p_F-measure 1 on the x-axis.

(B) Let m be defined by

$$m = \int_{-\infty}^{\infty} \phi(x)f(x, 0)dx. \tag{44}$$

* R. A. Fisher, loc. cit., p. 359.

H. Hotelling, loc. cit., pp. 856–858. Through an oversight, this theorem is stated, on p. 850, with the variance of p_n as $\sigma^2 n$.

Then

$$(45) \qquad P_F\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\phi(x_j) = m\right\} = 1$$

by the Corollary to Theorem 2, or by Theorem 3, so that

$$(46) \qquad \lim_{n\to\infty}\bar{P}_F\left\{\frac{|p_n|}{n}\sum_{j=1}^{n}\phi(x_j) \geqq \epsilon\right\} = 0$$

for any $\epsilon > 0$.*

$$(C) \qquad P_F\left\{\lim_{n\to\infty}\frac{-1}{\sigma^2 n}\sum_{j=1}^{n}\beta(x_j) = 1\right\} = 1,$$

by the Corollary to Theorem 2 or by Theorem 3.

Now from (43), if $0 < |p_n| \leqq a_2$ and if the denominator does not vanish,

$$(47) \qquad n^{1/2}\sigma p_n = \frac{1}{\sigma n^{1/2}}\sum_{j=1}^{n}\alpha(x_j) + R_n,$$

where

$$(48) \qquad R_n = \frac{\dfrac{1}{\sigma n^{1/2}}\displaystyle\sum_{j=1}^{n}\alpha(x_j)\left\{1 + \dfrac{1}{\sigma^2 n}\displaystyle\sum_{j=1}^{n}\beta(x_j) + \dfrac{1}{\sigma^2 n p_n}\displaystyle\sum_{j=1}^{n}\gamma_p(x_j, p_n)\right\}}{-\dfrac{1}{\sigma^2 n}\displaystyle\sum_{j=1}^{n}\beta(x_j) - \dfrac{1}{\sigma^2 n p_n}\displaystyle\sum_{j=1}^{n}\gamma_p(x_j, p_n)}.$$

We define R_n as 0 if $p_n = 0$.

Using (A), (B), (C), we shall show that

$$\lim_{n\to\infty}\bar{P}_F(|R_n| > \epsilon) = 0$$

for every $\epsilon > 0$. Since

$$\left|\frac{1}{np_n}\sum_{j=1}^{n}\gamma_p(x_j, p_n)\right| \leqq \frac{|p_n|}{n}\sum_{j=1}^{n}\phi(x_j),$$

and since by the Laplace-Liapounoff theorem†

* Equation (45) expresses the fact that a certain sequence $\{h_n\}$ of functions on $\Omega(F)$ converges to m almost everywhere on $\Omega(F)$. Then since the sequence $\{|p_n|\}$ converges in measure to 0 on $\Omega(F)$, by hypothesis, the sequence $\{|p_n|h_n\}$ converges in measure to 0 on $\Omega(F)$, which fact is expressed by (46).

† A. Khintchine, Ergebnisse der Mathematik, vol. 2, No. 4: *Asymptotische Gesetze der Wahrscheinlichkeitsrechnung*, pp. 1–8.

$$\lim_{n\to\infty} P_F\left\{\frac{1}{\sigma n^{1/2}}\sum_{j=1}^{n}\alpha(x_j) \leqq \lambda\right\} = \frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{\lambda} e^{-x^2/2}dx, \tag{49}$$

the numerator of R_n converges in measure to 0 and the denominator to 1 as n becomes infinite. Then R_n converges in measure to 0 on $\Omega(F)$ as n becomes infinite:

$$\lim_{n\to\infty} \bar{P}\left\{\left|\sigma n^{1/2}p_n - \frac{1}{\sigma n^{1/2}}\sum_{j=1}^{n}\alpha(x_j)\right| \geqq \epsilon\right\} = 0 \tag{50}$$

for every $\epsilon>0$. Now suppose that $\sigma n^{1/2}p_n<\lambda$ on the set E_n on $\Omega(F)$. Fix $\epsilon>0$ and suppose that the difference in (50) is less than ϵ on the set

F_n: $$\lim_{n\to\infty} \underline{P}_F(F_n) = 1.$$

Then the points of $\Omega(F)$ where $\sigma n^{1/2}p_n<\lambda$ for any constant λ are included in the points of the complement of F_n or in the points common to F_n and the set on which

$$\frac{1}{\sigma n^{1/2}}\sum_{j=1}^{n}\alpha(x_j) < \lambda + \epsilon.$$

The points of $\Omega(F)$ where $\sigma n^{1/2}p_n<\lambda$ include the points where

$$\frac{1}{\sigma n^{1/2}}\sum_{j=1}^{n}\alpha(x_j) < \lambda - \epsilon$$

which also belong to F_n. These considerations show that (38) is true, since (49) is uniform in λ.

Theorem 6 requires a slight modification if the parameter p is replaced by several parameters, $p^{(1)}, \cdots, p^{(r)}$. Theorem 5 evidently needs no essential change in this case. In Theorem 6 we replace (32) by

$$\begin{aligned}\log f(x, p^{(1)}, \cdots, p^{(r)}) &= \log f(x, p)\\ &= \log f(x, 0) + \sum_{i=1}^{r} p^{(i)}\alpha_i(x) + \tfrac{1}{2}\sum_{i,k=1}^{r} p^{(i)}p^{(k)}\beta_{ik}(x) + \gamma(x, p),\end{aligned} \tag{32'}$$

$$\beta_{ik}(x) = \beta_{ki}(x),$$

where we take the true set of parameters as $(0, \cdots, 0)$, and where we suppose that the first partial derivatives of $\gamma(x, p^{(1)}, \cdots, p^{(r)})$ exist in a neighborhood of the origin in the r-dimensional p-space, and are continuous at the origin. Conditions (ii) and (iii) are modified in an obvious way, and (36) becomes

$$\int_{-\infty}^{\infty}\alpha_i(x)\alpha_k(x)f(x, 0)dx + \int_{-\infty}^{\infty}\beta_{ik}(x)f(x, 0)dx = 0, \tag{36'}$$

proved as before. If we set

$$\sigma_{ik} = \int_{-\infty}^{\infty} \alpha_i(x)\alpha_k(x)f(x, 0)dx,$$

the theorem states that the joint distribution of $p_n^{(1)}, \cdots, p_n^{(r)}$, the nth approximation of maximum likelihood, approaches normality, where the matrix of the variances and covariances of the $p_n^{(i)}$ becomes $1/n$ times the inverse matrix of $\|\sigma_{ik}\|$, which we assume non-singular. The proof will be sketched briefly. The theorem is stated in a way invariant under non-singular linear transformations of $p^{(1)}, \cdots, p^{(r)}$. We can assume that a linear transformation has been performed already, if necessary, reducing the positive definite quadratic form

$$(51) \qquad -\sum_{i,j=1}^{r} p^{(i)}p^{(j)} \int_{-\infty}^{\infty} \beta_{ij}(x)f(x, 0)dx = \int_{-\infty}^{\infty} \left[\sum_{i=1}^{r} p^{(i)}\alpha_i(x) \right]^2 f(x, 0)dx$$

to canonical form, so that

$$(52) \qquad -\int_{-\infty}^{\infty} \beta_{ij}(x)f(x, 0)dx = \delta_{ij}$$

where δ_{ij} is the usual Kronecker delta. Equation (43) becomes

$$(43') \qquad \left.\frac{\partial L_n}{\partial p^{(k)}}\right|_{p_n(k)} = \sum_{j=1}^{n} \alpha_k(x_j) + \sum_{i=1}^{r} \sum_{j=1}^{n} p^{(i)}\beta_{ki}(x_j) + \gamma_{p(k)}(x, p_n) = 0,$$

and (47) becomes

$$(47') \qquad (\sigma_{ii}n)^{1/2}p_n^{(i)} = \frac{1}{(\sigma_{ii}n)^{1/2}} \sum_{j=1}^{n} \alpha_i(x_j) + R_n^{(i)}.$$

It is shown as before that R_n approaches 0 in probability as n becomes infinite. For large n, the estimates $p_n^{(i)}$ are then distributed nearly normally, with variances and covariances obtained from $1/n$ times the inverse of the matrix $\|\sigma_{ik}\|$.

Columbia University,
New York, N. Y.