ARKIV FÖR MATEMATIK Band 1 nr 17

Communicated 25 January 1950 by HARALD CRAMÉR and TRYGVE NAGELL

Stochastic processes and statistical inference

By ULF GRENANDER

Introduction

The purpose of this thesis is, partly to show the possibility of applying statistical concepts and methods of inference to stochastic processes, and partly to obtain practically working methods of this kind by studying special cases of inference.

Time-series have been subjected to statistical treatment in a more or less systematical way for a very long time, but unlike the case of finite dimensional samples, there exists no unified theory. The extensive literature on stochastic processes has but rarely touched upon questions of inference. On the other hand, the attempts to treat time-series data do not seem to have been much influenced by the theory of stochastic processes. This is specially the case when considering a continuous time-parameter, which will be our main interest in the following chapters. The treatment of the problem^ in the present dissertation is based on the general idea outlined in CRAMÉR: Mathematical methods of statistics — to base statistical methods on the mathematical theory of probability.

In the first two chapters we shall give a short survey of some fundamental facts about stochastic processes and statistical inference. The third and fourth chapters will deal with the problem of testing hypotheses and the fifth with estimation. Finally in the sixth chapter we shall show very shortly that prognosis and filtering of time-series are questions similar to testing and estimation and can be treated on analogous lines.

Some topics in the theory of stochastic processes

1.1. Measure of probability. Let us consider an abstract space Ω with the following properties. The points in Ω are denoted by ω. In Ω is defined a Borelfield of sets containing also Ω. On this Borelfield there is defined a completely additive, non-negative setfunction P for which $P(\Omega) = 1$. Then P is said to be a probability-measure on Ω. It is sometimes convenient to close the measure by defining every set, which can be enclosed by a set (belonging to the Borelfield) of measure zero, as measurable with measure zero.

If $f(\omega)$ is a real function defined on Ω and measurable with respect to P, $f(\omega)$ is called a stochastic variable. The mean value operator E is defined as

$$E\,f(\omega) = \int_{\Omega} f(\omega)\,d\,P(\omega)$$

if $f(\omega)$ is integrable with respect to P. The modifications necessary in order to treat complex valued stochastic variables are evident.

Let $f_1(\omega)$, $f_2(\omega)$, . . . $f_n(\omega)$ be n stochastic variables defined on Ω. If it is true for every choice of n Borelsets E_1, E_2, . . . E_n on the real axis that

$$P\,\{f_i(\omega) \in E_i;\ i = 1, 2, \ldots n\}_\omega = \prod_{i=1}^{n} P\,\{f_i(\omega) \in E_i\}_\omega$$

the variables are said to be independent.

Let A be an arbitrary measurable set and $x_1(\omega)$, $x_2(\omega)$, . . . $x_n(\omega)$ n stochastic variables. If M is a "cylinderset" with an arbitrary measurable set in the n-dimensional Euclidean space $R_n(x_1, x_2, \ldots x_n)$ as basis, there exists one, and but for equivalence, only one function $P(A\,|\,x_1, x_2, \ldots x_n)$ that satisfies

$$P(A\,M) = \int_{M} P(A\,|\,x_1, x_2, \ldots x_n)\,d\,P(x_1, x_2, \ldots x_n)$$

for every M with the said properties (see KOLMOGOROFF 1, p. 42). $P(A\,|\,x_1, x_2, \ldots x_n)$ is called the conditional (with regard to x_1, x_2, . . . x_n) probability of A. The conditional expectation is defined in an analogous way (see KOLMOGOROFF 1, p. 47). It has been shown that the conditional probability has usually the same properties as the absolute probabilities. Let A be a fixed set. Then we have

$$0 \leq P(A\,|\,x_1, \ldots x_n) \leq 1$$

for almost all x_1, . . . x_n and other similar analogies. More generally it can be shown under some conditions that $P(A\,|\,x_1, \ldots x_n)$ can be defined so that it is almost certainly a probability distribution.

1.2. Stochastic processes. Let T be the whole or a part of the real axis. In the set of time-points T we observe a quantity depending upon time and in some way or other containing a random component. By repeating this experiment a large number of times we get a population of functions defined on T. The idealization of such a population together with a measure of probability, which we shall define more precisely later on in this section, is called a stochastic process. The elements in the population are called the realizations or the sample-functions.

When we want to define in a more rigorous way what is meant by a stochastic process, we find at least three different ways of doing so. Consider the quantity under observation at a fixed time-point t_0. The result of a large number of such experiments can in the usual way be described by a stochastic variable, which we denote by $x(t_0)$, leaving out the ω as is usually done in this connection. Later we shall see that it is convenient to consider the stochastic variable as a point in an abstract space, which of course is different from the sample-space Ω, on which the stochastic variable is defined. When t_0 takes on every value in T, we get a one-parameter family of stochastic variables. This curve in the abstract space is defined as the stochastic process under consideration.

The second alternative is obtained by fixing a realization, and regarding this as a function of t. Let us denote the realization by ω and the function space consisting of all real functions defined on T by Ω. It is then possible to define the process as the family $x_\omega(t)$ of real functions, where ω plays the rôle of a parameter.

This dualism depends evidently on the fact that the value of the process is a function of two variables: the time t and the realization ω. We get the third alternative definition by defining the process as a random function $f(t, \omega)$, where $f(t, \omega)$ for fixed t shall be a measurable function of ω, i. e. a stochastic variable.

1.3. Stochastic processes as families of stochastic variables. Processes regarded from the first point of view have usually been described with the aid of the first two moments. Suppose that $E\,x(t)^2 < \infty$ for all $t \in T$, and introduce the quantities

$$\begin{cases} m(t) = E\,x(t) \\ r(s, t) = E\,[x(s) - m(s)]\,[x(t) - m(t)]. \end{cases}$$

$m(t)$ is called the mean value function, and $r(s, t)$ the covariance function. By considering the process $x(t) - m(t)$ instead we can suppose that $m(t) \equiv 0$.

Form all finite linear combinations

$$\sum_{i=1}^{n} c_i\,x(t_i)$$

with real coefficients c_i and $t_i \in T$, and close this set by convergence in the mean. We then obtain a Hilbert space $L_2(X)$, if we define the inner product as

$$(f, g) = E\,f\,g.$$

The first systematic treatment of stochastic processes with the aid of Hilbert space methods is due to KARHUNEN. In some connections it is convenient to introduce complex-valued processes and the inner product is then defined as

$$(f, g) = E\,f\,\bar{g}.$$

If, for every $z \in L_2(X)$, it is true that the real function $E\,z\,x(t)$ is Lebesgue measurable, the process is said to be measurable (K). If further, for every $z \in L_2(X)$, $E\,z\,x(t)$ is Lebesgue integrable over T and the expression

$$\sup_{z \in L_2(X)} \frac{1}{\|z\|} \left| \int_T E\,z\,x(t)\,dt \right|$$

is finite, there exists a unique element $X \in L_2(X)$ satisfying the equation

$$E\,z\,X = \int_T E\,z\,x(t)\,dt.$$

Then the process is said to be integrable (K) over T with the integral X (see KARHUNEN 3, Satz 5).

If $\|x(t) - x(t_0)\|$ is a continuous function of t at $t = t_0$, the process is said to be continuous in the mean at $t = t_0$. If the corresponding holds for every $t_0 \in T$, the process is said to be continuous in the mean on T.

Suppose that $x(t)$ is continuous in the mean on the finite interval (a, b). If $a = t_0^n < t_1^n < \cdots < t_n^n = b$ and $\operatorname{Max}_{\nu} (t_\nu^n - t_{\nu-1}^n) \to 0$ when n tends to infinity, it can be shown that the expression

$$S_n = \sum_{\nu=1}^{n} x(t_\nu^n)(t_\nu^n - t_{\nu-1}^n)$$

converges in the mean to a stochastic variable I when n tends to infinity, irrespective of the choice of the points t_ν^n. I is said to be the integral (C) of $x(t)$ over (a, b)

$$I = \int_a^b x(t)\,dt.$$

(See CRAMÉR 2, lemma 3.)

If $x(t)$ is continuous in the mean on (a, b), it is evidently measurable (K), and using Schwarz' inequality it is seen that

$$\sup_{z \in L_2(X)} \frac{1}{\|z\|} \left| \int_T E\,z\,x(t)\,dt \right|$$

is finite. Thus the process is integrable (K) over (a, b) with a uniquely determined integral. But S_n and the limit element I belong to $L_2(X)$ and, as convergence in the mean implies weak convergence, we get

$$E\,z\,I = \lim_{n\to\infty} \sum_{\nu=1}^{n} E\,z\,x(t_\nu^n)(t_\nu^n - t_{\nu-1}^n).$$

As $E\,z\,x(t)$ is continuous it is Riemann integrable and thus

$$E\,z\,I = \int_a^b E\,z\,x(t)\,dt$$

so that in this case the two definitions of integration coincide.

Suppose that the process $Z(\lambda)$, $-\infty < \lambda < \infty$, has mean value zero and finite variance. If for every pair of disjoint intervals (λ_1, λ_2) and (λ_3, λ_4) it is true that

$$E\,[Z(\lambda_2) - Z(\lambda_1)]\,[Z(\lambda_4) - Z(\lambda_3)] = 0$$

$Z(\lambda)$ is said to be an orthogonal process. Then

$$r(\lambda, \lambda) = E\,Z(\lambda)^2 = F(\lambda)$$

is a non-decreasing function of λ. Suppose that $f(\lambda)$ is a real function with $\int\limits_{-\infty}^{\infty} f(\lambda)^2\, d\,F(\lambda) < \infty$. Then it is possible to define $\int\limits_{-\infty}^{\infty} f(\lambda)\, d\,Z(\lambda)$ by the aid of the Riemann-Stieltjes partial sums. (See KARHUNEN 3.) The analogous holds for complex-valued orthogonal processes.

KARHUNEN (3, Satz 10 which is more generally formulated) has given the following important theorem on representation of a stochastic process. Let $x(t)$ be a process that can take complex values and with mean value zero and

$$r(s, t) = \int\limits_{-\infty}^{\infty} f(s, \lambda)\, \overline{f(t, \lambda)}\, d\,\sigma(\lambda)$$

where σ is a measure on the real axis. This measure shall have the property that the whole axis is the denumerable sum of sets of finite σ-measure. $f(s, \lambda)$ shall be quadratically integrable with respect to σ for every s. Then there exists an orthogonal process $Z(\lambda)$ so that

$$x(t) = \int\limits_{-\infty}^{\infty} f(t, \lambda)\, d\,Z(\lambda).$$

Consider a process which is continuous in the mean and has mean value zero. If the covariance function $r(s, t)$ depends only upon the difference $s - t$, the process is said to be stationary in the wide sense. According to a well-known theorem of KHINTCHINE 1 there exists a bounded non-decreasing function $F(\lambda)$ so that

$$r(s, t) = r(s - t) = \int\limits_{-\infty}^{\infty} e^{i(s-t)\lambda}\, d\,F(\lambda).$$

Then the process itself has an analogous representation

$$\begin{cases} x(t) = \int\limits_{-\infty}^{\infty} e^{it\lambda}\, d\,Z(\lambda) \\ E\,|Z(\lambda)|^2 = F(\lambda), \end{cases}$$

(see CRAMÉR 3), where $Z(\lambda)$ is an orthogonal process. According to the mean ergodic theorem (see HOPF 1) the expression $\frac{1}{2T}\int\limits_{-T}^{T} x(t)\, dt$ converges in the mean to a stochastic variable $\hat{x}$ when T tends to infinity.

$$E\,|\hat{x}|^2 = F(+0) - F(-0)$$

so that for $\hat{x}$ to be identically zero with probability one it is necessary and sufficient that there is no discrete spectral mass at $\lambda = 0$.

Suppose that $x(t)$ is real and continuous in the mean on the finite interval (a, b) with mean value zero and covariance function $r(s, t)$. The covariance function is positive semi-definite, and by considering the integral equation

$$\varphi(t) = \lambda \int_a^b r(t, s)\, \varphi(s)\, ds$$

one gets the representation

$$r(s, t) = \sum_1^\infty \frac{\varphi_\nu(s)\, \varphi_\nu(t)}{\lambda_\nu}$$

with uniform convergence according to Mercer's theorem. Here $\varphi_\nu(t)$ are the eigen-functions of the integral equation, and λ_ν the corresponding eigen-values. From Karhunen's general theorem on representation of stochastic processes it then follows that

$$x(t) = \sum_1^\infty z_\nu \frac{\varphi_\nu(t)}{\sqrt{\lambda_\nu}}$$

with convergence in the mean for every $t \in (a, b)$. (See KARHUNEN 1.) The z's are stochastic variables with

$$\begin{cases} E\, z_\nu = 0 \\ E\, z_\nu z_\mu = \delta_{\nu\mu}. \end{cases}$$

If the contrary is not stated we will always in the following suppose that the kernel $r(s, t)$ is non-degenerate, i. e. $\lambda_\nu < \infty$ for all ν.

Another type of representation is obtained in the following way. If $Z(\lambda)$ is an orthogonal process of bounded variance

$$E\,|\,Z(\lambda)\,|^2 < k, \quad -\infty < \lambda < \infty,$$

(a bounded orthogonal process), we define the measure σ corresponding to the non-decreasing function $E\,|\,Z(\lambda)\,|^2$ in the usual way. The set of all functions which are quadratically integrable over $(-\infty, \infty)$ with respect to σ is a Hilbert space if the usual quadratic metric is used. In this space we choose a CON system of functions $\{\varphi_\nu(\lambda)\,;\, \nu = 1, 2, \ldots\}$. Let $\varepsilon_{\lambda_0}(\lambda)$ denote the function which is one for $\lambda \leq \lambda_0$ and zero for $\lambda > \lambda_0$. Then we obtain using the completeness of the system and Parseval's relation

$$\sum_1^\infty \overline{(\varphi_\nu\,;\, \varepsilon_{\lambda_0})}\, (\varphi_\nu\,;\, \varepsilon_{\lambda_1}) = (\varepsilon_{\lambda_0}\,;\, \varepsilon_{\lambda_1})$$

and hence

$$\sum_1^\infty \int_{-\infty}^{\lambda_0} \overline{\varphi_\nu(\lambda)}\, d\sigma(\lambda) \int_{-\infty}^{\lambda_1} \varphi_\nu(\lambda)\, d\sigma(\lambda) = \int_{-\infty}^{\mathrm{Min}\,(\lambda_0,\, \lambda_1)} d\sigma(\lambda).$$

But $E|Z(\min \lambda_0, \lambda_1)|^2$ is the covariance function $r(\lambda_0, \lambda_1)$ of the orthogonal process, so that by Karhunen's theorem there exists a sequence of stochastic variables $\{z_\nu\}$ for which

$$\begin{cases} E z_\nu = 0 \\ E z_\nu \bar{z}_\mu = \delta_{\nu\mu} \\ Z(\lambda) = \sum\limits_1^\infty z_\nu \int\limits_{-\infty}^{\lambda} \varphi_\nu(\lambda)\, d\sigma(\lambda). \end{cases}$$

Take as a special case σ as the Lebesgue measure in (0, 1) and vanishing outside this interval. On (0, 1) the system $\{e^{2\pi i\nu\lambda};\ \nu = 0, \pm 1, \ldots\}$ is a CON system and we get the representation

$$Z(\lambda) = z_0\lambda + \sum_{-\infty}^{\infty}{}' z_\nu \frac{e^{2\pi i\nu\lambda} - 1}{2\pi\nu i};\ 0 \leq \lambda \leq 1;$$

where the $'$ in the summation symbol indicates that the term corresponding to $\nu = 0$ is left out at the summation. In the case of a normal process this is the Wiener random function (see PALEY-WIENER 1).

The derivative of a stochastic process can be defined either as the strong or as the weak limit of $\frac{x(t+h) - x(t)}{h}$ when h tends to zero. The latter will in general be convenient when dealing with linear differential equations. In this way one can show e. g. that the equation of Langevin

$$\frac{dx(t)}{dt} + \beta x(t) = \frac{dB(t)}{dt}$$

where β is a constant and $B(t)$ is the process of Einstein-Smoluchovsky has the solution

$$x(t) = e^{-\beta t} x(0) + e^{-\beta t} \int\limits_0^t e^{\beta\tau} dB(\tau).$$

The same solution has been given by DOOB 4 using another interpretation of the differential equation. As is known, one obtains in this way the stationary, normal Markoff process.

We have hitherto supposed that the variance is finite. If we only suppose that $E|x(t)|^p < \infty$, $p \geq 1$ we can still use similar methods. We form all finite linear combinations just as before and close this set with respect to strong convergence according to the metric

$$\|x - y\| = \sqrt[p]{E|x - y|^p}.$$

We get a Banach space X. X is evidently situated in $L_p(\Omega)$. To define the integral of a stochastic process we use the theory of integration in Banach space developed by PETTIS 1. The process is said to be measurable if for every linear functional $F \in \bar{X}$ (where $\bar{X}$ is the adjoined space to X) $F[x(t)]$ is a Lebesgue measurable function of t. If there is a unique element $I \in X$ so that

$$F(I) = \int_T F[x(t)]\, dt$$

for every $F \in \bar{X}$ the process is said to be integrable with the integral I. Using the general form for the said functionals (see e. g. BANACH 1) we get a method of integration which for $p = 2$ evidently gives us the (K) integration.

As most processes met with in practical applications have finite variance and as it is possible to develop most of the theory in X in a manner which is analogous to that used in Hilbert space we now leave this subject.

1.4. Stochastic processes as families of real functions. In the above method the process is considered as a curve in an abstract space. In the investigation of the process only the metric properties of this curve have usually been used, disregarding the concrete meaning of the points of the curve. The method has been applied with great success to many problems, especially to linear ones, but in certain cases they are not sufficient. This is particularly the case when considering properties of the individual realizations, e. g. continuity and measurability. When we are going to make statements of inferential nature, we clearly want to use all available knowledge about the process and not only its linear properties. This is why we shall in most cases interpret the concept of a stochastic process in the sense used by Doob in a series of papers (see DOOB 1, 3, 4).

Let $t_1, t_2, \ldots t_n$ be a finite number of time-points in T and $a_1, a_2, \ldots a_n$, $b_1, \ldots b_n$ real numbers. Then the set

$$\{a_i \leq x(t_i) < b_i;\ i = 1, 2, \ldots n\}_\omega \in \Omega'$$

is called a finite dimensional interval in Ω'. Ω' is the space of all real functions defined on T. Suppose now that the probabilities of all finite dimensional intervals are given in a consistent way. Then it is possible according to a theorem of KOLMOGOROFF 1 to extend the probability-measure to the Borel field constructed from all finite dimensional intervals. The measure is closed in the way mentioned in 1.1 and the measure thus obtained is denoted by P'.

In order to be able to consider the probability of sets depending essentially upon the values of the process in a non-denumerable set of time-points, e. g. all continuous or bounded realizations, Doob proceeds in the following way. Let Ω be a subset of Ω' with $\bar{P}'(\Omega) = 1$. If $\Lambda = \Lambda' \Omega$ where Λ' is measurable with respect to P', we define $P(\Lambda) = P'(\Lambda')$. It can be shown that this definition leads to a unique determination of the measure. This means that we have to confine ourselves to a smaller sample-space, consisting of appropriate functions.

If there is a subset $\Omega < \Omega'$ with $\bar{P}'(\Omega) = 1$ such that the function $x(t, \omega)$, where $\omega \in \Omega$, is measurable with respect to the product-measure introduced on $\Omega \times T$ in the usual way, the process is said to be measurable (D). In all the cases we are going to consider, the processes are continuous in the mean. If T is a finite interval (a, b), we obtain using Schwarz' inequality

$$\int_T \int_\Omega |x(t, \omega)| \, dP \, dt < \infty$$

so that according to Fubini's theorem $x(t)$ is Lebesgue measurable and integrable with probability one. Further $\int_T x(t) \, dt$ is a measurable function on Ω, i.e. a stochastic variable. This variable is defined as the integral (D) of the process over (a, b). If $z(\omega)$ is an arbitrary quadratically integrable function on Ω it can be shown by applying Fubini's theorem that

$$E\left\{z \int_T x(t) \, dt\right\} = \int_T E z x(t) \, dt$$

which implies that the (D) integral coincides with the (C) and (K) integral in the case when the process is continuous in the mean. This identity can be shown under considerably more general conditions.

Some criteria for deciding whether a process is (D) measurable or not have been given. The following one is due to KOLMOGOROFF (see AMBROSE 1). In order that a process $x(t)$ shall be measurable (D) it is necessary and sufficient that for every $\varepsilon > 0$ and almost every t

$$P\{|x(t+h) - x(t)| > \varepsilon\} \to 0$$

when h tends to zero in a set that may depend upon t but not upon ε, and of metric density 1 in $h = 0$. This condition is satisfied e.g. when the process is continuous in the mean.

Another result in this direction is the following which we shall use later on. On Ω' is defined a measure P' belonging to a time-homogeneous process with

$$P'\{a < x(t+\delta) - x(t) < b\} = \frac{1}{\sqrt{2\pi\delta}} \int_a^b e^{-\frac{x^2}{2\delta}} \, dx; \; \delta > 0.$$

Let Ω_c be the set of all continuous functions. Then $\overline{P}'(\Omega_c) = 1$ and Ω_c is the sample-space of a measurable process (see DOOB 1).

If $x(t)$ is a normal process with mean value zero and covariance function $\sigma^2 e^{-\beta|t-s|}$ $(\beta > 0)$ (i.e. the process considered in connection with the equation of LANGEVIN in 1.3) we make the transformation (see DOOB 4)

$$y(t) = \sqrt{t}\, x\left(\frac{1}{2\beta} \log t\right); \; t > 0.$$

This process is of the type described above, and thus $x(t)$ itself is a measurable process with Ω_c as its sample-space.

We introduce the translation operator T_h operating on the individual functions $x(t)$ in the sample-space

$$T_h x(t) = x(t+h).$$

If for every set S it is true that $P(S) = P(T_h S)$ for all h, the process is said to be stationary in the strict sense. Let further $x(t)$ be (D) measurable. If $f(\omega)$ is integrable on Ω, Birkhoff's ergodic theorem states that the limit

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} f(\omega_t)\, dt = \hat{f}(\omega)$$

exists almost certainly. $\hat{f}(\omega)$ is measurable and integrable on Ω (see e.g. HOPF 1).

If, for every integrable $f(\omega)$ it is true that $\hat{f}(\omega)$ is identically constant with probability one, the process is said to be ergodic.

A measurable set A satisfying $T_h A = A$ for all h is called invariant with respect to the translation operator. If the process has the property that every invariant set has probability one or zero, the process is said to be metrically transitive. It can be shown that ergodicity and metric transitivity are equivalent concepts because of the finite measure of the sample-space (see HOPF 1).

In the third approach the process is considered as a function of two variables $x(t, \omega)$, $\omega \in \Omega$, $t \in T$, with a given measure of probability on a reference space Ω. DOOB and AMBROSE have shown that this is substantially equivalent to Doob's method. In particular problems one of these may be more suitable than the other, but it has to be decided in each case according to convenience which to use.

Elementary notions in the theory of statistical inference

2.1. Power properties of a test. In this chapter we shall give some basic facts about statistical inference in the case of finite dimensional samples. Let us regard the observed values $x_1, x_2, \ldots x_n$ as representative of a population X, whose probability distribution P_0 is completely specified by the hypothesis H_0. Such a hypothesis which completely specifies the probability distribution, is called simple.

Having observed $x_1, x_2, \ldots x_n$ we want to make a statement about the truth or falsehood of H_0. The methods used for this purpose shall have the property of giving correct results in the majority of cases in the long run. Let us form a measurable region W with $P_0(W) = \varepsilon$. If $(x_1, x_2, \ldots x_n) \in W$ we reject the hypothesis, otherwise accept it. W is called a critical region of size ε. It is evident that when H_0 is true we shall reject it with the probability ε.

In this way we can form an infinity of tests corresponding to the different critical regions of size ε. To be able to choose between them Neyman and Pearson take an alternative hypothesis H_1 into consideration. The ε introduced above is called the error of the first kind. Another way of committing an error is to accept H_0 when H_1 is true. The probability of this is called the error of the second kind and is $P_1(W^*)$. Having fixed ε we now want to find the region W of size ε and of minimum $P_1(W^*)$. This clearly gives us a test of optimum character.

The case usually dealt with is when the probability distributions P_0 and P_1 are given by frequency functions $f_0(x_1, \ldots x_n)$ and $f_1(x_1, \ldots x_n)$ respectively.

The likelihood ratio is defined as

$$l(x_1, x_2, \ldots x_n) = \frac{f_1(x_1, x_2, \ldots x_n)}{f_0(x_1, x_2, \ldots x_n)}.$$

The set where both numerator and denominator are zero have probability zero according to both hypotheses, so that we can define the likelihood ratio as e.g. one in this set. It can now be proved that the best critical region in the sense explained above is given by

$$W = \{l(x_1, \ldots x_n) \geq c\}_{x_1, \ldots x_n}$$

where the constant c is chosen to satisfy $P_0(W) = \varepsilon$. For the trivial difficulty when this equation has no solution we refer to CRAMÉR 4. The case when the probability distributions are of the discrete type is dealt with in the same way.

Usually the alternative hypothesis is not simple but may depend upon a real parameter a, H_0 itself corresponding to the value $a = a_0$. Then to every fixed a we get a best critical region for H_0 against H_a. If all these critical regions coincide, the corresponding test is called uniformly most powerful. Unfortunately this is almost never the case when $a - a_0$ takes both positive and negative values (see KENDALL 1).

Then it is customary to consider only a subset of all possible tests. It is fairly evident that $P(W; a) \geq P(W; a_0) = \varepsilon$ is a desideratum for a good test. Such a test is called unbiased. In the class of all unbiased tests we try to find a region of size ε for which $P(W^*, a)$ is minimum for a fixed a. Under some regularity conditions it can be shown that the test

$$W = \{f(x_1, \ldots x_n; a) \geq c f(x_1, \ldots x_n; a_0) + c_1 f_1(x_1, \ldots x_n; a_0)\}_{x_1, \ldots x_n},$$

where $f_1(x_1, \ldots x_n; a) = \dfrac{\partial f(x_1, \ldots x_n; a)}{\partial a}$, has the desired property. It can happen that for every a we get the same region W. Then the corresponding test is called the uniformly most powerful unbiased test.

For more complicated situations, e.g. when it is required to test one composite hypothesis against another, we refer to KENDALL 1, where also a list of the original papers may be found. In the following chapters we shall show the possibility of using the above methods on stochastic processes. The principal difficulty of transferring these concepts to the infinite dimensional case thus being solved, it seems easy, at least in principle, to extend the results to composite hypotheses in the same way as in the finite dimensional sample case.

2.2. Some desiderata for an estimate. Suppose now that the hypotheses H_a are completely specified by their probability distributions P_a, when a is known. a is a real parameter in some interval A. We want to form a sample-function $a^*(x_1, \ldots x_n)$ which can be used as an estimate of a. One has several possible ways of describing desirable properties of a^*.

Let the sample-number n tend to infinity. Then our information about the population is increased, and if the sequence $a_n^*(x_1, \ldots x_n)$ converges to a in probability with regard to P_a for every $a \in A$, we say that the estimate (or more correctly the sequence of estimates) is consistent.

Irrespective of this asymptotic behaviour we can describe the goodness of an estimate $a^* (x_1, \ldots x_n)$ for a fixed n, by studying its two first moments. If for every $a \in A$ it is true that

$$E_a a^* = a$$

we say that a^* is an unbiased estimate of a. This is certainly a desirable property.

The fluctuation of an estimate about its true value can be measured by the expression $E_a (a^* - a)^2$. Put

$$E_a a^* = a + b(a),$$

where $b(a)$ is called the bias of a^*. Then it can be shown under some regularity conditions that (see CRAMÉR 4)

$$E_a (a^* - a)^2 \geq \frac{\left(1 + \frac{d b}{d a}\right)^2}{E_a \left(\frac{\partial \log f}{\partial a}\right)^2}.$$

If a^* is unbiased, we define the efficiency $e(a^*)$ of a^* as

$$e(a^*) = \frac{1}{D^2(a^*)\, E_a \left(\frac{\partial \log f}{\partial a}\right)^2},$$

and thus have

$$0 \leq e(a^*) \leq 1.$$

In the case $e(a^*) = 1$, we say that the estimate is efficient. It can be shown that if two estimates a_1^* and a_2^* are both efficient, then $a_1^* = a_2^*$ almost certainly.

When we are considering a sequence of estimates a_n^* it may happen that it has desirable properties although the two first moments do not exist. The following definition (WALD 1) takes this possibility into account. The estimate is said to be asymptotically efficient if there is a sequence of stochastic variables u_n with

$$\lim_{n \to \infty} E_a u_n = 0; \quad \lim_{n \to \infty} E_a u_n^2 = 1$$

so that

$$\sqrt{E_a \left(\frac{\partial \log f}{\partial a}\right)^2}\,(a_n^* - a) - u_n \to 0, \quad n \to \infty,$$

in probability with regard to P_a for every $a \in A$.

The most important method of estimation is the method of maximum likelihood. If $x_1, x_2, \ldots x_n$ are independent stochastic variables, each of them having the frequency function $f(x; a)$, we form the simultaneous frequency function

$$f(x_1, \ldots x_n; a) = f(x_1; a) f(x_2; a) \ldots f(x_n; a).$$

As an estimate of a we take a non-identically constant solution $a^*(x_1, \ldots x_n)$ of the equation

$$\frac{\partial \log f(x_1, \ldots x_n; a)}{\partial a} = 0.$$

Under some regularity conditions it can be shown that this equation has a solution, which converges in probability, with regard to P_a, to a for every $a \in A$ when n tends to infinity. This estimate is asymptotically normal and asymptotically efficient. The analogous result holds for the discrete type of distribution.

2.3. Confidence regions. The preceding section deals with point-estimation. In many cases we do not want to assign a single value to the unknown parameter but an interval or some more general region, which, for a given ε, contains the true value of the parameter with the probability $1 - \varepsilon$, for all $a \in A$. It is possible to do so in the following way (see CRAMÉR 4).

For every fixed value of a we choose a set $S(a) \in X$, such that $P_a[S(a)] = 1 - \varepsilon$. For every $(x_1, \ldots x_n) \in X$ we denote by $\Sigma(x_1, \ldots x_n) < A$ the set of all a such that for the element $(a; x_1, \ldots x_n)$ in the product space $A \times X$ the relation

$$(a; x_1, \ldots x_n) \in D$$

holds, where $D < A \times X$ is determined as the set of all $(a; x_1, \ldots x_n)$ such that $(x_1, \ldots x_n) \in S(a)$. Then we have for every a

$$\{x_1, \ldots x_n) \in S(a)\} = \{a \in \Sigma(x_1, \ldots x_n)\}$$

so that

$$P_a[a \in \Sigma(x_1, \ldots x_n)] = 1 - \varepsilon.$$

$\Sigma(x_1, \ldots x_n)$ is called a confidence region for a with the confidence coefficient $1 - \varepsilon$. If, in particular, $\Sigma(x_1, \ldots x_n)$ is an interval in A, we call it a confidence interval for the parameter.

Observable coordinates of a stochastic process

3.1. In the following we shall try to transfer the classical methods of statistical inference to stochastic processes. Let us consider specially the case when it is required to estimate a single parameter a. From the preceding chapters it is evident that the natural way of doing this is to form a function of the observed realization, and choose this function in order to make it a good estimate of a in some sense. As a function of a function the estimate is a functional defined on the sample-space. Now the theory of functionals has mainly been developed for linear functionals or some special types. Usually (but with some exception to be studied later) there is nothing in our problem that gives us reason to confine ourselves to these special types of functionals.

Therefore we shall consider quite general types of functionals, restricted only by some natural regularity conditions.

We get the information obtained by observing a stochastic process in the form of one or more real functions. For our purpose it will be more convenient to translate this information into the form of a sequence of real numbers $(c_1, c_2, \ldots)$, if possible, which implies that we have to use a sample-space of smaller power than the space Ω' consisting of all real functions. This will be the case in the problems we are going to study. How to choose this sequence is only partly a mathematical question. One has to take into account what properties of the realizations that really can be observed in the practical application under consideration. The c's are called the observable coordinates of the process. We shall see that it is very important to choose these coordinates in a way that facilitates the construction of estimates, test functions and so on.

Consider the following important case. $x(t)$ is a normal process that is observed in the finite time interval $T = (a, b)$. The process is continuous in the mean with mean value function $m(t)$ and covariance function $r(s, t)$. As shown in 1.3, we have

$$x(t) = m(t) + \sum_1^\infty z_\nu \frac{\varphi_\nu(t)}{\sqrt{\lambda_\nu}}$$

with convergence in the mean for every $t \in T$. λ_ν and $\varphi_\nu(t)$ are the eigen-values and the corresponding eigen-functions of the integral equation

$$\varphi(s) = \lambda \int_a^b r(s, t)\, \varphi(t)\, dt$$

and

$$\begin{cases} E z_\nu = 0. \\ E z_\nu z_\mu = \delta_{\nu\mu}. \end{cases}$$

In this case the stochastic variables z_ν will obviously be normally distributed, being limits of finite linear combinations of the values of $x(t)$ at time-points in T.

Now we represent the process by the following random function

$$x(t, \omega) = x(t, z_1, z_2, \ldots) = m(t) + \sum_1^\infty z_\nu \frac{\varphi_\nu(t)}{\sqrt{\lambda_\nu}},$$

where the quantities in the right hand member have the same properties as before. KAC and SIEGERT 1 have shown that the sum converges for almost all (t, ω) in $T \times \Omega$. Further, they have shown that the expression converges in the mean with regard to Lebesgue measure on T almost certainly. Taking a quadratically integrable function $f(t) \in L_2(T)$ it is now possible to form the integral $\int_T f(t)\, x(t)\, dt$ which is a measurable function on Ω, because of the random function being measurable on $T \times \Omega$.

We now want to represent the information contained in a certain realization in a convenient way. One rather natural way of doing this is to form the Fourier-coefficients $(c_1, c_2, \ldots)$ of the realization with regard to a CON system. It is particularly convenient to take $\{\varphi_\nu(t)\}$ as this system if it is complete, otherwise we make it complete by adding another orthonormal system to it in the usual way. The probability distribution of these c's is easily obtained when we only consider a finite number of them. Then the usual procedure is used to extend the measure to the Borel field. When $(c_1, c_2, \ldots)$ is given we know the realization completely if we consider two functions differing at most in a set of points of measure zero as identical. From the practical point of view this seems quite enough.

In cases when the covariance function is of some simple form it may be better to use another system of coordinates. Let us consider the normal stationary Markoff process. We have seen in 1.4 that it is possible to choose the set of all continuous real functions on T as the sample-space of the process. Let $\{t_n\}$ be a denumerable set of points everywhere dense in T. Then, with probability one, the realization is completely specified when the values $x(t_n) = c_n$ are known for all n. We shall see that the choice of this system of observable coordinates will prove advantageous in the treatment of the process.

An important type of processes is the class of pointprocesses with adjoined stochastic variables. As the natural sample-space of these processes consists of step-functions, the following system of coordinates seems appropriate. If a realization has the form

$$\begin{cases} x(t) = x_0 & \text{for} \quad a \leq t \leq t_1 \\ x(t) = x_i & \text{for} \quad t_i < t \leq t_{i+1}, \qquad i = 1, 2, \ldots n-1, \\ x(t) = x_n & \text{for} \quad t_n < t \leq b \end{cases}$$

where $n, t_1, \ldots t_n, x_0, \ldots x_n$ are stochastic variables, we use as coordinates the sequence $\{n; x_0, t_1, x_1, \ldots t_n, x_n\}$. To get the coordinates in a form symmetrical to the above, we add an infinite sequence of real numbers. To these we assign some simple probability distribution, e.g. such that they are independent of each others and of the preceding coordinates (except of n of course), each having a normal distribution (0, 1). This means just a formal simplification. We shall only deal with the case when $P(n < \infty) = 1$.

When considering other types of processes it can of course happen that we have to introduce a different type of coordinates. In the following we shall always suppose that the information given by a realization can be expressed with the aid of a denumerable sequence of real numbers. We shall sometimes denote the coordinates $(x_1, x_2, \ldots) = \omega \in \Omega$, using the letter ω to symbolize the information observed in a realization. Ω is then called the coordinate-space.

The problem of testing statistical hypotheses

4.1. Existence of a most powerful test. We shall now begin to study the problem of testing statistical hypotheses in the case of a stochastic process. In this chapter it will be shown that it is possible to transfer the fundamental ideas and methods of the Neyman-Pearson theory to this case. By a simple hypothesis will be understood a completely specified measure of probability on the coordinate-space Ω.

Let us test a simple hypothesis H_0 corresponding to the measure of probability P_0 against a simple alternative H_1 corresponding to P_1. In the same way as in the classical theory we want to form a critical region W of size ε such that the error of the second kind $P_1(W^*)$ is as small as possible. In the case of a finite dimensional sample this was done by choosing the region where the likelihood function was as large as possible. In our case we have no frequency functions in Ω at our disposal, but we shall get an analogous concept serving the same purpose. Usually in the classical case one considers only the situation when the probability distributions are either of the continuous or the discrete type. These restrictions are not essential as will be seen in the following.

The analytical tools we are going to use for our purpose are the Lebesgue decomposition of an additive set-function and the Radon-Nikodym theorem (see SAKS 1). Applying these to our problem we get the following: there is a set H of P_0-measure zero and a non-negative function $f(\omega)$ integrable over Ω with respect to P_0 such that for every measurable set $E < \Omega$

$$P_1(E) = \int_E f(\omega)\, d\, P_0(\omega) + P_1(E\, H)$$

It is evident that $f(\omega)$ plays the same rôle here as the likelihood ratio in the classical case. We form the set

$$S_k = \{f(\omega) \geq k\}_\omega + H$$

and determine k so that $P_0(S_k) = \varepsilon$. One can deal with the trivial difficulty of this equation having no solution in k in the same way as in the classical case (see CRAMÉR 4) but we suppose that such a solution exists. Then we have the following

Theorem: *The test corresponding to the critical region S_k is the most powerful test of H_0 against H_1 on the level ε.*

To prove this, let E be another set with $P_0(E) = \varepsilon$, and introduce

$$E\, H^* = A, \quad E\, H = B, \quad H^* \{f(\omega) \geq k\} = F.$$

Then

$$P_1(F) = P_1(F - F\, A) + P_1(F\, A) \geq k\, P_0(F) - k\, P_0(F\, A) +$$
$$+ P_1(F\, A) = k\, P_0(A) - k\, P_0(F\, A) + P_1(F\, A) \geq P_1(A - A\, F) + P_1(F\, A) = P_1(A)$$

and

$$P_1(S_k) = P_1(F) + P_1(H) \geq P_1(A) + P_1(B) = P_1(E)$$

which proves the theorem.

4.2. Construction of a most powerful test. The above theorem has the character of a proof of existence. We shall now proceed to show how to construct H and $f(\omega)$. In order not to complicate the proof we suppose that the probability distributions of a finite number of coordinates is of the absolutely continuous type so that it is given by frequency functions $g_0(x_1, x_2, \ldots x_n)$ and $g_1(x_1, x_2, \ldots x_n)$ respectively according to the two hypotheses. It is possible to prove the same results also when the coordinates have a discrete distribution or a combination of these two simple cases. We shall consider the possible alternatives.

A: Suppose that $P_1(H) = 0$. *This will be called the regular case.* Take an arbitrary cylinderset $C_n < \Omega$ with a basis $B_n < R_n(x_1, x_2, \ldots x_n)$. Put

$$l_n(\omega) = \frac{g_1(x_1, \ldots x_n)}{g_0(x_1, \ldots x_n)}.$$

If the numerator and denominator in this expression both vanish we put $l_n(\omega) = 1$. Then

$$\int_{C_n} f(\omega)\, dP_0(\omega) = P_1(C_n) = \int_{B_n} l_n(x_1, \ldots x_n)\, g_0(x_1, \ldots x_n)\, dx_1, \ldots dx_n.$$

According to the definition of conditional expectation (see 1.1) we get

$$l_n(x_1, \ldots x_n) = E_0[f(\omega) \mid x_1, \ldots x_n]$$

almost certainly with respect to P_0. Using a theorem by Levy which has been generalized by Doob (see DOOB 3) we have

$$f(\omega) = \lim_{n \to \infty} \frac{g_1(x_1, \ldots x_n)}{g_0(x_1, \ldots x_n)}$$

almost certainly with respect to P_0, and also with respect to P_1 as $P_1(H) = 0$.

B: Suppose that $0 < P_1(H) < 1$. As we always have $P_0(H) = 0$, it is possible to cover the set H by a denumerable sum of disjoint finite dimensional intervals I_ν in the coordinate space so that $P_0\left\{\sum_1^\infty I_\nu\right\} < \varepsilon$. As $\sum_1^\infty P_1(I_\nu) \leq 1$, we can choose an integer N so that $\sum_{N+1}^\infty P_1(I_\nu) < \varepsilon$. We get

$$P_1\{H; l_n(\omega) \leq k\} \leq P_1\left\{\sum_1^\infty I_\nu;\ l_n(\omega) \leq k\right\} \leq P_1\left\{\sum_1^N I_\nu;\ l_n(\omega) \leq k\right\} + \varepsilon.$$

Choose n greater than the largest index used in determining the intervals $I_1, I_2, \ldots I_N$. Then the set

$$C = \left\{\sum_1^N I_\nu;\ l_n(\omega) \leq k\right\} < \Omega$$

is a cylinderset with a basis C' in $R_n(x_1, x_2, \ldots x_n)$, and we have

$$P_1 \{H ; l_n(\omega) \leq k\} \leq \varepsilon + \int\limits_{C'} l_n(x_1, \ldots x_n)\, g_0(x_1, \ldots x_n)\, dx_1, \ldots dx_n \leq$$

$$\leq \varepsilon + k P_0 \left\{ \sum_1^N I_\nu \right\} \leq \varepsilon (1 + k).$$

As this can be made arbitrarily small for given k we have proved that $l_n(\omega)$ converges in probability with regard to P_1 to $+\infty$ in H, as n tends to infinity.

Consider the measure of probability

$$F(S) = \frac{P_1(SH^*)}{P_1(H^*)} = \int\limits_S \frac{f(\omega)}{P_1(H^*)}\, dP_0(\omega)$$

which is possible as $P_1(H^*) > 0$. We have $F(S) = P_1(S \,|\, H^*)$, and the frequency function of $(x_1, x_2, \ldots x_n)$ with respect to F is

$$g(x_1, x_2, \ldots x_n \,|\, H^*) = \frac{g_1(x_1, \ldots x_n)\, P_1(H^* \,|\, x_1, \ldots x_n)}{P_1(H^*)}$$

because of

$$\int\limits_B \frac{P_1(H^* \,|\, x_1, \ldots x_n)}{P_1(H^*)}\, g_1(x_1, \ldots x_n)\, dx_1 \ldots dx_n = \frac{P_1(H^* B)}{P_1(H^*)} = F(B)$$

which is true for every set $B < R_n(x_1. \ldots x_n)$ according to the definition of conditional probability. As the case P_0 against F is regular, we can use the above result and get

$$\frac{f(\omega)}{P_1(H^*)} = \lim_{n\to\infty} \frac{g_1(x_1, \ldots x_n)}{g_0(x_1, \ldots x_n)} \frac{P_1(H^* \,|\, x_1, \ldots x_n)}{P_1(H^*)}$$

almost certainly. But if $\omega \in H^*$ we have

$$\lim_{n\to\infty} P_1(H^* \,|\, x_1, \ldots x_n) = 1$$

almost certainly with respect to P_1 (see DOOB 1), i.e. we have proved that in H^*

$$f(\omega) = \lim_{n\to\infty} \frac{g_1(x_1, \ldots x_n)}{g_0(x_1, \ldots x_n)}$$

almost certainly with respect to P_1. If there is a set $E < H^*$ where the above is not true, we have just seen that $P_1(E) = 0$. But if $P_0(E) > 0$, then we can use our previous result applied to P_1 and P_0 (in changed order) and get

$$\frac{g_0(x_1, \ldots x_n)}{g_1(x_1, \ldots x_n)} \to +\infty$$

in probability with respect to P_0 in E, i.e.

$$\frac{g_1(x_1, \ldots x_n)}{g_0(x_1, \ldots x_n)} \to 0$$

in probability with respect to P_0 in E. But in E we must have $f(\omega) = 0$ almost certainly with respect to P_0 so that we have in E with convergence in probability with respect to P_0

$$f(\omega) = \lim_{n \to \infty} \frac{g_1(x_1, \ldots x_n)}{g_0(x_1, \ldots x_n)}.$$

C: Suppose finally that $P_1(H) = 1$. Then we have shown that $l_n(\omega)$ converges to $+\infty$ in probability with respect to P_1. It is easily seen that $l_n(\omega)$ tends to $f(\omega)$ in probability with respect to P_0.

Summing up we have:

With respect to P_0: $l_n(\omega)$ converges to $f(\omega)$ in probability.
With respect to P_1: $l_n(\omega)$ converges to $f(\omega)$ in probability in H^*.
With respect to P_1: $l_n(\omega)$ converges to $+\infty$ in probability in H.

Now we can take as usual a sub-sequence $l_{n_\nu}(\omega)$ converging almost certainly with respect to both measures to $f(\omega)$ and to $+\infty$ in H^* and H respectively. We thus have

$$S_k = \{\lim_{\nu \to \infty} l_{n_\nu}(\omega) \geq k\}_\omega.$$

In the applications it will hardly be necessary to choose such a sub-sequence for different reasons. In a certain type of application the coordinates can be chosen in such a way that they are independent stochastic variables, and according to the zero or one law we will have probability zero or one for the convergence of $l_n(\omega)$. We will thus have either the regular case or the extreme singular case. For the pointprocesses to take another important case $l_\nu(\omega)$ will be independent of ν when ν is greater than the first coordinate n, so that we will have no difficulty with regard to the convergence of the sequence. Almost all cases we shall meet will be of the regular type.

4.3. Tests for composite alternatives. In the preceding sections we have shown how to construct the most powerful region for testing a simple hypothesis against another simple one. In practice, however, one meets more complicated situations. To deal with these several types of critical regions have been proposed. Difficult as these questions are, *it is evident that the principal difficulty of transferring the concept of the best critical region to the case of stochastic processes has been solved in 4.1—4.2.* Therefore we will only treat two more cases.

Suppose that we still want to test the simple hypothesis H_0 but against a family H_α of simple alternatives. Here α is a real parameter, which may be normed so that $\alpha = 0$ corresponds to H_0.

Fix α and construct the best critical region $S_\alpha < \Omega$ for testing H_0 against H_α on the level ε. If we get the same S for all values of α under consideration, we call the test corresponding to S uniformly most powerful. Unfortunately this is seldom the case, except sometimes for one-sided alternatives, where α takes values only of one sign.

In the classical case it is then sometimes possible to find a uniformly most powerful unbiased test. Using the result of 4.1–4.2 it is easily seen how to transfer this to our case. Suppose for simplicity, that the singular part H vanishes. If then the derivative $\frac{\partial f(\omega, \alpha)}{\partial \alpha}$ exists almost certainly and is dominated

$$\left| \frac{\partial f(\omega, \alpha)}{\partial \alpha} \right| < F(\omega)$$

for all α, where $F(\omega)$ is integrable with respect to P_0, we take as critical region the set

$$S = \left\{ f(\omega, \alpha) \geq c + c_1 \left(\frac{\partial f}{\partial \alpha} \right)_{\alpha=0} \right\}_{\omega}.$$

We suppose that this set is independent of α and that it is possible to choose the constants c and c_1 to satisfy

$$\begin{cases} P_0(S) = \varepsilon \\ \left(\frac{\partial P_\alpha(S)}{\partial \alpha} \right)_{\alpha=0} = 0. \end{cases}$$

If E is another unbiased region of the same size, we get

$$P_\alpha(S - ES) = \int_{S-ES} f(\omega, \alpha)\, dP_0(\omega) \geq c P_0(S - ES) + c_1 \left(\frac{\partial P_\alpha(S - ES)}{\partial \alpha} \right)_{\alpha=0} =$$

$$= c P_0(E - ES) + c_1 \left(\frac{\partial P_\alpha(E - ES)}{\partial \alpha} \right)_{\alpha=0} \geq P_\alpha(E - ES)$$

and

$$P_\alpha(S) \geq P_\alpha(E).$$

Thus S is really the uniformly most powerful unbiased region.

4.4. Tests for the mean value function in the normal case. Now we are going to apply the obtained results to the following problem. $x(t)$ is the normal process considered in 3.1 with known covariance function $r(s, t)$. We want to test the hypothesis

$$H_0 : E_0\, x(t) = m_0(t)$$

against the alternative

$$H_1 : E_1\, x(t) = m_1(t).$$

As described in 3.1 we take as the observable coordinates of the process

$$x_\nu = \int_a^b x(t)\, \varphi_\nu(t)\, dt; \quad y_\nu = \int_a^b x(t)\, \psi_\nu(t)\, dt; \quad \nu = 1, 2, \ldots.$$

Here $\{\varphi_\nu(t)\}$ is the orthonormal system corresponding to the integral equation of the process. In most practical cases $r(s, t)$ is a positive definite kernel so that $\{\varphi_\nu(t)\}$ is a complete system. It may, however, happen that the system is not complete and we then add another orthonormal system $\{\psi_\nu(t)\}$ with $\{\psi\} \perp \{\varphi\}$. It is immediately seen that the y's have a normal distribution with the parameters

$$\begin{cases} E_i y_\nu = \int\limits_a^b m_i(t)\, \psi_\nu(t)\, dt; \quad i = 0, 1. \\ D_i y_\nu = 0; \quad i = 0, 1. \end{cases}$$

If there is an integer ν such that $E_0 y_\nu \neq E_1 y_\nu$ we take as the critical region

$$S = \left\{ \int\limits_a^b x(t)\, \psi_\nu(t)\, dt = E_1 y_\nu \right\},$$

which has $P_0(S) = 0$, $P_1(S) = 1$. *We have thus arrived at the extreme end of the singular case and can determine the true hypothesis from the knowledge of one realization,* if we disregard events of probability zero.

In the following we exclude this case by supposing $E_0 y_\nu = E_1 y_\nu$ for all ν. Then we only have to take the x's into account. They have independent normal distributions with parameters

$$H_0 : \begin{cases} E_0 x_\nu = a_\nu^0 \\ D_0^2 x_\nu = \dfrac{1}{\lambda_\nu} \end{cases} \qquad H_1 : \begin{cases} E_1 x_\nu = a_\nu^1 \\ D_1^2 x_\nu = \dfrac{1}{\lambda_\nu}, \end{cases}$$

where $a_\nu^i = \int\limits_a^b m_i(t)\, \varphi_\nu(t)\, dt$. We have the frequency functions in R_n

$$\begin{cases} g_0(x_1, \ldots x_n) = \dfrac{\sqrt{\lambda_1 \ldots \lambda_n}}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum\limits_1^n \lambda_\nu (x_\nu - a_\nu^0)^2} \\ g_1(x_1, \ldots x_n) = \dfrac{\sqrt{\lambda_1 \ldots \lambda_n}}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum\limits_1^n \lambda_\nu (x_\nu - a_\nu^1)^2} \end{cases}$$

and

$$l_n(\omega) = e^{-\frac{1}{2}\sum\limits_1^n \lambda_\nu (a_\nu^{1\,2} - a_\nu^{0\,2}) - \sum\limits_1^n x_\nu (a_\nu^0 - a_\nu^1)\, \lambda_\nu}.$$

Suppose now that $\sum\limits_1^\infty \lambda_\nu (a_\nu^0 - a_\nu^1)^2 < \infty$. We have, putting

$$z_\nu = -\tfrac{1}{2} \lambda_\nu (a_\nu^{1\,2} - a_\nu^{0\,2}) - \lambda_\nu x_\nu (a_\nu^0 - a_\nu^1)$$

that

$$\begin{cases} E_0 z_\nu = -\frac{1}{2}\lambda_\nu (a_\nu^0 - a_\nu^1)^2 \\ D_0^2 z_\nu = \lambda_\nu (a_\nu^0 - a_\nu^1)^2. \end{cases} \qquad \begin{cases} E_1 z_\nu = \frac{1}{2}\lambda_\nu (a_\nu^0 - a_\nu^1)^2 \\ D_1^2 z_\nu = \lambda_\nu (a_\nu^0 - a_\nu^1)^2. \end{cases}$$

Then using a known theorem of Kolmogoroff we see that $\sum_1^\infty z_\nu$ is converging almost certainly with respect to both P_0 and P_1. *Thus we have the regular case* and almost certainly

$$f(\omega) = \lim_{n\to\infty} e^{\sum_1^n z_\nu}.$$

It is convenient to use another form. Putting

$$f_n(t) = \sum_1^n (a_\nu^0 - a_\nu^1)\lambda_\nu \varphi_\nu(t)$$

we get

$$l_n(\omega) = e^{-\int_a^b f_n(t)\left[x(t) - \frac{m_0(t)+m_1(t)}{2}\right] dt}$$

and using the result of 4.1 *we get the following most powerful region to test H_0 against H_1*

$$\lim_{n\to\infty} \int_a^b f_n(t)\left[x(t) - \frac{m_0(t) + m_1(t)}{2}\right] dt \leq k.$$

Suppose now that $\sum_1^\infty \lambda_\nu (a_\nu^0 - a_\nu^1)^2 = \infty$. Using Tchebychef's inequality and putting

$$t_n(\omega) = \int_a^b f_n(t)\left[x(t) - \frac{m_0(t) + m_1(t)}{2}\right] dt$$

we then have for large n

$$P_0(t_n \leq a) = P_0(t_n - E_0 t_n \leq a - E_0 t_n) \leq P_0(|t_n - E_0 t_n| \geq |a - E_0 t_n|) \leq$$

$$\leq \frac{\sum_1^n \lambda_\nu (a_\nu^0 - a_\nu^1)^2}{\left[a - \frac{1}{2}\sum_1^n \lambda_\nu (a_\nu^0 - a_\nu^1)^2\right]^2}$$

which tends to zero for every a when n tends to infinity. Thus $l_n(\omega) = e^{-t_n(\omega)}$ converges to zero in probability with respect to P_0 when n tends to infinity. In the same way

$$P_1(t_\nu \geq a) = P_1(t_n - E_1 t_n \geq a - E_1 t_n) \leq P_1(|t_n - E_1 t_n| \geq |a - E_1 t_n|) \leq$$

$$\leq \frac{\sum\limits_1^n \lambda_\nu (a_\nu^0 - a_\nu^1)^2}{\left[a + \frac{1}{2}\sum\limits_1^n \lambda_\nu (a_\nu^0 - a_\nu^1)^2\right]^2}$$

which tends to zero for every a when n tends to infinity. Thus $l_n(\omega)$ converges to $+\infty$ in probability with respect to P_1 when n tends to infinity. From this follows as is easily seen by applying the result of 4.2 that we have $P_0(H) = 0$, $P_1(H) = 1$ and thus *again come to the case of extreme singularity.* In the chapter on estimation we will give a more explicit expression for the true hypothesis. It is interesting to note the two distinct ways in which the singular case has been met with. The first one appeared already when only a finite number of the coordinates were taken into account. The second is essentially dependent on the property of convergence of a sequence involving an infinite number of coordinates.

4.5. Continuation; composite alternatives. Now when we are going to consider composite hypotheses, we shall only deal with the regular case. Suppose that we want to test

$$\begin{cases} H_0 : E_0\, x(t) = 0 \\ H_1 : E_\alpha\, x(t) = m_\alpha(t). \end{cases}$$

As $\sum\limits_1^\infty \lambda_\nu a_\nu^2(\alpha) < \infty$, the best critical regions corresponding to the values of α have the form

$$\lim_{n\to\infty} \sum_1^n x_\nu \lambda_\nu a_\nu(\alpha) \geq c(\alpha)$$

which follows from the above. It is easily seen that in order to get a uniformly most powerful region we must have

$$a_\nu(\alpha) = k(\alpha)\, a_\nu$$

where $k(\alpha)$ is of constant sign. If this is true we get the test

$$\lim_{n\to\infty} \int_a^b f_n(t)\, x(t)\, dt \geq c$$

where $f_n(t) = \sum\limits_1^n a_\nu \lambda_\nu \varphi_\nu(t)$ if $k(\alpha) > 0$ and the $\leq$ sign is used if $k(\alpha) < 0$. This test has the character of one-sidedness, and is uniformly most powerful with respect to the alternatives $k(\alpha) > 0$ (or $k(\alpha) < 0$).

We now want to construct a uniformly most powerful unbiased test under the conditions

$$\begin{cases} H_0 : E_0\, x(t) = 0 \\ H_\alpha : E_\alpha\, x(t) = \alpha\, a(t). \end{cases}$$

We get

$$\frac{\partial f(\omega, \alpha)}{\partial \alpha} = e^{-\frac{1}{2}\alpha^2 \sum\limits_1^\infty \lambda_\nu a_\nu^2 + \alpha \sum\limits_1^\infty x_\nu \lambda_\nu a_\nu} \left\{ -\alpha \sum_1^\infty \lambda_\nu a_\nu^2 + \sum_1^\infty x_r \lambda_\nu a_\nu \right\}.$$

Introducing $X = \sum\limits_1^\infty x_\nu \lambda_\nu \alpha_\nu$ which is a normally distributed stochastic variable, we get the domination valid for $|\alpha| < 1$

$$\left| \frac{\partial f(\omega, \alpha)}{\partial \alpha} \right| < k\, e^{|X|} \left\{ \sum_1^\infty a_\nu^2 \lambda_\nu + |X| \right\}.$$

As the last expression has existing mean value, the assumptions made in 4.3 are fulfilled. Using the method described there we get the region

$$e^{-\frac{1}{2} o^2 \alpha^2 + \alpha X} \geq C + c_1 X.$$

Exactly as in the corresponding classical case this gives us the uniformly most powerful unbiased region

$$\left| \lim_{n \to \infty} \int_a^b f_n(t)\, x(t)\, dt \right| \geq k.$$

4.6. Existence and determination of the test-function. As almost certainly $x(t)$ is quadratically integrable, we should get, assuming that $f_n(t)$ converges in the mean to a function $f(t) \in L_2(T)$, that

$$\lim_{n \to \infty} \int_a^b f_n(t)\, x(t)\, dt = \int_a^b f(t)\, x(t)\, dt$$

almost certainly. This would be a very convenient form of the test. This is, however, not always the case. Because of the theorem of Fisher-Riesz it is necessary and sufficient for the convergence in the mean of $f_n(t)$ that $\sum\limits_1^\infty a_\nu^2 \lambda_\nu^2 < \infty$. Assuming this and using the bilinear form of the covariance function we get

$$\int_a^b r(s, t)\, f(s)\, ds = \sum_1^\infty a_\nu \varphi_\nu(t)$$

with uniform convergence for $t \in T$. Of course, we have

$$\int_a^b \varphi_\nu(t) \left[\sum_1^\infty a_\nu \varphi_\nu(t) - a(t) \right] dt = 0$$

for all ν. If $\{\varphi_\nu(t)\}$ is not complete we add as before an orthonormal system $\{\psi_\nu\}$. If for any ν

$$\int_a^b \psi_\nu(t)\left[\sum_1^\infty a_\nu \varphi_\nu(t) - a(t)\right] dt = -\int_a^b \psi_\nu(t)\, a(t)\, dt \neq 0$$

we consider the region

$$S_\alpha = \left\{\int_a^b \psi_\nu(t)\, x(t)\, dt = \alpha \int_a^b \psi_\nu(t)\, a(t)\, dt\right\}_\omega$$

We have immediately $P_\beta(S_\alpha) = 0$ for $\alpha \neq \beta$ and $P_\alpha(S_\alpha) = 1$. As we have excluded the singular case we must have

$$a(t) = \sum_1^\infty a_\nu \varphi_\nu(t)$$

for almost all $t \in T$, and hence

$$\int_a^b r(t, s)\, f(s)\, ds = a(t)$$

for almost all $t \in T$. The left hand side is a continuous function of t because of the properties of the covariance function, and $a(t)$ is also a continuous function being the mean value function of a process which is continuous in the mean. Thus we have equality for all $t \in T$.

If there is a quadratically integrable solution $f(t)$ to this equation, we get immediately

$$\int_a^b f(t)\, \varphi_\nu(t)\, dt = a_\nu \lambda_\nu, \qquad f(t) = \sum_1^\infty a_\nu \lambda_\nu \varphi_\nu(t)$$

so that using the Fisher-Riesz theorem we have $\sum_1^\infty a_\nu^2 \lambda_\nu^2 < \infty$. Thus *for the existence of a quadratically integrable best test-function $f(t)$ it is necessary and sufficient that the equation*

$$\int_a^b r(t, s)\, f(s)\, ds = a(t)$$

has a quadratically integrable solution. As the test-function we take the projection of the solution on the space spanned by $\{\varphi_\nu\}$. *The question considered has thus been reduced to finding a "quellenmässig" representation of the mean value function by means of the covariance function.*

The most interesting case is of course $a(t) = 1$, especially in the stationary case. Let us consider a stationary process $x(t)$. Then it seems intuitively that if the process is of some strongly regular behaviour, e.g. analytic in $t = 0$, which means that $r(t)$ is analytic in $t = 0$, usually no best test of the said simple sort will exist. Using the spectral representation of the covariance function

$$r(t) = \int_{-\infty}^{\infty} e^{it\lambda}\, d\,F(\lambda)$$

we get because of the absolute integrability

$$\int_{a}^{b} r(t-s)\, f(s)\, d\,s = \int_{-\infty}^{\infty} e^{it\lambda}\, \varphi(\lambda)\, d\,F(\lambda)$$

where

$$\varphi(\lambda) = \int_{a}^{b} e^{-it\lambda} f(t)\, d\,t.$$

As $r(z)$ is analytic in $|z| < r$ for some positive r, we know that it is analytic in $|I(z)| < r$ (see LÉVY 1). Then

$$\int_{-\infty}^{\infty} e^{it\lambda}\, \varphi(\lambda)\, dF(\lambda) = 1.$$

for all real t. As $\varphi(0) = \int_{a}^{b} f(t)\, d\,t = \sum_{1}^{\infty} a_\nu^2 \lambda_\nu > 0$, we can write for all real t

$$\int_{-\infty}^{\infty} e^{it\lambda}\, \varphi(\lambda)\, d\,F_1(\lambda) = 0,$$

with

$$F_1(\lambda) = F(\lambda) - \frac{\varepsilon(\lambda)}{\varphi(0)}.$$

Now using the same method of approximation as Karhunen has done in 3, p. 64–65, we get $\varphi(\lambda) = 0$ almost everywhere with respect to $F(\lambda)$ except possibly for $\lambda = 0$. But as $\varphi(\lambda)$ is a non-identically vanishing integral function, it has a discrete set of zeroes, and $F(\lambda)$ must be a step function

$$F(\lambda) = \sum_{\lambda_\nu \leq \lambda} F_\nu$$

where we have put $\lambda_0 = 0$.[1] Thus if the spectrum contains any absolutely continuous or singular component no best test of the above type exists. If it is a pure pointspectrum we form the Hilbert space Λ_2 of functions which are quadratically integrable on (a, b) spanned by the elements

$$\{e^{it\lambda_\nu},\ a \leq t \leq b,\ \nu \neq 0\}.$$

The frequencies λ_ν must not be so dense that Λ_2 includes the constant function 1, because then

$$1 = \underset{n\to\infty}{\mathrm{l.\,i.\,m.}} \sum_{1}^{n} c_\nu^n e^{it\lambda_\nu}$$

[1] The λ's do not denote the eigen-values.

and

$$0 < \int_a^b f(t)\,dt = \lim_{n\to\infty} \sum_1^n c_\nu^n \int_a^b e^{-it\lambda_\nu} f(t)\,dt = 0.$$

If $1 \notin \Lambda_2$ we can write

$$1 = \xi(t) + \eta(t)$$

with $\xi(t) \in \Lambda_2$, $\eta(t) \perp \Lambda_2$ and $\eta(t) \not\equiv 0$. Take $f(t) = \eta(t)$. Then

$$\int_a^b f(t)\,dt = \int_a^b \overline{[\xi(t) + \eta(t)]}\,\eta(t)\,dt = \int_a^b |\eta(t)|^2\,dt > 0$$

and

$$\int_a^b r(t-s)\,f(s)\,ds = \sum_0^\infty F_\nu e^{it\lambda_\nu} \int_a^b e^{-it\lambda_\nu} \eta(t)\,dt = F_0 \int_a^b |\eta(t)|^2\,dt \equiv \text{const.}$$

We shall return to similar questions in connection with estimation.

4.7. Test for the "covariance-factor" in the normal case. Another type of hypotheses for a normal process is the following. Suppose that the mean value function is known, say, identically zero. The covariance function is known but for a multiplicative constant. Put

$$\begin{cases} H_0 : E_0\,x(t) = 0;\; E_0\,x(s)\,x(t) = r(s,t). \\ H_\sigma : E_\sigma\,x(t) = 0;\; E_\sigma\,x(s)\,x(t) = \sigma^2 r(s,t);\; \sigma^2 \neq 1. \end{cases}$$

With the same coordinates as before we get the frequency functions in $R_n(x_1, \ldots x_n)$

$$\begin{cases} g_0(x_1, \ldots x_n) = \dfrac{\sqrt{\lambda_1 \ldots \lambda_n}}{(2\pi)^{\frac{n}{2}}}\, e^{-\frac{1}{2}\sum\limits_1^n x_\nu^2 \lambda_\nu} \\ g_1(x_1, \ldots x_n) = \dfrac{\sqrt{\lambda_1 \ldots \lambda_n}}{(2\pi)^{\frac{n}{2}}\sigma^n}\, e^{-\frac{1}{2}\sum\limits_1^n \lambda_\nu \frac{x_\nu^2}{\sigma^2}} \end{cases}$$

and

$$l_n(\omega) = \frac{1}{\sigma^n}\, e^{-\frac{1}{2}\sum\limits_1^n x_\nu^2 \lambda_\nu \left[\frac{1}{\sigma^2} - 1\right]}$$

If H_0 is true we get, using the fact that $\frac{1}{n}\sum\limits_1^n \lambda_\nu x_\nu^2$ converges almost certainly to 1, which will be shown later,

$$\frac{1}{n}\log l_n(\omega) = -\frac{1}{2n}\sum_1^n x_\nu^2 \lambda_\nu \left[\frac{1}{\sigma^2} - 1\right] - \log\sigma \to \tfrac{1}{2} - \frac{1}{2\sigma^2} - \log\sigma < 0.$$

Thus $l_n(\omega)$ converges against zero with probability one. If H_σ is true we get in the same manner that $l_n(\omega)$ converges against $+\infty$ almost certainly with respect to P_σ. We have the interesting situation that for these hypotheses we always get the singular case. It is possible to get an explicit expression to determine what hypothesis is true. Regard the expression

$$\frac{1}{N}\sum_{1}^{N}\lambda_\nu x_\nu^2 = \frac{1}{N}\sum_{1}^{N} z_\nu^2$$

with

$$\begin{cases} E_0\, z_\nu^2 = 1 \\ D_0^2 z_\nu^2 = 2 \end{cases} \qquad \begin{cases} E_\sigma\, z_\nu^2 = \sigma^2 \\ D_\sigma^2 z_\nu^2 = 2\,\sigma^4. \end{cases}$$

As the z's are independent stochastic variables, we can apply the convergence theorem of Kolmogoroff and get the result that the limit

$$\lim_{N\to\infty}\frac{1}{N}\sum_{1}^{N}\lambda_\nu\left\{\int_a^b x(t)\,\varphi_\nu(t)\,dt\right\}^2$$

exists almost certainly according to both hypotheses and that its value is 1 or σ^2 according to whether H_0 or H_σ is true.

Consider the transformation T_λ operating on the elements of the sample space with

$$T_\lambda\, x(t) = \lambda\, x(t)$$

where λ is a real constant different from one. Then the set $E < \Omega$ for which

$$\lim_{N\to\infty}\frac{1}{N}\sum_{1}^{N}\lambda_\nu\left\{\int_a^b x(t)\,\varphi_\nu(t)\,dt\right\}^2 = 1$$

has P_0-measure one. Evidently the set $T_\lambda E$ is disjoint from E and has P_0-measure zero. Take especially the case $a = 0$, $b = 1$ and $r(s, t) = \min(s, t)$, and we have the time-homogeneous differential normal process, the Wiener process. The surprising result that there exists a set E having the said properties has been shown by Cameron and Martin starting from another point than that of testing statistical hypotheses.

4.8. Several observations. It is now natural to continue the construction of best tests for other types of hypotheses, e.g. to test two given covariance functions against each other, when the mean value function is known, or to consider a composite null hypothesis and try to find similar regions. In practice we have often more knowledge about the type of realization, and using this we can sometimes get simple tests. We want to stress the importance of choosing an appropriate sample space and shall see the advantages of doing this in the following. The difficulty of solving the integral equation of the process can then sometimes be avoided. As the results of the preceding sections have made clear how to proceed in the manner of 4.4–4.7 we shall restrict ourselves to consider only a simple generalization of the above before leaving this topic.

Suppose that we want to test the same hypothesis as before

$$\begin{cases} H_0 : E_0\, x(t) = 0 \\ H_1 : E_1\, x(t) = a(t) \end{cases}$$

when the covariance function $r(s, t)$ is known, but that we now have observed N independent realizations $x_1(t), x_2(t), \ldots x_N(t)$. This sample is naturally described by the coordinates

$$x_{\nu\mu} = \int_a^b x_\nu(t)\, \varphi_\mu(t)\, dt\,; \quad \nu = 1, 2, \ldots N\,; \quad \mu = 1, 2, \ldots.$$

Forming the approximation to the likelihood function, and still supposing the case to be regular, we get

$$l_n(\omega) = e^{-\frac{1}{2} N \sum\limits_1^n \lambda_\mu a_\mu^2 + \sum\limits_{\substack{1 \le \mu \le n \\ 1 \le \nu \le N}} \lambda_\mu x_{\nu\mu} a_\mu}$$

The most powerful test is thus obtained by using as critical region the set where

$$\lim_{n \to \infty} \int_a^b [x_1(t) + \cdots x_N(t)]\, f_n(t)\, dt \ge k$$

where $f_n(t)$ is defined as above. Uniformly most powerful onesided tests and uniformly most powerful unbiased tests can be found in the same way.

4.9. A pointprocess with adjoined variables. The method described in the preceding section leads to integral equations which in practical cases can but occasionally be solved explicitly, though we shall see in 5.3–5.5 that the problem can be dealt with in the most important cases. Though approximate numerical methods are available, it still seems desirable to find tests of simpler structure. As already briefly mentioned, this might be possible when our knowledge of the nature of the realizations allows us to consider more restricted functional spaces than $L_2(T)$.

Let us consider the following process which will also appear in the chapter on estimation. We are observing a Poisson process with intensity β in the time interval $(0, T)$. We get a series of points $0 < t_1 < t_2 < \cdots < t_n < T$. To every interval $t_i < t \le t_{i+1}$ is adjoined a normally distributed stochastic variable x_i with mean value m (which is unknown) and standard deviation 1. Furthermore, these variables are supposed to be independent. This process has been used in applications of stochastic processes to the theory of servomechanisms (see James, Nichols, Phillips 1). The realizations are of the form considered at the end of 3.1, and we choose the same set of coordinates. We want to test the hypothesis $m = 0$ against the alternative value m. Now of course n has a discrete probability distribution, but as mentioned in 4.2, this will add no complication when constructing the test. As is easily seen we get the likelihood function

$$f(\omega) = \frac{e^{-\frac{1}{2}\sum\limits_0^n (x_\nu - m)^2}}{e^{-\frac{1}{2}\sum\limits_0^n x_\nu^2}} = e^{m\sum\limits_0^n x_\nu - \frac{n+1}{2} m^2}$$

The most powerful critical region then has the simple form

$$m \sum_0^n x_\nu - \frac{n+1}{2} m^2 \geq k.$$

This set has the same form as the best test of the mean value of n independent normally distributed stochastic variables $x_1, x_2, \ldots x_n$. We observe, however, that this test could not have been obtained in that manner, because in our case n is not a fixed number but a stochastic variable. On the other hand, one could have got the best conditioned test (by fixing n) of the hypothesis in that way.

To calculate the covariance function which will be needed later we fix $t > s$ and get

$$E[x(s) - m][x(t) - m] = P\{\text{no time point } t_i \text{ in } (s, t)\} \cdot$$
$$\cdot E[x(s) - m]^2 + P\{\text{some time point } t_i \text{ in } (s, t)\} \cdot 0.$$

As the covariance function of a real process is symmetrical in the both arguments, we have

$$r(s, t) = e^{-\beta|t-s|}.$$

It has to be observed that this is not a normal process, which can be seen by considering the simultaneous distribution of $x(s)$ and $x(t)$.

4.10. Tests for pointprocesses. A type of process which is commonly met with in practice is the point-process. We shall study some of these types in connection with test problems. Let $x(t)$ be a generalized Poisson process with probability intensity $\lambda(t)$ or $\mu(t)$ according to whether H_0 or H_1 is true. The process is observed in the time-interval $(0, T)$. Using the coordinates $(n, t_1, \ldots t_n)$ we easily find the likelihood function

$$f(\omega) = \frac{\mu(t_1) \ldots \mu(t_n)}{\lambda(t_1) \ldots \lambda(t_n)} e^{-\int\limits_0^T [\mu(t) - \lambda(t)]\, dt}$$

We get the critical region

$$S = \left\{\prod_1^n \frac{\mu(t_\nu)}{\lambda(t_\nu)} \geq k\right\}.$$

Here we have supposed that $\lambda(t)$ is different from zero almost everywhere in the set where $\mu(t) \neq 0$. Otherwise we should have got a singular part H which does not appear now, as no question of convergence turns up (we have $P(n < \infty) = 1$).

We now restrict the alternative hypotheses to be of the form

$$\begin{cases} H_0 : \text{the prob. intensity is } \lambda(t) \\ H_\mu : \text{the prob. intensity is } \mu\,\lambda(t). \end{cases}$$

Then we get the critical region

$$S = \{\mu^n \geq k\},$$

and if we confine ourselves to the alternatives $\mu > 1$, $(\mu < 1)$ we get the one-sided uniformly most powerful test

$$S = \{n \geq n_0\} \qquad (\text{or } S = \{n \leq n_0\}).$$

Because n has a discrete distribution, it may be impossible to solve n_0 exactly from the equation

$$P_0(S) = \varepsilon$$

with an arbitrary choice of ε, but as this question is of little practical interest, we will not deal with it here.

To get a uniformly most powerful unbiased test we put according to 4.3

$$\mu^n \geq c + c_1 n,$$

and because of the convexity of the exponential function the critical region has the form

$$S = \{n < n_1\} + \{n > n_2\}.$$

Also here it might be impossible to solve the equations determining n_1 and n_2 exactly, but due to the same reason as above we dismiss the question.

Another choice of hypotheses is the following. Let

$$\begin{cases} H_0 \; : \lambda(t) = e^{\alpha t} \\ H_\mu : \mu(t) = e^{(\alpha+\mu)t}. \end{cases}$$

The most powerful critical region S is easily obtained

$$S = \{e^{\mu \sum\limits_1^n t_i} \geq k\}.$$

For the one-sided alternative $\mu > 0$, we thus get the uniformly most powerful test

$$S = \left\{\sum_1^n t_i \geq k'\right\}.$$

Here $\sum\limits_1^n t_i$ has a continuous distribution so that we have no difficulty in determining k for a given value of ε.

Suppose now that $x(t)$ is a Pólya process (see LUNDBERG 1) defined in $(0, T)$ with parameter β. Then the conditional probability intensity when n events have taken place up to the time t is given by

$$P_n(t) = \frac{1 + \beta n}{1 + \beta t}.$$

We want to test whether $\beta = 0$ against the simple alternative value $\beta > 0$. To do this we form the expression for the probability element corresponding to the coordinates $(n, t_1, \ldots t_n)$ for $n > 0$.

$$g_\beta(\omega) = \prod_0^{n-1} \frac{1 + \beta \nu}{1 + \beta t_{\nu+1}} e^{-\int\limits_0^{t_1} \frac{dt}{1+\beta t} - \int\limits_{t_1}^{t_2} \frac{1+\beta}{1+\beta t} dt - \cdots - \int\limits_{t_n}^{T} \frac{1+n\beta}{1+\beta t} dt}.$$

We thus have

$$\log g_\beta(\omega) = -\frac{1}{\beta} \log (1 + \beta t_1) - \frac{1+\beta}{\beta} \log \frac{1 + \beta t_2}{1 + \beta t_1} - \cdots - \frac{1 + n\beta}{\beta} \log \frac{1 + \beta T}{1 + \beta t_n} +$$

$$+ \sum_0^{n-1} \log \frac{1 + \beta \nu}{1 + \beta t_{\nu+1}} = -\frac{1}{\beta} \log (1 + \beta T) - \sum_1^{n} \nu \log \frac{1 + \beta t_{\nu+1}}{1 + \beta t_\nu} + \sum_0^{n-1} \log \frac{1 + \beta \nu}{1 + \beta t_{\nu+1}}$$

where we have put $t_{n+1} = T$. Reducing the second term we get

$$\log g_\beta(\omega) = -\frac{1}{\beta} \log (1 + \beta T) + \sum_1^{n} \log (1 + \beta t_\nu) - n \log (1 + \beta T) +$$

$$+ \sum_0^{n-1} \log \frac{1 + \beta \nu}{1 + \beta t_{\nu+1}} = -\frac{1}{\beta} \log (1 + \beta T) + \sum_0^{n-1} \log (1 + \beta \nu) - n \log (1 + \beta T).$$

This expression is valid for $n > 0$, and for $n = 0$ we get immediately

$$\log g_\beta(\omega) = -\frac{1}{\beta} \log (1 + \beta T).$$

For the hypothesis $\beta = 0$ one gets

$$g_0(\omega) = e^{-T},$$

and hence

$$f(\omega) = \frac{g_\beta(\omega)}{g_0(\omega)} = \frac{c(\beta)}{(1 + \beta T)^n} \prod_0^{n-1} (1 + \beta \nu),$$

where the product shall be assigned the value one for $n = 0$. We get the most powerful region

$$S = \{f(\omega) \geq k\}$$

and as $\log \prod_{0}^{n-1} (1 + \beta \nu)$ is a convex function of n, while $n \log (1 + \beta T)$ is a linear one, the best region gets the form

$$S = \{n < n_1\} + \{n > n_2\}$$

with $n_1 < n_2$. This implies that we reject H_0 when we find very small or very large values of n, which intuitively seems to be in accordance with the fact that

$$D_\beta^2 n = T(1 + \beta T) > T = D_0^2 n.$$

It is of interest that the position of the time-points of the events are not used in the best test. Hence it follows that if we observe only n, we can make just as strong statements as when we are considering $t_1, t_2, \ldots t_n$ also.

4.11. The stationary Markoff process. In 4.4—4.7 we have seen how to construct best tests for the mean value of a normal process. Unfortunately the test functions only occasionally get the simple form $\int_a^b f(t)\, x(t)\, dt$, but one can sometimes obtain simple test functions by specializing the sample space in an appropriate manner. This is true for the perhaps most important type of normal processes, viz. the stationary Markoff process. Let $x(t)$ be such a process with mean value m and covariance function

$$r(s, t) = e^{-\beta |t-s|}.$$

We want to test the hypothesis $H_0 : m = 0$ against a simple alternative value of m to begin with. We shall show that the corresponding test functions $f_n(t)$ do not converge to a function in $L_2(T)$. We take $T = (0, 1)$. The kernel $r(s, t)$ is positive definite which can be seen by the same argument as we shall use to show the divergence of $f_n(t)$. Then, if $f_n(t)$ converges in the mean to a function $f(t) \in L_2(T)$, we must have

$$\int_0^1 e^{-\beta |t-s|} f(t)\, dt = \int_0^s e^{\beta (t-s)} f(t)\, dt + \int_s^1 e^{\beta (s-t)} f(t)\, dt \equiv 1$$

for all $s \in T$. Then for almost all s

$$0 = -\beta e^{-\beta s} \int_0^s e^{\beta t} f(t)\, dt + \beta e^{\beta s} \int_s^1 f(t)\, e^{-\beta t}\, dt,$$

and by subtraction

$$1 = 2 e^{-\beta s} \int_0^s f(t)\, e^{\beta t}\, dt$$

and as both sides are continuous, this holds for every $s \in T$. We get

$$f(s) = \frac{\beta}{2}$$

for almost all $s \in T$. But this function does not satisfy our integral equation. Now we shall show that it is still possible to get a simple best test.

To this end we choose the set of continuous functions in $(0, T)$ as sample space, which is possible according to 1.4. As coordinates we use $x(t_n)$, $n = 1, 2, \ldots$, where $\{t_n\}$ is a denumerable sequence of points everywhere dense in $(0, T)$. It is convenient to choose $t_1 = 0$, $t_2 = T$ and the following as $\frac{1}{2}T$, $\frac{1}{4}T$, $\frac{3}{4}T, \ldots$ and so on by repeated dichotomy. We want to calculate the frequency function of $x(t_1), x(t_2), \ldots x(t_n)$ under the hypothesis H_m. To do this we rearrange the t's in their natural order, and using this new numbering we define

$$\begin{cases} x(t_i) = x_i \\ e^{-\beta(t_{i+1}-t_i)} = \varrho_i. \end{cases} \qquad i = 1, 2, \ldots n.$$

We then get

$$g_m^{(n)}(\omega) = \frac{1}{(2\pi)^{\frac{n}{2}} \prod\limits_1^{n-1} (1-\varrho_i^2)^{\frac{1}{2}}} e^{-\frac{1}{2}(x_1-m)^2 - \frac{1}{2}\sum\limits_1^{n-1} \frac{[x_{i+1}-\varrho_i x_i - m(1-\varrho_i)]^2}{1-\varrho_i^2}}$$

and

$$\log \frac{g_m^{(n)}(\omega)}{g_0^{(n)}(\omega)} = m x_1 - \frac{m^2}{2} + m \sum_1^{n-1} \frac{x_{i+1} - \varrho_i x_i}{1+\varrho_i} - \frac{m}{2} \sum_1^{2\,n-1} \frac{1-\varrho_i}{1+\varrho_i} =$$

$$= m x_1 - \frac{m^2}{2} + m \sum_2^{n-1} \frac{x_i}{1+\varrho_{i-1}} - m \sum_2^{n-1} \frac{\varrho_i x_i}{1+\varrho_i} +$$

$$+ m \frac{x_n}{1+\varrho_{n-1}} - m \frac{\varrho_1 x_1}{1+\varrho_1} - \frac{m^2}{2} \sum_1^{n-1} \frac{1-\varrho_i}{1+\varrho_i}.$$

If n is large $\Delta_n = \max\limits_i (t_{i+1} - t_i)$ is small, and we get

$$\log l_n(\omega) = m x(0) - \frac{m^2}{2} + m \frac{x(T)}{1+\varrho_{n-1}} - m \frac{\varrho_1 x(0)}{1+\varrho_1} +$$

$$+ \frac{m}{2} \sum_2^{n-1} x_i \left\{ \frac{\beta}{2}(t_i - t_{i-1}) + \frac{\beta}{2}(t_{i+1} - t_i) + O(\Delta^2) \right\} - \frac{m^2}{2} \sum_2^{n-1} \beta \frac{t_{i+1} - t_i + O(\Delta^2)}{2}.$$

As $\Delta_n \rightarrow 0$ and as the realizations are continuous, we get almost certainly

$$\log f(\omega) = \lim_{n \to \infty} \log l_n(\omega) = m \frac{x(0)}{2} + m \frac{x(T)}{2} - \frac{m^2}{2} + \frac{m}{2} \beta \int_0^T x(t)\, dt - \frac{m^2}{4} \beta T.$$

We thus have the regular case, and in the same way as before *we get the following simple form for the uniformly most powerful test of H_0 against the one-sided alternative $m > 0$*

$$x(0) + x(T) + \beta \int_0^T x(t)\,dt \geq k.$$

The uniformly most powerful unbiased test is obtained in the same way. We shall later return to this process in connection with a problem of estimation.

4.12. Approximation of tests. As has already been seen, it may happen that the form of the best test-function is too complicated to be used in practical applications. Then one has to construct a simpler but less powerful test, or, if a test of high power is required, one can approximate to the best test-function in some appropriate way. Suppose for simplicity that we have the regular case. Then

$$l_n(\omega) = \frac{g_1(x_1, \ldots x_n)}{g_0(x_1, \ldots x_n)}$$

converges to the likelihood function $f(\omega)$ in probability. The best region is

$$S = \{f(\omega) \geq k\}_\omega,$$

and we can use the following approximation

$$S_n = \{l_n(\omega) \geq k\}_\omega.$$

For these regions we have

$$P_i(S_n) \to P_i(S); \quad i = 0, 1;$$

which means that the errors of the first and second kinds corresponding to S_n can be made as near as desired to those corresponding to the best region S. Choosing n sufficiently large, we thus have a test which from the practical point of view differs little from the best possible.

We now leave the problem of testing statistical hypotheses regarding stochastic processes. To continue the construction of practical tests it seems important to consider the demands of the applications.

The problem of estimation

5.1. Unbiased estimates. Suppose now that a process $x(t)$ has one of the distributions P_α, where α is a real parameter in a finite interval $A = (a, b)$. To avoid the singular case we suppose further that for every pair $\alpha_1, \alpha_2 \in A$ there does not exist any set S with $P_{\alpha_1}(S) = 0$, $P_{\alpha_2}(S) \neq 0$. Using the knowledge given by a realization of the process we want to decide which of the hypotheses is true, i.e. we want to form a function $t(\omega)$ estimating α. We have for an arbitrary set S

$$P_\alpha(S) = \int_S f(\omega, \alpha)\,dP_0(\omega)$$

where P_0 is the measure corresponding to a_0, a fixed value of a. We now regard the likelihood function as a stochastic process given by the random function $f(\omega, a)$ with a in the rôle of a time-parameter, and with a measure of probability P_0. The mean value of this process is easily obtained

$$E_0 f(\omega, a) = \int_{\Omega} f(\omega, a)\, d P_0(\omega) = P_a(\Omega) = 1.$$

We now make the natural assumption that $f(\omega, a)$ has a finite variance and that it is continuous in the mean. Putting $f(\omega, s) = f(s)$, we have the covariance function

$$\varrho(s, t) = E_0 [f(s) - 1][f(t) - 1].$$

Form the usual integral equation corresponding to this kernel and denote its eigen-values by λ_ν and its eigen-functions by $\varphi_\nu(a)$. We then have for every $a \in A$

$$f(\omega, a) = \underset{n \to \infty}{\mathrm{l.\,i.\,m.}} \sum_1^n \varphi_\nu(a) \frac{\psi_\nu(\omega)}{\sqrt{\lambda_\nu}} + 1,$$

where $\{\psi_\nu(\omega)\}$ is an orthonormal system in $L_2(\Omega)$, and the convergence is taken with respect to P_0.

We shall now study the existence of unbiased estimates of minimum variance, i.e. functions $t(\omega)$ satisfying

$$\begin{cases} E_a t(\omega) = a \\ E_a t(\omega)^2 < \infty \end{cases}$$

for every $a \in A$. If the system $\{\psi_\nu\}$ is not complete in $L_2(\Omega)$ with respect to P_0, we add its orthogonal complement

$$L_2(\Omega) \ominus \{\psi_\nu\} = \{\psi'_\mu\},$$

where $\{\psi'_\mu\}$ is an orthonormal system. By means of the systems $\{\psi_\nu\}$ and $\{\psi'_\mu\}$ we can develop an arbitrary estimate with the above properties in a series converging in the mean (with respect to P_0)

$$t(\omega) = \sum_1^\infty t_\nu \psi_\nu(\omega) + \sum_1^\infty t'_\mu \psi'_\mu(\omega).$$

We thus obtain

$$a = E_a t(\omega) = \int_{\Omega} t(\omega) f(\omega, a)\, d P_0(\omega) = \sum_1^\infty t_\nu \frac{\varphi_\nu(a)}{\sqrt{\lambda_\nu}} + E_0 t(\omega).$$

The convergence is uniform because

$$\left| \sum_n^\infty t_\nu \frac{\varphi_\nu(a)}{\sqrt{\lambda_\nu}} \right|^2 \leq \sum_n^\infty t_\nu^2 \sum_n^\infty \frac{\varphi_\nu(a)^2}{\lambda_\nu} \leq \sum_n^\infty t_\nu^2 \sum_1^\infty \frac{\varphi_\nu(a)^2}{\lambda_\nu}$$

and as

$$\sum_1^\infty \frac{\varphi_\nu(a)^2}{\lambda_\nu} = \varrho(a, a)$$

which is a continuous function of a, we obtain the stated result. Thus we can integrate termwise and get

$$\gamma_\nu = \int_a^b (a - a_0)\, \varphi_\nu(a)\, d\,a = \frac{t_\nu}{\sqrt{\lambda_\nu}}.$$

In order that an unbiased estimate of finite variance shall exist, it is thus necessary that $\sum_1^\infty \lambda_\nu \gamma_\nu^2 < \infty$.

Suppose for simplicity that $\{\varphi_\nu(a)\}$ is complete in $L_2(A)$. Then the condition $\sum_1^\infty \lambda_\nu \gamma_\nu^2 < \infty$ is also a sufficient condition for the existence of an unbiased estimate. For, consider

$$t(\omega) = \sum_1^\infty \sqrt{\lambda_\nu}\, \gamma_\nu\, \psi_\nu(\omega) + \sum_1^\infty t'_\mu\, \psi'_\mu(\omega)$$

where the t'_μ are arbitrary real numbers but for the condition $\sum_1^\infty t'^2_\mu < \infty$. This series converges in the mean with respect to P_0. We get immediately

$$E_\alpha t(\omega) = \sum_1^\infty t_\nu\, \varphi_\nu(a) + E_0\, t(\omega).$$

Using the same method as before, we can show that $\sum_1^\infty t_\nu\, \varphi_\nu(a) = a - a_0$ for all $a \in A$. Thus $t(\omega) + a_0 - E_0\, t(\omega)$ is an unbiased estimate of a. Summing up we have:

In order that an unbiased estimate with finite variance shall exist it is necessary that

$$\sum_1^\infty \lambda_\nu \left\{ \int_a^b (a - a_0)\, \varphi_\nu(a)\, d\,a \right\}^2 < \infty.$$

If the system $\{\varphi_\nu\}$ is complete, this condition implies the existence of a family of unbiased estimates whose dimensionality is given by Dim $L_2(\Omega) \ominus \{\psi_\nu\}$.

If there is more than one unbiased estimate, it seems not unnatural to choose the one for which $D_0(t)$ is minimum, if one has reason to believe that the true value of a lies in the neighbourhood of a_0. In some cases it might happen that in the class of unbiased estimates there is one with minimum variance for all a.

The method described has certain theoretical advantages and could be extended, but it is not quite suitable for applications because of which we shall try to find other ways of estimation in the following sections. The following simple example gives a procedure that in some special cases may be applied. Consider the Poisson process of 4.10 with a constant probability intensity β. We want to find an unbiased estimate β^* of β which has minimum variance. Putting $\beta^* = \beta^* (n, t_1, \ldots t_n)$ we shall have

$$\beta = E_\beta \beta^* = \sum_0^\infty P(n=\nu) E_\beta [\beta^* \mid n = \nu] = \sum_0^\infty \frac{e^{-\beta t}(\beta t)^\nu}{\nu !} E_\beta [\beta^* \mid n=\nu],$$

i.e.

$$\beta e^{\beta t} = \sum_0^\infty \frac{(\beta t)^\nu}{\nu !} E_\beta [\beta^* \mid n = \nu]$$

valid for β in a certain interval. The conditional frequency function for $t_1, t_2, \ldots t_n$ when n is known is

$$\frac{e^{-\beta t}\beta^n}{\frac{e^{-\beta t}(\beta t)^n}{n!}} = \frac{n!}{t^n} \qquad (0 < t_1 < t_2 < \cdots < t_n < t)$$

so that $E_\beta [\beta^* \mid n = \nu]$ is independent of β. As $\beta e^{\beta t}$ is an integral function of β, we get equating the Taylor coefficients

$$E_\beta [\beta^* \mid n = \nu] = \frac{\nu}{t}\cdot$$

But

$$E_\beta (\beta^* - \beta)^2 = \sum_0^\infty P_n E_\beta [(\beta^* - \beta)^2 \mid n] = \sum_0^\infty P_n \left\{ E_\beta \left[\left(\beta^* - \frac{n}{t}\right)^2 \middle| n \right] + \right.$$

$$\left. + 2 E_\beta \left[\left(\beta^* - \frac{n}{t}\right)\left(\frac{n}{t} - \beta\right) \middle| n \right] + E_\beta \left[\left(\frac{n}{t} - \beta\right)^2 \middle| n \right] \right\}\cdot$$

As

$$E_\beta \left[\left(\beta^* - \frac{n}{t}\right) \middle| n \right] = 0,$$

the second term vanishes. We thus get the unique unbiased estimate with minimum variance by choosing β^* so that the first term (which otherwise would be > 0) vanishes

$$\beta^* = \frac{n}{t}\cdot$$

5.2. A class of linear estimates. The approach of the preceding section demands complete specification of the probability distributions P_α. It often happens that we do not want or are not able to specify the distributions completely. It may still be possible to find unbiased estimates of minimum variance in some restricted class of estimates. Consider e.g. the following situation, where we want to estimate the mean value m of a stochastic process

$x(t)$ which is supposed to be continuous in the mean with covariance function $r(s, t)$. Regard the class of linear estimates

$$m^* = \int_a^b f(t)\, x(t)\, dt$$

where $f(t)$ is quadratically integrable in (a, b). We either take the integral in the sense of 1.3 or choose a sample-space such that $x(t)$ is quadratically integrable in the sense of Doob. To get an unbiased estimate in this class we must demand that

$$\int_a^b f(t)\, dt = 1.$$

The variance is then

$$E\,(m^* - m)^2 = \int_a^b \int_a^b r(s, t)\, f(s)\, f(t)\, ds\, dt.$$

Introducing the eigen-values λ_ν and the eigen-functions φ_ν of the integral equation of the process (we suppose that $\{\varphi_\nu\}$ forms a complete set in $L_2(a, b)$) and using the bilinear expression of the covariance function we get

$$E\,(m^* - m)^2 = \sum_1^\infty \frac{c_\nu^2}{\lambda_\nu},$$

where

$$c_\nu = \int_a^b f(t)\, \varphi_\nu(t)\, dt, \quad \nu = 1, 2, \ldots .$$

This should be minimized subject to the condition

$$\sum_1^\infty c_\nu a_\nu = 1,$$

where

$$a_\nu = \int_a^b \varphi_\nu(t)\, dt.$$

But using the Schwarz' inequality we obtain, if $\sum\limits_1^\infty a_\nu^2 \lambda_\nu < \infty$

$$1 = \left[\sum_1^\infty c_\nu a_\nu\right]^2 \leq \left[\sum_1^\infty |c_\nu a_\nu|\right]^2 \leq \sum_1^\infty \frac{c_\nu^2}{\lambda_\nu} \sum_1^\infty a_\nu^2 \lambda_\nu,$$

i.e.

$$E\,(m^* - m)^2 \geq \frac{1}{\sum\limits_1^\infty a_\nu^2 \lambda_\nu},$$

where the sign of equality is realized by

$$c_\nu = \frac{a_\nu \lambda_\lambda}{\sum_1^\infty a_\nu^2 \lambda_\nu}.$$

Put

$$f_N(t) = \frac{\sum_1^N a_\nu \lambda_\nu \varphi_\nu(t)}{\sum_1^N a_\nu^2 \lambda_\nu}$$

and

$$m_N^* = \int_a^b x(t) f_N(t)\, dt = m + \frac{\sum_1^N a_\nu \sqrt{\lambda_\nu}\, x_\nu}{\sum_1^N a_\nu^2 \lambda_\nu},$$

where $\{x_\nu\}$ is a set of non-correlated stochastic variables with mean value zero and standard deviation 1. When N tends to infinity, *this sequence of estimates evidently converges in the mean to an estimate m^* which is unbiased and of minimum variance $\frac{1}{\sum_1^\infty a_\nu^2 \lambda_\nu}$ in the class under consideration* (see GRENANDER 1). If so desired one can extract a subsequence converging almost certainly. It should be noted that the limit of this sequence is not always of the type introduced above. This obnoxious property of this class of estimates will be dealt with later.

If on the other hand $\sum_1^\infty a_\nu^2 \lambda_\nu = \infty$, it is easily seen that there is a subsequence $m_{N_\nu}^*$ converging almost certainly to the true value m when ν tends to infinity. *Thus we are able to state, though we know only the covariance function of the process, that if $\sum_1^\infty a_\nu^2 \lambda_\nu = \infty$ we get the singular case.* We have seen in 4.4 that in the normal case this is also necessary if we only consider positive definite covariance functions.

As seen in 4.6, the convergence of $f_n(t)$ to a function in $L_2(T)$ implies the existence of a quadratically integrable solution of the equation

$$\int_a^b r(s, t)\, f(t)\, dt = 1.$$

This being seldom the case, we are naturally led to consider the following form of linear estimates

$$m^* = \int_a^b x(t)\, d\, F(t)$$

where $F(t)$ is a function of bounded variation, and the integral is interpreted in some appropriate sense (e.g. that of Karhunen).

We demand analogously to the above that

$$\begin{cases} E\, m^* = m \int_a^b d\, F(t) = m \\ E\,(m^* - m)^2 = \int_a^b \int_a^b r(s, t)\, d\, F(s)\, d\, F(t) = min. \end{cases}$$

Suppose that $F(t)$ satisfies these conditions and let α and β be two points in (a, b). If

$$G = \varepsilon\,(t - \alpha) - \varepsilon\,(t - \beta)$$

and δ is a real number, the weight function $F(t) + \delta\, G(t)$ gives an unbiased estimate, as we have

$$\int_a^b d\,[F(t) + \delta\, G(t)] = 1.$$

Further we have, denoting

$$\int_a^b r(s, t)\, d\, F(t) = R(s)$$

that

$$\int_a^b \int_a^b r(s, t)\, d\,[F(s) + \delta\, G(s)]\, d\,[F(t) + \delta\, G(t)] = \int_a^b \int_a^b r(s, t)\, d\, F(s)\, d\, F(t) + \\ + 2\,\delta \int_a^b R(t)\, d\, G(t) + \delta^2 \int_a^b \int_a^b r(s, t)\, d\, G(s)\, d\, G(t).$$

As this is to be larger than $\int_a^b \int r(s, t)\, d\, F(s)\, d\, F(t)$ for all δ, we must have

$$\int_a^b R(t)\, d\, G(t) = R(\alpha) - R(\beta) = 0,$$

so that

$$R(s) = \int_a^b r(s, t)\, d\, F(t) \equiv c$$

for $s \in T$. Evidently the minimum variance is just the constant figuring in the right member.

Suppose on the other hand that $F(t)$ satisfies this integral equation and that

$\int_a^b dF(t) = 1$. If $H(t)$ is another function of bounded variation in (a, b) with $\int_a^b dH(t) = 1$, we have, putting $H = F + G$

$$\int_a^b dG(t) = \int_a^b dH(t) - \int_a^b dF(t) = 0,$$

and

$$\int_a^b \int_a^b r(s, t)\, dH(s)\, dH(t) = \int_a^b \int_a^b r(s, t)\, dF(s)\, dF(t) + \\ + 2\int_a^b \int_a^b r(s, t)\, dF(s)\, dG(t) + \int_a^b \int_a^b r(s, t)\, dG(s)\, dG(t).$$

The last term is non-negative, and further

$$\int_a^b \int_a^b r(s, t)\, dF(s)\, dG(t) = c \int_a^b dG(t) = 0.$$

Thus we have

$$D^2(m_H^*) \geq D^2(m_F^*).$$

We have previously seen that the covariance function $e^{-\beta|t-s|}$ (in the normal case corresponding to a stationary Markoff process) does not admit any best test function in $L_2(T)$ for the mean. But by considering the equation

$$\int_0^T e^{-\beta|t-s|}\, dF(t) = \frac{2}{2 + \beta T}$$

it is easily verified that the function of bounded variation

$$F(t) = \frac{\varepsilon(t) + \varepsilon(t - T) + \beta t}{2 + \beta T}$$

satisfies the equation considered. Thus

$$m^* = \frac{x(0) + x(T) + \beta \int_0^T x(t)\, dt}{2 + \beta T}$$

is an unbiased estimate of m with minimum variance in the class of estimates which has been considered. We shall later on see that, if the process is normal, this estimate is the best one out of a larger class of estimates.

Another case is obtained by considering a time-homogeneous orthogonal process $x(t)$ in the time-interval (a, b) with constant though unknown mean value m and covariance function $r(s, t) = \min(s, t)$. The equation

$$\int_a^b \min(s, t)\, dF(t) = c$$

evidently has the solution

$$F(t) = \varepsilon(t - a)\frac{c}{a}$$

so that we get the best estimate

$$m^* = x(a).$$

5.3. The equidistributed estimate. When the process is stationary, the problem dealt with in 5.2 shows some special features. If the spectral energy in $\lambda = 0$ is zero it is known that the estimate

$$m^* = \frac{1}{2T}\int_{-T}^{T} x(t)\, dt,$$

which is a priori unbiased, converges in probability to m (see 1.3), i.e. m^* is a consistent estimate. This estimate which will be called the equidistributed estimate of m has another optimal property, which shall be studied in this section. The result obtained seems to be capable of generalization which the author wishes to consider in a later publication.

Suppose that the spectrum is absolutely continuous with a spectral density $h(\lambda)$ which is continuous at the origin and bounded. Consider unbiased estimates of the type

$$m_T^* = \int_{-T}^{T} f(t)\, x(t)\, dt; \quad \int_{-T}^{T} f(t)\, dt = 1.$$

We get, putting

$$f(t) = \frac{1}{T} g\left(\frac{t}{T}\right)$$

that

$$m_T^* = \int_{-1}^{1} g(u)\, x(uT)\, du; \quad \int_{-1}^{1} g(u)\, du = 1.$$

Thus $g(u)$ measures the relative weight given to different values of t. Confining ourselves to the regular class of unbiased estimates obtained, when $g(u)$ belongs to a class C of functions, which are uniformly continuous and uniformly bounded in $(-1,1)$, *we shall see that the equidistributed estimate is of minimum variance asymptotically in this class when T tends to infinity.* More precisely, we define the efficiency of the estimate

$$\mu_T^* = \frac{1}{2T}\int_{-T}^{T} x(t)\, dt$$

as

$$e_T = \frac{\inf\limits_{m^* \in C} D^2 m_T^*}{D^2 \mu_T^*};\quad 0 \leq e_T \leq 1.$$

This concept of efficiency is different from the one used in the classical theory, as it takes only the linear properties of the process into consideration. We shall show that e_T tends to unity as T tends to infinity. Put $\overline{\lim\limits_{T\to\infty}}\, e_T = e$. Then there is a sequence $T_\nu \to \infty$ and corresponding estimates $m_{T_\nu}^* \in C$ so that

$$\frac{D^2\, m_{T_\nu}^*}{D^2\, \mu_{T_\nu}^*} \to e$$

as $\nu \to \infty$. Introduce the functions

$$\gamma_\nu(\lambda) = \int\limits_{-1}^{1} e^{iu\lambda}\, g_\nu(u)\, du.$$

Then

$$D^2(m_{T_\nu}^*) = \frac{1}{T_\nu^2} \int\limits_{-T_\nu}^{T_\nu} \int\limits_{-T_\nu}^{T_\nu} r(s,t)\, g_\nu\left(\frac{s}{T_\nu}\right) g_\nu\left(\frac{t}{T_\nu}\right) ds\, dt =$$

$$= \int\limits_{-\infty}^{\infty} |\gamma_\nu(T_\nu \lambda)|^2\, h(\lambda)\, d\lambda = \frac{1}{T_\nu} \int\limits_{-\infty}^{\infty} |\gamma_\nu(\mu)|^2\, h\left(\frac{\mu}{T_\nu}\right) d\mu.$$

For the equidistributed estimate one gets in the same way

$$D^2\, \mu_{T_\nu}^* = \frac{1}{T_\nu} \int\limits_{-\infty}^{\infty} \frac{\sin^2 \mu}{\mu^2}\, h\left(\frac{\mu}{T_\nu}\right) d\mu$$

and using a property of the Fejér kernel,

$$T_\nu D^2\, \mu_{T_\nu}^* \to \pi\, h(0)$$

as ν tends to infinity. Because of the uniform continuity of the g's it is possible to choose a subsequence converging to a continuous function $g(u)$. Supposing this to be already done we get

$$\int\limits_{-1}^{1} |g(u) - g_\nu(u)|^2\, du \to 0$$

as $\nu \to \infty$, because of the theorem of Lebesgue on bounded convergence. Thus, using the theorem of Plancherel

$$\int\limits_{-\infty}^{\infty} |\gamma_\nu(\mu) - \gamma(\mu)|^2\, d\mu \to 0,$$

and

$$\int_{-\infty}^{\infty}\left|\gamma_\nu(\mu)\sqrt{h\left(\frac{\mu}{T_\nu}\right)}-\gamma(\mu)\sqrt{h\left(\frac{\mu}{T_\nu}\right)}\right|^2 d\mu \leq H\int_{-\infty}^{\infty}|\gamma_\nu(\mu)-\gamma(\mu)|^2 d\mu \to 0$$

as $\nu \to \infty$. Hence

$$\int_{-\infty}^{\infty}|\gamma_\nu(\mu)|^2 h\left(\frac{\mu}{T_\nu}\right)d\mu - \int_{-\infty}^{\infty}|\gamma(\mu)|^2 h\left(\frac{\mu}{T_\nu}\right)d\mu \to 0.$$

But

$$\left|\int_{-\infty}^{\infty}|\gamma(\mu)|^2 h\left(\frac{\mu}{T_\nu}\right)d\mu - h(0)\int_{-\infty}^{\infty}|\gamma(\mu)|^2 d\mu\right| \leq$$

$$\leq \left|\int_{|\mu|<\varepsilon T_\nu}|\gamma(\mu)|^2\left[h\left(\frac{\mu}{T_\nu}\right)-h(0)\right]d\mu\right| + \left|\int_{|\mu|\geq\varepsilon T_\nu}|\gamma(\mu)|^2\left[h\left(\frac{\mu}{T_\nu}\right)-h(0)\right]d\mu\right|$$

and choosing ε so small that $|h(\mu)-h(0)|<\delta$ for $|\mu|<\varepsilon$, and then T_ν so large that $\int_{|\mu|\geq\varepsilon T_\nu}|\gamma(\mu)|^2 d\mu<\delta$, we get

$$\int_{-\infty}^{\infty}|\gamma_\nu(\mu)|^2 h\left(\frac{\mu}{T_\nu}\right)d\mu \to h(0)\int_{-\infty}^{\infty}|\gamma(\mu)|^2 d\mu$$

as $\nu \to \infty$. But according to Plancherel's theorem

$$\int_{-\infty}^{\infty}|\gamma(\mu)|^2 d\mu = 2\pi\int_{-1}^{1}|g(u)|^2 du$$

and thus, using the Schwarz' inequality

$$e = 2\int_{-1}^{1}|g(u)|^2 du \geq \left\{\int_{-1}^{1}|g(u)|\,du\right\}^2 \geq \left\{\int_{-1}^{1}g(u)\,du\right\}^2 = 1.$$

According to the definition of e only the equality sign is possible, and we have thus proved our proposition.

Corollary: The most important consequence of this seems to be that *it is impossible to get any asymptotically more efficient estimate of m than the equidistributed one by constructing estimates of the type*

$$m^* = \frac{1}{T}\int_{-T}^{T}g\left(\frac{t}{T}\right)x(t)\,dt$$

where $\int_{-1}^{1} g(t)\,dt = 1$, *and* $g(u)$ *is a function defined in* $(-1, 1)$ *not depending upon* T. The above proof shows further that the best of these estimates actually is the equidistributed one.

It has to be noted that it is not possible to remove the condition that $g_\nu \in C$ without restricting in some way the type of the process. Take e.g. the process with correlation function $e^{-\frac{t^2}{2}}$. If a realization is observed in a non-degenerate interval we know the realization for all values of t because of the fact that the process is analytic for all t. Then we can form $\frac{1}{2A}\int_{-A}^{A} x(t)\,dt$ which, as $A \to \infty$, tends to m in the mean because the spectrum is continuous. Thus the equidistributed estimate has efficiency zero.

5.4. Doob's elementary processes. We have seen that in order to test the mean value of a process it is not sufficient to consider estimates of the type $\int x(t)\,f(t)\,dt$ where $f(t)$ is quadratically integrable, but we have had to introduce Stieltjes integrals. We shall see that in other cases the best test is not even of the Stieltjes integral type. Because of this it is appropriate to consider this problem from another point of view.

We observe the process $y(t) = m + x(t)$ during the time (a, b) and suppose that $E\,x(t) \equiv 0$. It is required to find an unbiased linear estimate of m with minimum variance. If m^* is an unbiased estimate and

$$m^* = \underset{n\to\infty}{\mathrm{l.\,i.\,m.}}\; m_n^*;\;\; m_n^* = \sum_1^n e_\nu^{(n)}\, y\,(t_\nu^{(n)});\;\; t_\nu^{(n)} \in (a, b);$$

we have

$$E\,m_n^* = m \sum_1^n c_\nu^{(n)} \to m \quad \text{as} \quad n \to \infty.$$

Thus we can write

$$m^* = \underset{n\to\infty}{\mathrm{l.\,i.\,m.}} \left\{ \sum_1^n c_\nu^{(n)}\, y\,(t_\nu^{(n)}) + \left[1 - \sum_1^n c_\nu^{(n)}\right] y\,(a) \right\}$$

because the quantity in rectangular brackets tends to zero as n tends to infinity. In this way every unbiased linear estimate can be regarded as the limit of finite sums

$$\sum_1^n c_\nu^{(n)}\, y\,(t_\nu^{(n)}) \quad \text{with} \quad \sum_1^n c_\nu^{(n)} = 1.$$

Consider the set $M_0 < L_2\,(X;\; a, b)$ consisting of all elements of the form

$$\sum_1^n c_\nu\, x\,(t_\nu) \quad \text{with} \quad \sum_1^n c_\nu = 1,\;\; t_\nu \in (a, b).$$

Closing this set with respect to convergence in the mean we obtain a set $M < L_2(X;\ a, b)$. As this set is closed and convex there exists at least one element μ^* with

$$\| \mu^* \| = \inf_{z \in M} \| z \|.$$

There cannot be more than one such element for suppose μ_1^* is another one. Then also $\frac{\mu^* + \mu_1^*}{2} \in M$ and

$$\left\| \frac{\mu^* + \mu_1^*}{2} \right\|^2 = \frac{\| \mu^* \|^2}{4} + \frac{\| \mu_1^* \|^2}{4} + 2\, Re \frac{E\, \mu^*\, \bar{\mu}_1^*}{4} < \| \mu^* \|^2$$

which contradicts the definition of μ^*. Thus *there is a uniquely determined estimate of minimum variance in the class of unbiased linear estimates.*

Denote this estimate by m^* and regard the expression $E\, m^* \overline{x(t)}$ as $a \leq t \leq b$. If

$$E\, m^* \overline{x(\alpha)} \neq E\, m^* \overline{x(\beta)};\ a \leq \alpha,\ \beta \leq b;$$

we introduce

$$m_1^* = m^* + \varepsilon\, [x(\alpha) - x(\beta)]$$

which is also unbiased. Then

$$\| m_1^* - m \|^2 = \| m^* - m \|^2 + 2\, R_e \{ \bar{\varepsilon}\, E\, [m^* - m]\, \overline{[x(\alpha) - x(\beta)]} \} + \\ + | \varepsilon |^2 \| x(\alpha) - x(\beta) \|^2$$

and it is possible to choose ε such that $D\, m_1^* < D\, m^*$ contrary to the definition of m^*. *Hence the function $E\, m^* \overline{x(t)}$ is a constant for t in the interval (a, b)*

$$E\, m^* \overline{x(t)} = c;\ t \in (a, b).$$

But as $m^* = \underset{n \to \infty}{\mathrm{l.\,i.\,m.}}\ m_n^*$ with

$$m_n^* = \sum_1^n c_\nu^{(n)}\, y\,(t_\nu^{(\nu)})$$

we get

$$D^2\, m^* = \| m^* - m \|^2 = \lim_{n \to \infty} E\, m^*\, \overline{m_n^* - m} = c \sum_1^n \bar{c}_\nu^{(n)} = c$$

so that *the constant c is equal to the variance of the estimate m^*.*

The solution of this equation always gives us the uniquely determined unbiased estimate of minimum variance. For if

$$\left. \begin{aligned} E\, m^* \overline{x(t)} &= c \\ E\, m_1^* \overline{x(t)} &= c_1 \end{aligned} \right\};\ a \leq t \leq b;$$

we see that if $c = c_1$ the two estimates coincide. If $c \neq c_1$ we can suppose that $c \neq 0$ and we then have

$$m_1^* = \frac{c_1}{c} m^*$$

and as both estimates are unbiased we get $c_1 = c$.

Let us apply the above to the important class of stationary processes introduced by DOOB (5). We shall deal with the non-deterministic type of these. The correlation function can then be written as

$$r(t) = \int_{-\infty}^{\infty} e^{it\lambda} \frac{d\lambda}{|a_n (i\lambda)^n + a_{n-1} (i\lambda)^{n-1} + \cdots + a_0|^2}$$

where the denominator has its zeroes λ_ν in the upper half plane and the coefficients a_ν are real

$$a_n (i\lambda)^n + \cdots + a_0 = a_n i^n \prod_1^n (\lambda - \lambda_\nu).$$

This process can also be obtained as a solution of a linear differential equation with constant coefficients (see KARHUNEN 2). For $n = 1$ we get a correlation function of the type $e^{-\beta |t|}$.

It is immediately seen that the process has strong derivatives up to the order $n - 1$. Consider the estimate (where $x(t)$ is the observed process)

$$m^* = \frac{\sum_0^{n-1} \{\alpha_\nu x^{(\nu)}(0) + \beta_\nu x^{(\nu)}(T)\} + a_0 \int_0^T x(s)\, ds}{2a_1 + a_0 T}$$

where

$$\left.\begin{aligned} \alpha_\nu &= (-1)^\nu a_{\nu+1} \\ \beta_\nu &= a_{\nu+1} \end{aligned}\right\}; \ \nu = 0, 1, \ldots n - 1.$$

This is possible because

$$i a_1 = (-1)^{n-1} i^n a_n \sum_1^n \frac{\lambda_1 \ldots \lambda_n}{\lambda_\nu} = - a_0 \sum_1^n \frac{1}{\lambda_\nu} = - a_0 \sum_1^n \frac{\bar{\lambda}_\nu}{|\lambda_\nu|^2}$$

so that $i \frac{a_1}{a_0} (a_0 \neq 0)$ has positive imaginary part, i.e. $\frac{a_1}{a_0} > 0$. The estimate is, of course, unbiased and we shall show that it has minimum variance. Consider the expression

$$E\, m^* \overline{x(t) - m} = \frac{\sum_0^{n-1} \{\alpha_\nu r^{(\nu)}(-t) + \beta_\nu r^{(\nu)}(T - t)\} + a_0 \int_0^T r(s - t)\, dt}{2a_1 + a_0 T}.$$

But

$$a_n \frac{d^n r(t)}{d t^n} + a_{n-1} \frac{d^{n-1} r(t)}{d t^{n-1}} + \cdots + a_0 = \int_{-\infty}^{\infty} e^{i t \lambda} \frac{a_n (i \lambda)^n + \cdots + a_0}{|a_n (i \lambda)^n + \cdots + a_0|^2} d \lambda =$$

$$= \int_{-\infty}^{\infty} e^{i t \lambda} \frac{d \lambda}{a_n (-i)^n \prod_{1}^{n} (\lambda - \bar{\lambda}_\nu)}$$

which according to Cauchy's theorem is equal to zero for $t > 0$. For $t < 0$ one gets in the same way

$$a_n \frac{d^n r(t)}{d t^n} - a_{n-1} \frac{d^{n-1} r(t)}{d t^{n-1}} + - \cdots (-1)^n a_0 r(t) = 0.$$

Thus

$$a_0 \int_0^T r(s-t)\, d s = a_0 \lim_{\varepsilon \to 0} \left\{ \int_0^{t-\varepsilon} r(s-t)\, d s + \int_{t+\varepsilon}^{T} r(s-t)\, d s \right\} =$$

$$= \lim_{\varepsilon \to 0} \{ a_1 [r(-\varepsilon) - r(-t)] - a_2 [r'(-\varepsilon) - r'(-t)] + \cdots (-1)^{n+1} a_n [r^{(n-1)}(-\varepsilon)$$

$$- r^{(n-1)}(-t)] - a_1 [r(T-t) - r(\varepsilon)] - a_2 [r'(T-t) - r'(\varepsilon)] - \cdots - a_n [r^{(n-1)}(T-t)$$

$$- r^{(n-1)}(\varepsilon)]\} = 2 a_1 r(0) + 2 a_3 r''(0) + \cdots + 2 a_\mu r^{(\mu-1)}(0)$$

$$- a_1 r(-t) - a_1 r(T-t) + a_2 r'(-t) - a_2 r'(T-t) + - \cdots$$

$$+ (-1)^n a_n r^{(n-1)}(-t) - a_n r^{(n-1)}(T-t).$$

Here we have put $\mu = n$ if n is odd, otherwise $\mu = n - 1$. We have

$$E\, m^* \overline{x(t) - m} = \frac{2 a_1 r(0) + 2 a_2 r''(0) + \cdots + 2 a_\mu r^{(\mu-1)}(0)}{2 a_1 + a_0 T} \equiv \text{const.}$$

As the right hand member does not depend upon t we know that *m^* is the unique unbiased estimate of minimum variance.* To find the variance we only have to calculate the constant. We get

$$2 a_1 r(0) + 2 a_3 r''(0) + \cdots 2 a_\mu r^{(\mu-1)}(0) = \lim_{A \to \infty} a_0 \int_{-A}^{A} r(t)\, d t$$

by integrating the two differential equations for $r(t)$ between $-A$ and A and letting A tend to infinity because the functions $r(t), r'(t), \ldots r^{(n-1)}(t)$ tend to zero as t tends to infinity according to Lebesgue's theorem on Fourier coefficients. Hence the variance is

$$\frac{a_0\, 2 \pi \frac{1}{a_0^2}}{2 a_1 + a_0 T} = \frac{2 \pi}{a_0 (2 a_1 + a_0 T)}.$$

5.5. Purely non-deterministic processes. We have seen that if the process can be completely extrapolated when we know the realization on an interval of length A, the equidistributed estimate has efficiency zero if the length of the interval of observation exceeds A. To avoid this it seems natural to consider the purely non-deterministic processes (see HANNER 1 and KARHUNEN 4). Then the spectrum is absolutely continuous and there is a quadratically integrable function $f(\lambda)$ with

$$|f(\lambda)|^2 = F'(\lambda)$$

and such that the function $g(a)$ defined by

$$g(a) = \underset{A\to\infty}{\text{l. i. m.}} \frac{1}{\sqrt{2\pi}} \int_{-A}^{A} e^{ia\lambda} f(\lambda)\, d\lambda$$

vanishes for negative arguments. We shall suppose that for small values of λ

$$f(\lambda) = f_0 + f_1 \lambda (1 + 0(\lambda))$$

where f_0 is real and different from zero.

Consider functions of λ with the inner product

$$(\varphi, \psi) = \int_{-\infty}^{\infty} \varphi(\lambda) \overline{\psi(\lambda)}\, dG(\lambda) \quad \text{where} \quad G(\lambda) = F(\lambda) + \varepsilon(\lambda).$$

The function

$$H(\lambda) = \frac{e^{-iT\lambda} \dfrac{1}{\overline{f(\lambda)}} - e^{iT\lambda} \dfrac{1}{f(\lambda)}}{\lambda}$$

clearly has a finite norm. The functions $e^{it\lambda}$; $-T \leq t \leq T$; also have finite norms. The Hilbert space spanned by these will be denoted by $\lambda_2(T)$. Put

$$H(\lambda) = e^{-iT\lambda} \frac{\dfrac{1}{\overline{f(\lambda)}} - \dfrac{1}{f(0)}}{\lambda} + \frac{e^{-iT\lambda} - e^{iT\lambda}}{f(0)\lambda} + e^{iT\lambda} \frac{\dfrac{1}{f(0)} - \dfrac{1}{f(\lambda)}}{\lambda} = H_1(\lambda) + H_2(\lambda) + H_3(\lambda),$$

where H_1, H_2 and H_3 have finite norms and $H_2 = \dfrac{1}{i f(0)} \displaystyle\int_{-T}^{T} e^{i\lambda t}\, dt \in \lambda_2(T)$. Put

$$\begin{cases} H^*_{1,T} = P_{\lambda_2(T)} H_1 \\ H^*_{3,T} = P_{\lambda_2(T)} H_3 \end{cases}$$

and

$$H_T^*(\lambda) = H_{1,T}^*(\lambda) + H_2(\lambda) + H_{3,T}^*(\lambda) \in \lambda_2(T).$$

If

$$x(t) = \int_{-\infty}^{\infty} e^{it\lambda} d Z(\lambda)$$

then

$$y(t) = m + x(t) = \int_{-\infty}^{\infty} e^{it\lambda} d Z'(\lambda) \quad \text{where} \quad Z'(\lambda) = Z(\lambda) + m\,\varepsilon(\lambda).$$

As we have observed $y(t)$ in the interval $(-T, T)$ we can form

$$m_0^* = \int_{-\infty}^{\infty} H_T^*(\lambda)\, d Z'(\lambda).$$

For if

$$H_T^*(\lambda) = \underset{n\to\infty}{\text{l.i.m.}} \sum_1^n c_\nu^{(n)} e^{i t_\nu^{(n)} \lambda}; \quad -T \leq t_\nu^{(n)} \leq T;$$

we get

$$m_0^* = \underset{n\to\infty}{\text{l.i.m.}} \sum_1^n c_\nu^{(n)} y(t_\nu^{(n)}).$$

We have

$$E\, m_0^* = m\, H_T^*(0) = m\,[H_{1,T}^*(0) + H_2(0) + H_{3,T}^*(0)].$$

But the projection $H'_{1,T}(\lambda)$ of

$$\frac{1}{\lambda}\left[\frac{1}{\overline{f(\lambda)}} - \frac{1}{f(0)}\right]$$

on the space spanned by $\{e^{i(T+t)\lambda};\ -T \leq t \leq T;\}$ is $e^{iT\lambda} H_{1,T}^*(\lambda)$ and as T tends to infinity $H'_{1,T}(\lambda)$ tends to a limit $H'_{1,\infty}(\lambda)$. Thus

$$(H_1; 1) = (H_{1,T}^*; 1) = (e^{-iT\lambda} H'_{1,T}(\lambda) - e^{-iT\lambda} H'_{1,\infty}(\lambda); 1) + (e^{-iT\lambda} H'_{1.\infty}; 1).$$

When T tends to infinity the left member

$$(H_1; 1) = \int_{-\infty}^{\infty} H_1(\lambda)\, d G(\lambda) = \int_{-\infty}^{\infty} \frac{f(\lambda) - \dfrac{|f(\lambda)|^2}{f(0)}}{\lambda} e^{-iT\lambda} d\lambda + H_1(0) \to H_1(0) = \frac{\bar{f}_1}{f_0^2}$$

and the first term of the right member

$$|(e^{-iT\lambda} H'_{1,T} - e^{-iT\lambda} H'_{1,\infty}; 1)| \leq \|1\| \cdot \|H'_{1,T} - H'_{1,\infty}\| \to 0.$$

The second term

$$(e^{-iT\lambda} H'_{1,\infty}(\lambda); 1) = \int_{-\infty}^{\infty} H'_{1,\infty}(\lambda)\, e^{-iT\lambda} |f(\lambda)|^2 d\lambda + H'_{1,\infty}(0) \to H'_{1,\infty}(0)$$

from which follows

$$\lim_{T\to\infty} H_{1,T}^*(0) = \lim_{T\to\infty} H'_{1,T}(0) = H'_{1,\infty}(0) = H_1(0).$$

In the same way it is shown that

$$\lim_{T\to\infty} H_{3,T}^{*}(0) = H_3(0).$$

Hence

$$m^* = \frac{\int\limits_{-\infty}^{\infty} H_T^*(\lambda)\, d Z'(\lambda)}{H_{1,T}^*(0) + \dfrac{2T}{i f(0)} + H_{3,T}^*(0)},$$

where the denominator is different from zero when T is sufficiently large, is an unbiased estimate. We shall show that it has minimum variance.

We have

$$E\, m^* \overline{x(t)} = E\,[m^* - m]\, \overline{x(t)}$$

and

$$\left[H_{1,T}^*(0) + \frac{2T}{i f(0)} + H_{3,T}^*(0)\right] E\,[m^* - m]\,\overline{x(t)} =$$

$$= \int\limits_{-\infty}^{\infty} H_T^*(\lambda)\, e^{-it\lambda} F'(\lambda)\, d\lambda = \int\limits_{-\infty}^{\infty} H_T^*(\lambda)\, e^{-it\lambda}\, d\, G(\lambda) - H_T^*(0) =$$

$$= \int\limits_{-\infty}^{\infty} H(\lambda)\, e^{-it\lambda}\, d\, G(\lambda) - H_T^*(0) = \int\limits_{-\infty}^{\infty} e^{-it\lambda} H(\lambda)\, F'(\lambda)\, d\lambda + H(0) - H_T^*(0).$$

But putting

$$n(t) = \int\limits_{-\infty}^{\infty} H(\lambda)\, e^{-it\lambda} F'(\lambda)\, d\lambda$$

we get for $-T \leq t_1,\ t_2 \leq T$

$$n(t_2) - n(t_1) = \int\limits_{-\infty}^{\infty} \frac{e^{i(-T-t_2)\lambda} - e^{i(-T-t_1)\lambda}}{\lambda} f(\lambda)\, d\lambda - \int\limits_{-\infty}^{\infty} \frac{e^{i(T-t_2)\lambda} - e^{i(T-t_1)\lambda}}{\lambda} \overline{f(\lambda)}\, d\lambda.$$

The first term is according to Plancherel's theorem

$$i \int\limits_{-T-t_1}^{-T-t_2} g(a)\, d a = 0$$

because

$$\begin{cases} -T \leq t_2 \\ -T \leq t_1. \end{cases}$$

The other term is also zero because

$$\begin{cases} t_2 \leq T \\ t_1 \leq T. \end{cases}$$

Then

$$n(t) = C(T) \quad \text{when} \quad -T \leq t \leq T$$

and hence m^* has minimum variance. In the same way it is shown that $C(T)$ is independent of T. To calculate the variance we regard

$$C = \int_{-\infty}^{\infty} H(\lambda) F'(\lambda)\, d\lambda = 2\, i \int_{-\infty}^{\infty} \cos T\lambda \frac{I f(\lambda)}{\lambda}\, d\lambda - 2\, i \int_{-\infty}^{\infty} \frac{\sin T\lambda}{\lambda} Re f(\lambda)\, d\lambda.$$

The first integral tends to zero as T tends to infinity (it is the Fourier-coefficient of an integrable function) and the second one tends to $-2\, i\, \pi f(0)$. The variance of the unbiased estimate of minimum variance is then

$$D^2 m^* = \frac{-2\, i f(0)\, \pi + H(0) - H_T^*(0)}{H_{1,T}^*(0) + \dfrac{2T}{i f(0)} + H_{3,T}^*(0)} \sim \frac{\pi f(0)^2}{T} = \frac{\pi F'(0)}{T}.$$

Hence we have shown that *for the purely non-deterministic process the equidistributed estimate is asymptotically efficient without having restricted the class of estimates as in 5.3.*

5.6. Efficiency of estimates. In this and the following sections we shall deal with the method of maximum likelihood. Though it is possible to transfer this method to the case of a stochastic process with a continuous time-parameter, some very interesting and essential complications will turn up. A first step in the direction of solving these problems will be taken in 5.7—5.9.

We still suppose that for a in the finite interval A under consideration we have the regular case, and that $f(\omega, a)$ almost certainly has a derivative which is dominated

$$\left| \frac{\partial f(\omega, a)}{\partial a} \right| < F(\omega)$$

where $F(\omega)$ is a stochastic variable of finite variance with respect to P_0, and further that

$$E_a \left(\frac{\partial \log f(\omega, a)}{\partial a} \right)^2 < \infty.$$

Considering estimates $a^*(\omega)$ of finite variance and using theorem 15.1 in Saks 1, *we get an expression of the minimum variance which is analogous to the one obtained in the finite dimensional case* (see Cramér 4).

If $b(a)$ is the bias of the estimate, we have

$$a + b(a) = E_a\, a^* = E_0 \left[a^*(\omega) f(\omega, a)\right]$$

and

$$1 + \frac{d\, b(a)}{d\, a} = E_0 \left[a^*(\omega) \frac{\partial f(\omega, \alpha)}{\partial a} \right]$$

because the integrand is less than $|a^*(\omega)|\,F(\omega)$ in absolute value, and this majorant is according to Schwarz' inequality integrable with respect to P_0. In the same way we get

$$1 = E_a 1 = E_0 f(\omega, a)$$

and

$$0 = E_0\left(\frac{\partial f(\omega, a)}{\partial a}\right).$$

Thus

$$\left(1 + \frac{d b}{d a}\right)^2 = \left[E_0 (a^* - a)\frac{\partial f(\omega, a)}{\partial a}\right]^2 \leq$$

$$\leq E_0\left[(a^* - a)^2 f(\omega, a)\right] E_0\left[\left(\frac{\partial \log f(\omega, a)}{\partial a}\right)^2 f(\omega, a)\right] = E_a (a^* - a)^2 E_a\left(\frac{\partial \log f}{\partial a}\right)^2,$$

which gives us

$$E_a (a^* - a)^2 \geq \frac{\left(1 + \frac{d b}{d a}\right)^2}{E_a\left(\frac{\partial \log f(\omega, a)}{\partial a}\right)^2}.$$

Passing by we want to point out the formal similarity of this result when using coordinates of the type used in the study of point-processes with adjoined variables, to a theorem on minimum variance of sequential estimates by WOLFOWITZ 1.

As is easily seen we obtain the equality sign in the above formula if and only if

$$\frac{\partial \log f(\omega, a)}{\partial a} = k(a)\left[a^*(\omega) - a\right].$$

In the same way as in the classical case we thus see that if there is an efficient estimate it is obtained as the unique, non-identically constant, solution of the maximum likelihood-equation.

Consider now the estimation problem studied in 5.2, and suppose we have the normal, regular case. We get

$$f(\omega, m) = e^{-\frac{m^2}{2}\sum\limits_1^\infty a_\nu^2 \lambda_\nu + m \sum\limits_1^\infty a_\nu \lambda_\nu x_\nu}$$

which satisfies the regularity conditions. If m^* is an unbiased estimate of finite variance of m, we thus have

$$D_m^2 m^* \geq \frac{1}{E_m\left[\sum\limits_1^\infty a_\nu \lambda_\nu x_\nu - m \sum\limits_1^\infty \lambda_\nu a_\nu^2\right]^2} = \frac{1}{\sum\limits_1^\infty a_\nu^2 \lambda_\nu}$$

But as

$$\frac{\partial \log f(\omega, m)}{\partial m} = m \left\{ \frac{\sum_1^\infty a_\nu \lambda_\nu x_\nu}{\sum_1^\infty a_\nu^2 \lambda_\nu} - m \right\} \sum_1^\infty a_\nu^2 \lambda_\nu$$

we therefore have an efficient estimate

$$m^* = \frac{\sum_1^\infty a_\nu \lambda_\nu x_\nu}{\sum_1^\infty a_\nu^2 \lambda_\nu}$$

as is easily verified by direct calculation.

We are especially interested in the case of a stationary Markoff process. Then

$$f(\omega, m) = e^{\frac{m}{2}\left[x(0)+x(T)+\beta \int_0^T x(t)\,dt\right] - \frac{m^2}{2}\left(1+\frac{\beta T}{2}\right)}$$

and in the same way we obtain the estimate

$$m^* = \frac{x(0) + x(T) + \beta \int_0^T x(t)\,dt}{2 + \beta T}$$

which is unbiased and of minimum variance. Hence it is the best in this sense in the class of all estimates of finite variance.

5.7. The method of maximum likelihood. It is now possible to prove properties of estimates by methods quite similar to those of the finite dimensional case: e.g. if two estimates of the same parameter are efficient, they coincide with probability one. Also the case of several parameters can be treated in the same way. On the other hand, we shall encounter some difficulties when trying to apply the maximum likelihood method to stochastic processes. But the following result is easily obtained:

Suppose that the conditions 1—3 are satisfied.

1) $\frac{\partial^\nu \log f(\omega, \alpha)}{\partial \alpha^\nu}$; $\nu = 1, 2, 3$; exist almost certainly.

2) For every $\alpha \in A$ we shall have $\left|\frac{\partial f}{\partial \alpha}\right| < F_1(\omega)$, $\left|\frac{\partial^2 f}{\partial \alpha^2}\right| < F_2(\omega)$ and $\left|\frac{\partial^3 \log f}{\partial \alpha^3}\right| < H(\omega)$ where $E_0 F_1 < \infty$, $E_0 F_2 < \infty$ and $E_\alpha H < k$.

3) For every $\alpha \in A$ $E_\alpha \left(\frac{\partial \log f}{\partial \alpha}\right)^2$ shall be positive and finite.

We consider the case when we have observed N independent realizations of the process and want to estimate α. Denoting the realizations by $\omega_1, \omega_2, \dots \omega_N$ we form the simultaneous likelihood function

$$f(\omega_1, \omega_2, \dots \omega_N; \alpha) = f(\omega_1; \alpha) f(\omega_2; \alpha) \dots f(\omega_N; \alpha).$$

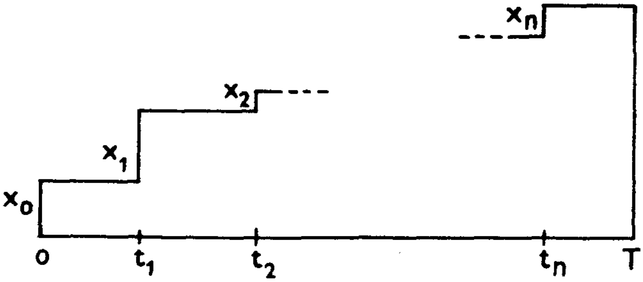
Then it is possible to show that the likelihood-equation

$$\frac{\partial \log f(\omega_1, \omega_2, \dots \omega_N; \alpha)}{\partial \alpha} = 0$$

has a solution $\alpha^*(\omega_1, \dots \omega_N)$ which is a consistent, asymptotically normal and asymptotically efficient estimate of α as N tends to infinity.

This is proved in the same manner as in CRAMÉR 4 p. 500–503.

As an application of this we consider the stochastic process used in EINSTEIN 1 to describe the movement of pebbles on the bottom of a water channel. The stone which is observed during the time interval $(0, T)$, has two possible states: either it rests or moves. The movement takes place without loss of time. At $t = 0$ the stone is supposed to be in movement. The other time points of movement $t_1, t_2, \dots t_n$ are distributed according to a Poisson process with probability intensity $\frac{1}{\vartheta}\cdot$ As is easily seen, ϑ is the mean resting time. In the movement at the time instants $0, t_1, \dots t_n$ the stone is transported over lengths $x_0, x_1, \dots, x_n$. Here the x's are independent stochastic variables taking positive values with the frequency function $\frac{1}{\xi} e^{-\frac{x}{\xi}}$. ξ is the mean length of transport. N stones are independently observed. We want to estimate ϑ.



We get the probability element

$$\frac{1}{\xi} e^{-\frac{x_1}{\xi}} dx_0 \frac{1}{\vartheta} e^{-\frac{t_1}{\vartheta}} dt_1 \dots \frac{1}{\vartheta} e^{-\frac{t_n - t_{n-1}}{\vartheta}} dt_n \frac{1}{\xi} e^{-\frac{x_n}{\xi}} dx_n e^{-\frac{T-t_n}{\vartheta}} =$$

$$= \xi^{-(n+1)} e^{-\frac{X}{\xi}} \vartheta^{-n} e^{-\frac{T}{\vartheta}} dx_0 \dots dx_n dt_1 \dots dt_n,$$

where we have put $X = x_0 + x_1 + \dots + x_n$. Labelling the coordinates of the different realizations with the index $i = 1, 2, \dots N$, we get the simultaneous probability element

$$\xi^{-\sum\limits_1^N (n_i+1)} e^{-\sum\limits_1^N \frac{X_i}{\xi}} \vartheta^{-\sum\limits_1^n n_i} e^{-\frac{N T}{\vartheta}}$$

leaving out the differentials. From this it is possible to calculate the distribution of X. We shall need only the first four moments, which, using the usual symbols, are

$$\begin{cases} \alpha_1 = \xi\left(1 + \frac{T}{\vartheta}\right) \\ \alpha_2 = \xi^2\left(2 + 4\frac{T}{\vartheta} + \frac{T^2}{\vartheta^2}\right) \\ \alpha_3 = \xi^3\left(6 + 18\frac{T}{\vartheta} + 9\frac{T^2}{\vartheta^2} + \frac{T^3}{\vartheta^3}\right) \\ \alpha_4 = \xi^4\left(24 + 96\frac{T}{\vartheta} + 72\frac{T^2}{\vartheta^2} + 16\frac{T^3}{\vartheta^3} + \frac{T^4}{\vartheta^4}\right). \end{cases} \qquad \begin{cases} \mu_2 = \xi^2\left(1 + 2\frac{T}{\vartheta}\right) \\ \mu_3 = \xi^3\left(2 + 6\frac{T}{\vartheta}\right) \\ \mu_4 = \xi^4\left(9 + 36\frac{T}{\vartheta} + 12\frac{T^2}{\vartheta^2}\right). \end{cases}$$

The maximum likelihood estimate is easily obtained

$$\vartheta^* = \frac{N T}{\sum\limits_1^N n_i}.$$

If $\sum\limits_1^N n_i = 0$, we get $\vartheta^* = \infty$, but when N tends to infinity, the probability of this tends to zero. This detail is of small practical importance. We have

$$\sqrt{N}(\vartheta^* - \vartheta) = -\vartheta \frac{\frac{1}{\sqrt{N}} \sum\limits_1^N \left(n_i - \frac{T}{\vartheta}\right)}{\frac{1}{N} \sum\limits_1^N n_i},$$

where n_i has a Poisson distribution with mean value $\frac{T}{\vartheta}\cdot$ Because of the central limit law $\frac{1}{\sqrt{N}} \sum\limits_1^N \left(n_i - \frac{T}{\vartheta}\right)$ is asymptotically normal with mean value zero and standard deviation $\sqrt{\frac{T}{\vartheta}}$. As $\frac{1}{N} \sum\limits_1^N n_i$ converges to $\frac{T}{\vartheta}$ in probability when N tends to infinity, we have (using theorem 20.6 in Cramér 4) that ϑ^* is asymptotically normal with mean value ϑ and variance $\frac{1}{N}\frac{\vartheta^3}{T}$ for large values of N. To calculate the efficiency we form

$$E_\vartheta\left(\frac{\partial \log f(\omega, \vartheta)}{\partial \vartheta}\right)^2 = E_\vartheta\left(\frac{n}{\vartheta} - \frac{T}{\vartheta^2}\right)^2 = \frac{T}{\vartheta^3}$$

and thus find that ϑ^* is asymptotically efficient.

Our observable coordinates have been the time points and the lengths of transport, but in the best estimate only the number of time-points is used. In EINSTEIN 1 $X_1, X_2, \ldots X_N$ are the only variables that have been observed. Presumably it would have been practically impossible to observe the numbers n_i, but as this might be possible in other applications of this process, it is of some interest to investigate the loss of efficiency when only observing the X's. Einstein uses the method of moments to estimate ϑ and gets (p. 38)

$$. \vartheta_0^* = T\left\{\frac{1}{\sqrt{1 - \frac{m_2}{a_1^2}}} - 1\right\},$$

where

$$\begin{cases} a_1 = \dfrac{1}{N}\sum\limits_1^N X_i \\ m_2 = \dfrac{1}{N}\sum\limits_1^N X_i^2 - a_1^2. \end{cases}$$

Regarding $\vartheta_0^* - \vartheta$ as a function $H(a_1, m_2)$ of the sample moments, we obtain

$$\begin{cases} H(\alpha_1, \mu_2) = 0 \\ \left(\dfrac{\partial H}{\partial a_1}\right)_0 = -\dfrac{\vartheta^3}{\xi T^2}(1 + 2\theta) \\ \left(\dfrac{\partial H}{\partial m_2}\right)_0 = \dfrac{\vartheta^3}{2\xi^2 T^2}(1 + \theta) \end{cases} \qquad \begin{cases} \mu_2(a_1) \sim \dfrac{\xi^2(1 + 2\theta)}{N} \\ \mu_{11}(a_1, m_2) \sim \dfrac{\xi^3(2 + 6\theta)}{N} \\ \mu_2(m_2) \sim \dfrac{\xi^4(8 + 32\theta + 8\theta^2)}{N} \\ \theta = \dfrac{T}{\vartheta}, \end{cases}$$

where the index 0 is used to denote the value obtained when putting $a_1 = \alpha_1$, $\mu_2 = m_2$, in the expression in the brackets. Using theorem 28.4 in CRAMÉR 4, we see that $\vartheta_0^* - \vartheta$ is asymptotically normal with mean value zero and with variance

$$\frac{\vartheta^6}{NT^4}(1 + 6\theta + 10\theta^2 + 8\theta^3 + 2\theta^4).$$

Then the efficiency is

$$e(\vartheta_0^*) = \frac{\theta^3}{1 + 6\theta + 10\theta^2 + 8\theta^3 + 2\theta^4}.$$

The following similar process has been met in connection with a problem of medical statistics. We investigate the occurrence of a phenomenon A in an interval which we denote as usual with $(0, T)$ although the parameter does not denote time but location. At $t = 0$ we suppose that A^* always occurs. As the sample-space we take functions taking the values 0 (for A^*) and 1 (for A) and having the value 0 for $t = 0$. As coordinates we take the numbers $n_1, n_2, t_1, t_2, \ldots t_{n_1+n_2}$, where the t's denote the time-points when the states change and n_1 denotes the number of starting points for the state A and n_2 the number of end points for A-intervals. Evidently $n_1 = n_2$ or $n_1 = n_2 + 1$ according as the state at $t = T$ is A^* or A. We suppose that the length of an A^*- (or A-)interval has the frequency function $\beta e^{-\beta t}$, $t > 0$ (or $\alpha e^{-\alpha t}$, $t > 0$) with independence in the usual manner between different intervals. We get the probability element

$$e^{-\beta t_1} \beta \, d t_1 \, e^{-\alpha (t_2 - t_1)} \alpha \, d t_2 \ldots e^{-\alpha (t_{n_1+n_2} - t_{n_1+n_2-1})} \alpha \, d t_{n_1+n_2} e^{-\beta (T - t_{n_1+n_2})} =$$

$$= e^{-\beta l'} \beta^{n_1} e^{-\alpha l''} \alpha^{n_2} d t_1 \ldots d t_{n_1+n}$$

if $n_1 = n_2$ and similarly

$$e^{-\beta t_1} \beta \, d t_1 \ldots e^{-\alpha (T - t_{n_1+n_2})} = e^{-\beta \lambda'} \beta^{n_1} e^{-\alpha \lambda''} \alpha^{n_2} d t_1 \ldots d t_{n_1+n_2}$$

if $n_1 = n_2 + 1$. Here we have put

$$\begin{cases} l' = t_1 + t_3 - t_2 + \cdots + T - t_{n_1+n_2} \\ l'' = t_2 - t_1 + \cdots t_{n_1+n_2} - t_{n_1+n_2-1} \end{cases} \qquad \begin{cases} \lambda' = t_1 + \cdots t_{n_1+n_2} - t_{n_1+n_2-1} \\ \lambda'' = t_2 - t_1 + \cdots + T - t_{n_1+n_2} . \end{cases}$$

Introducing the total length of the A^*-intervals L, we see that $L = l'$ in the first case and $L = l''$ in the second case. The analogous is true about the total length A of the A-intervals. In both cases we thus have the probability elements

$$\beta^{n_1} e^{-\beta L} \alpha^{n_2} e^{-\alpha A} \, d t_1 \ldots d t_{n_1+n_2} .$$

Repeating this experiment N times independently we get the maximum likelihood estimates

$$\beta^* = \frac{\sum_1^N n_{1,i}}{\sum_1^N L_i}; \quad \alpha^* = \frac{\sum_1^N n_{2,i}}{\sum_1^N A_i} .$$

The quantities in the denominator are the times of risk of the event that one state changes to the other, and the quantities in the numerator are the numbers of times this has happened.

5.8. Metric transitivity — consistent estimates. We have hitherto considered the case when N independent realizations of the process have been observed. As seen it is then possible to use the maximum likelihood method to obtain consistent and asymptotically efficient estimates when N tends to infinity. In the important case when the process is stationary one might hope that it would be possible to get such estimates by the maximum likelihood method using only one realization of length T, when T tends to infinity. This seems probable, because, for large T, we can split up the interval $(0, T)$ into a large number of intervals I_n, separated by other intervals $I_{n'}$, where the latter ones have negligible length in proportion to the former ones, but still so large that the values of the process in two different I_n are approximately independent. The validity of this statement seems to demand some condition of asymptotic independence between values of the process observed at timepoints which are separated by a long interval. The following example is intended to illustrate this.

Let $y(t)$ be a normal process of the kind studied several times before with mean value zero and covariance function $\varrho(s, t)$ and x a normal stochastic variable independent of $y(t)$ for all t. x shall have mean value zero and standard deviation σ. We are observing the process

$$x(t) = m + x + y(t); \quad 0 \leq t \leq T;$$

where m is an unknown real parameter. As has been shown the maximum likelihood estimate m_T^* has a variance given by the expression

$$\begin{cases} D^2 m_T^* = \inf \int\limits_0^T \int\limits_0^T r(s, t) f(s) f(t)\, ds\, dt \\ \int\limits_0^T f(s)\, ds = 1, \end{cases}$$

where $r(s, t)$ is the covariance function of $x(t)$. Thus

$$r(s, t) = \varrho(s, t) + \sigma^2$$

and

$$D^2 m_T^* \geq \sigma^2.$$

Hence m_T^* is not a consistent estimate of m when T tends to infinity. This depends evidently on the fact that the autocorrelation of the process is too strong. In order to avoid this *we want to impose some condition on the process, which ensures the existence of a consistent estimate. The property we are going to use for this purpose is metric transitivity.*

To avoid unessential difficulties we consider the situation dealt with in 4.1 where we had only two simple hypotheses corresponding to the probability distributions P_{α_1} and P_{α_2}, $\alpha_1 < \alpha_2$. We shall now show the existence of a consistent estimate of α (or rather a consistent test) when the length T of the interval of observation tends to infinity. Consider all finite dimensional intervals $\{I_n\}$ where n denotes the number of dimensions. If for all $I \in \{I_n\}$

$$P_{\alpha_1}(I) = P_{\alpha_2}(I)$$

the distributions are equivalent which case is trivial. In the other case there must be some interval I with

$$P_{\alpha_1}(I) \neq P_{\alpha_2}(I), \quad \text{say} \quad P_{\alpha_1}(I) > P_{\alpha_2}(I).$$

We shall see in 5.14 that, using the property of metric transitivity, it is possible to construct a consistent estimate $\pi_T(I)$ of $P(I)$. Forming the real function $f(x)$ defined by

$$\begin{cases} f(x) = \alpha_1 & \text{for} \quad x \leq \dfrac{P_{\alpha_1}(I) + P_{\alpha_2}(I)}{2} \\ f(x) = \alpha_2 & \text{for} \quad x > \dfrac{P_{\alpha_1}(I) + P_{\alpha_2}(I)}{2} \end{cases}$$

we thus see that $f[\pi_T(I)]$ is a consistent estimate of α.

5.9. The method of maximum likelihood. Besides the metric transitivity we shall need another condition to ensure that the method of maximum likelihood gives optimum estimates. This condition will restrict, not the degree of dependence of the process, but the type of dependence. Regard the values of the process during the time subsequent to t. When the realization is known for $a \leq s \leq b$, $(b < t)$, we get a conditional distribution for the process $x(s)$, $s \geq t > b$. If there is a number T such that this conditional distribution only depends on the values observed during the time $b - T \leq s \leq b$, $(a < b - T)$, we shall say that $x(t)$ is of generalized Markoff type. To this type belong inter alia the usual Markoff process and the processes where the knowledge of the derivatives of some order suffice to determine the conditional probabilities. In the case of discrete time we have the Markoff chain of finite order.

In the following we shall suppose that the conditional distributions can be so defined that they almost certainly are probability distributions and that the likelihood functions satisfy conditions analogous to those stated in 5.7 (we still suppose that we have the regular case). Let us regard the process during the time (O, NT), where N is a positive integer, and denote the realization during $((\nu - 1)T, \nu T)$ with ω_ν, $\nu = 1, 2, \ldots N$. The type of coordinates used shall not depend upon ν. The space corresponding to ω_ν is called Ω_ν.

Consider an arbitrary set $A < \Omega_N$ and another $S < \Omega_1 \times \cdots \times \Omega_{N-1}$. Then introducing the likelihood functions and using the definition of conditional probability, we get

$$P_\alpha(SA) = \int_S P_\alpha\{\omega_N \in A \mid \omega_1, \ldots \omega_{N-1}\} \, d P_\alpha(\omega_1, \ldots \omega_{N-1}) =$$

$$= \int_S P_\alpha\{\omega_N \in A \mid \omega_{N-1}\} \, d P_\alpha(\omega_1, \ldots \omega_{N-1}) = \int_S P_\alpha\{\omega_N \in A \mid \omega_{N-1}\} f(\omega_1 \ldots \omega_{N-1}; \alpha) \, .$$

$$. \, d P_0(\omega_1 \ldots \omega_{N-1}) = \int_{SA} f(\omega_1, \ldots \omega_N; \alpha) \, d P_0(\omega_1, \ldots \omega_N).$$

Denoting

$$\int_A f(\omega_1, \ldots \omega_N; \alpha) \, d P_0(\omega_N \mid \omega_1 \ldots \omega_{N-1}) =$$

$$= \int_A f(\omega_1, \ldots \omega_N; \alpha) \, d P_0(\omega_N \mid \omega_{N-1}) = g(\omega_1, \omega_2, \ldots \omega_{N-1}, A; \alpha)$$

we get (see DOOB 2)

$$\int_S P_\alpha\{\omega_N \in A \mid \omega_{N-1}\} \, f(\omega_1, \ldots \omega_{N-1}; \alpha) \, d P_0(\omega_1, \ldots \omega_{N-1}) =$$

$$= \int_S g(\omega_1, \ldots \omega_{N-1}, A; \alpha) \, d P_0(\omega_1, \ldots \omega_{N-1})$$

for every set $S < \Omega_1 \times \cdots \times \Omega_{N-1}$. Thus almost certainly

$$P_a(\omega_N \in A \mid \omega_{N-1}) = \int_A \frac{f(\omega_1, \ldots \omega_N; a)}{f(\omega_1, \ldots \omega_{N-1}; a)} \, d P_0(\omega_N \mid \omega_{N-1})$$

from which can be deduced that

$$\frac{f(\omega_1, \ldots \omega_N; a)}{f(\omega_1, \ldots \omega_{N-1}; a)}$$

almost certainly does not depend upon $\omega_1, \omega_2, \ldots \omega_{N-2}$. As we have supposed the regular case the denominator is different from zero with probability one. We write

$$f(\omega_1, \ldots \omega_N; a) = f(\omega_1; a) \frac{f(\omega_1, \omega_2; a)}{f(\omega_1; a)} \cdots \frac{f(\omega_1, \ldots \omega_N; a)}{f(\omega_1, \ldots \omega_{N-1}; a)}.$$

In the same way as above one shows that the ratios depend only on the two last ω's entering into the expressions. Thus we can write

$$f(\omega_1, \ldots \omega_N; a) = f(\omega_1; a) f_2(\omega_2 \mid \omega_1; a) \ldots f_N(\omega_N \mid \omega_{N-1}; a).$$

Because of the stationarity of the process we can leave out the indices of the f's

$$f(\omega_1, \ldots \omega_N; a) = f(\omega_1; a) f(\omega_2 \mid \omega_1; a) \ldots f_N(\omega_N \mid \omega_{N-1}; a).$$

Hence

$$\frac{1}{N} \frac{\partial \log f(\omega_1, \ldots \omega_N; a)}{\partial a} = \frac{1}{N} \frac{\partial \log f(\omega_1; a)}{\partial a} + \frac{1}{N} \sum_2^N \frac{\partial \log f(\omega_\nu \mid \omega_{\nu-1}; a)}{\partial a}$$

and now we use the method in CRAMÉR 4 p. 501–503 and have, with analogous notation, the likelihood equation

$$B_0 + B_1(a - a_0) + \tfrac{1}{2} \theta B_2 (a - a_0)^2 = 0$$

with

$$\begin{cases} B_0 = \dfrac{1}{N} \left(\dfrac{\partial \log f(\omega_1; a)}{\partial a} \right)_0 + \dfrac{1}{N} \displaystyle\sum_2^N \left(\dfrac{\partial \log f(\omega_\nu \mid \omega_{\nu-1}; a)}{\partial a} \right)_0 \\ B_1 = \dfrac{1}{N} \left(\dfrac{\partial^2 \log f(\omega_1; a)}{\partial a^2} \right)_0 + \dfrac{1}{N} \displaystyle\sum_2^N \left(\dfrac{\partial^2 \log f(\omega_\nu \mid \omega_{\nu-1}; a)}{\partial a^2} \right)_0 \\ B_2 = \dfrac{1}{N} H(\omega_1) + \dfrac{1}{N} \displaystyle\sum_2^N H(\omega_\nu \mid \omega_{\nu-1}). \end{cases}$$

Because of the ergodic theorem and the metric transitivity these expressions converge to their mean values in probability when N tends to infinity. But

$$E_0\left(\frac{\partial \log f(\omega_\nu \,|\, \omega_{\nu-1}; a)}{\partial a}\right)_0 = E_0\left(\frac{\partial \log f(\omega_1, \ldots \omega_\nu; a)}{\partial a}\right)_0 - \\ - E_0\left(\frac{\partial \log f(\omega_1, \ldots \omega_{\nu-1}; a)}{\partial a}\right)_0 = 0.$$

Putting

$$- E_0\left(\frac{\partial^2 \log f(\omega_2 \,|\, \omega_1; a)}{\partial a^2}\right)_0 = k$$

(we have to suppose that $k \neq 0$, otherwise we would get a trivial case), we get

$$E_0\left(\frac{\partial \log f(\omega_1, \ldots \omega_N; a)}{\partial a}\right)_0^2 = - E_0\left(\frac{\partial^2 \log f(\omega_1, \ldots \omega_N; a)}{\partial a^2}\right)_0 = \\ = - E_0\left(\frac{\partial^2 \log f(\omega_1; a)}{\partial a^2}\right)_0 + (N-1)\,k.$$

Thus

$$B_0 \to 0, \; B_1 \to -k, \; B_2 \to M < \infty \quad \text{in probability.}$$

In the same way as in CRAMÉR 4 we can show that there is a consistent maximum likelihood estimate and we have

$$\sqrt{N k}\,(a^* - a) = \frac{\sqrt{\frac{N}{k}}\,B_0}{u_N},$$

where u_N converges to unity in probability. But B_0 has mean value zero and variance

$$-\frac{1}{N^2} E_0\left(\frac{\partial^2 \log f(\omega_1; a)}{\partial a^2}\right)_0 + \frac{N-1}{N^2}\,k.$$

Using the definition of asymptotical efficiency which has been given by WALD 1 we have thus proved that *there is a maximum likelihood estimate which is consistent and asymptotically efficient.*

5.10. Criteria of metric transitivity. The concept of metric transitivity seems to be important in the problem of estimation in the case of a stationary stochastic process. The results given in DOOB 2 dealing with Markoff processes may be useful in this connection. We will give two other criteria of metric transitivity.

Theorem: *In order that a stationary normal process with a continuous correlation function $r(t)$ shall be metrically transitive, it is necessary and sufficient that the spectrum of the process is continuous.*

For the proof we shall utilize ideas due to DOOB 1 and ÎTO 1. We suppose the process to be (D)-integrable. Let

$$r(t) = \int_{-\infty}^{\infty} e^{it\lambda}\, dF(\lambda), \quad r(0) = 1,$$

where $F(\lambda)$ is the spectral function, which we have supposed to be continuous. If the process is not metrically transitive, there exists a set S which is invariant and has $P(S) = \varrho$, $0 < \varrho < 1$. We approximate S with a finite sum I of finite dimensional intervals in such a way that

$$\begin{cases} P(I) < \varrho + \varepsilon \\ P(S I^*) < \varepsilon \end{cases}$$

where ε is a positive number given in advance. It is evident that we can choose intervals of finite sides. Let $T_t I = I_t$. Denote the time points corresponding to I_t by $\tau_1 + t, \ldots \tau_n + t$. Introduce

$$\begin{cases} x_i = x(\tau_i); \; i = 1, 2, \ldots n. \\ x_{n+i} = x(\tau_i + t); \; i = 1, 2, \ldots n. \end{cases}$$

These stochastic variables have a $2n$-dimensional normal probability distribution determined by the moment matrix

$$\Lambda(t) = \left\{ \begin{array}{ccc|ccc} 1 & r(\tau_1 - \tau_2) \cdots & & r(\tau_1 - \tau_1 - t) \cdots & & \\ \cdots & & & \cdots & & \\ & & & & & \\ r(\tau_n - \tau_1) \cdots & & 1 & & & \\ \hline r(\tau_1 + t - \tau_1) \cdots & & & 1 \cdots & & \\ \cdots & & & \cdots & & \\ & & & & & \\ r(\tau_n + t - \tau_1) \cdots & & & & & 1 \end{array} \right\} = \begin{Bmatrix} A & B(t) \\ B(t) & A \end{Bmatrix},$$

where the matrix A does not depend on t. Now it is evident that no moment matrices can be singular because of the spectrum being continuous. We have (large values of t)

$$P(I I_t) = \frac{\Lambda(t)^{-\frac{1}{2}}}{(2\pi)^n} \int \cdots \int_{(x_1, \ldots x_n) \in I} \int \cdots \int_{(x_{n+1}, \ldots x_{2n}) \in I_t} e^{-\frac{1}{2} Q(x)} \, dx_1 \ldots dx_{2n},$$

where $Q(x)$ is the quadratic form in $x_1, x_2, \ldots x_{2n}$ corresponding to the inverse of $\Lambda(t)$. We arrange all numbers of the form $\tau_i - \tau_j$ as $t_1, t_2, \ldots t_N$. Then using the absolute integrability we get

$$\frac{1}{2T} \int_{-T}^{T} \sum_1^N |r(t_i + t)|^2 \, dt = \frac{1}{2T} \int_{-T}^{T} \sum_1^N \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{i(t_i+t)\lambda - i(t_i+t)\mu} \, dF(\lambda) \, dF(\mu) \, dt =$$

$$= \sum_1^N \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{e^{i(t_i+T)(\lambda-\mu)} - e^{i(t_i-T)(\lambda-\mu)}}{2Ti(\lambda-\mu)} \, dF(\lambda) \, dF(\mu).$$

By an argument of the usual type (see e.g. HOPF 1, p. 16) we get

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \sum_{1}^{N} | r(t_i + t) |^2 = 0,$$

and thus

$$\underline{\lim}_{t\to\infty} \sum_{1}^{N} | r(t_i + t) |^2 = 0.$$

Hence there is a sequence t_ν tending to infinity when ν tends to infinity such that $B_{t_\nu} \to 0$ and

$$\Lambda(t_\nu) \to \begin{Bmatrix} A & 0 \\ 0 & A \end{Bmatrix}.$$

Using Lebesgue's theorem on bounded convergence we have

$$P(I I_t) \to P(I)^2.$$

Thus for large values of ν

$$(\varrho + \varepsilon)^2 > P(I I_{t_\nu}) \geq P(S I I_{t_\nu}) \geq P(S) - P(S I^*) - P(S I^*_{t_\nu}) > \varrho - 2\varepsilon.$$

But in order that this shall be possible for an arbitrarily small ε, we must have $\varrho = 1$ or 0 contrary to our assumption, which proves the sufficiency of our condition.

To see that it is also necessary we consider the process $x(t)^2$. As the normal distribution has a finite moment of the fourth order this process has finite variance and further

$$\varrho(t) = E[x^2(s) - E x^2(s)]\,[x^2(s+t) - E x^2(s+t)] = 2 r^2(t).$$

We know that the limit

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x^2(t)\, dt = y$$

exists almost certainly and has the variance

$$D^2 y = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \varrho(t)\, dt = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} 2 r^2(t)\, dt.$$

But the last expression is according to CRAMÉR 1

$$D^2 y = 2 \sum_{1}^{\infty} \Delta_\nu F$$

where $\Delta_\nu F$ denote the saltuses in the points of discontinuity of $F(\lambda)$. In order that $x(t)$ shall be metrically transitive it is thus necessary that $F(\lambda)$ is continuous which completes our proof.

When we do not suppose the process to be normally distributed, we do not get complete knowledge of the process by specifying just the correlation function. It is still possible to give a criterion of metric transitivity which essentially is a generalization of the above.

The process is said to be mixing, if for every pair of measurable sets A, B it is true that

$$\lim_{t\to\infty} P(A B_t) = P(A)\,P(B) \quad \text{where} \quad B_t = T_t B.$$

One knows that the mixing property implies metric transitivity but the converse is not true (see HOPF 1). In the theorem just proved we have seen that if the process has no point spectrum it is metrically transitive. The spectrum can then consist of an absolutely continuous part and a singular part. If also the singular part disappears, Îto has shown in the normal case that the process is mixing. But we have seen, again in the normal case, that the process is metrically transitive even in the case of the spectrum having a singular component. But then there is a sequence $t_\nu \to \infty$ such that $\lim\limits_{\nu\to\infty} r(t_\nu) = 0$. This leads us to the following weakened concept of a mixing process.

The process is said to be *partially mixing* if for every measurable set A it is true that there is a sequence $t_\nu(A)$ such that

$$\lim_{\nu\to\infty} P(A A_{t_\nu}) = P(A)^2.$$

Theorem: *In order that a process shall be metrically transitive it is necessary and sufficient that it is partially mixing.*

If $x(t)$ is partially mixing, it is shown in the same way as above that it is metrically transitive. We only have to show the necessity of the condition. Take an arbitrary set A and call the characteristic function of A_t for $c(t, \omega)$. This is a stationary process with correlation function

$$r_A(t) = E\,c(s,\omega)\,c(s+t,\omega) - E\,c(s,\omega)\,E\,c(s+t,\omega) =$$
$$= P(A_s A_{s+t}) - P(A_s)\,P(A_{s+t}) = P(A A_t) - P(A)^2.$$

Because of the ergodicity we must have

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} r_A(t)\,dt = \lim_{T\to\infty} \frac{1}{T}\int_{0}^{T} r_A(t)\,dt = 0\,.$$

As the process is supposed to be (D)-integrable and -measurable, $P(A A_t)$ is a continuous function of t (see e.g. HOPF 1). Either $r_A(t)$ has a sequence of zeroes tending to infinity, or it is of constant sign for $t > t_0$. In both cases we can find a sequence $t_\nu(A)$ for which

$$\lim_{\nu\to\infty} P(A A_{t_\nu}) = P(A)^2,$$

which proves the theorem.

Remark: In the definition of the mixing property a certain condition is to be satisfied for every measurable set A. This is not a quite convenient formulation for applications. We shall show that it is sufficient to consider only finite dimensional intervals. Suppose that

$$\lim_{t\to\infty} P(I\,J_t) = P(I)\,P(J)$$

for every pair of finite dimensional intervals I, J. If A is an arbitrary measurable set, we can approximate to it by a finite sum $\Sigma = \Sigma\, I_\nu$ of disjoint intervals so that

$$P(A^*\Sigma) + P(A\,\Sigma^*) < \varepsilon.$$

Then

$$|P(A) - P(\Sigma)| < \varepsilon,$$

and

$$P\{A\,A_t(\Sigma\,\Sigma_t)^*\} + P\{(A\,A_t)^*\,\Sigma\,\Sigma_t\} \le P(A\,A_t\,\Sigma^*) + P(A^*\,\Sigma\,\Sigma_t) + P(A_t^*\,\Sigma\,\Sigma_t) + P(A\,A_t\,\Sigma_t^*) \le P(A\,\Sigma^*) + P(A^*\,\Sigma) + P(A_t\,\Sigma_t^*) + P(A_t^*\,\Sigma_t) < 2\,\varepsilon.$$

But

$$P(\Sigma\,\Sigma_t) = P(\sum_\nu I_\nu \sum_\mu I_\mu^t) = \sum_{\nu,\mu} P(I_\nu\, I_\mu^t)$$

which tends to $\sum_{\nu,\mu} P(I_\nu)\,P(I_\mu)$ as t tends to infinity and thus

$$\lim_{t\to\infty} P(\Sigma\,\Sigma_t) = P(\Sigma)^2.$$

We get

$$|P(A\,A_t) - P(A)^2| \le |P(A\,A_t) - P(\Sigma\,\Sigma_t)| + |P(\Sigma\,\Sigma_t) - P(\Sigma)^2| + |P(\Sigma)^2 - P(A)^2| \le 4\,\varepsilon + |P(\Sigma\,\Sigma_t) - P(\Sigma)^2|,$$

so that

$$\lim_{t\to\infty} P(A\,A_t) = P(A)^2.$$

5.11. Applications. We shall proceed to apply the method of maximum likelihood to two simple stationary stochastic processes. If $x(t)$ is a stationary normal Markoff process with mean value m and covariance function $e^{-\beta|t-s|}$, we know that the spectrum is absolutely continuous and hence the process is metrically transitive. The likelihood function is

$$f(\omega, m) = e^{-\frac{m^2}{2}\left(1+\frac{\beta T}{2}\right)+\frac{m}{2}\left\{x(0)+x(T)+\beta\int\limits_0^T x(t)\,dt\right\}},$$

and the maximum likelihood estimate is

$$m^* = \frac{x(0) + x(T) + \beta\int\limits_0^T x(t)\,dt}{2 + \beta\,T}$$

which thus is a consistent and asymptotically efficient estimate of m. In this case this is true a fortiori because we have seen in 5.6 that m^* is efficient for finite T.

Consider now the process in 4.9. It is stationary and of the Markoff type. The covariance function is also now $e^{-\beta|t-s|}$ but as the process is not normally distributed we can not apply the same result as before to show that it is metrically transitive. But consider an interval I with the corresponding time-points $t_1, t_2, \ldots t_n$ and another interval J with the time-points $t'_1, t'_2 \ldots t'_m$. Let t be a large positive number. Then

$$P(I J_t) = P_0(t) P(I J_t | 0) + P_1(t) P(I J_t | 1)$$

where the index 0 is used to denote the condition that no change has occurred during the time t_n, $t'_1 + t$ and the index 1 for the alternative condition.

$$P_0(t) = e^{-\beta(t'_1+t-t_n)} \to 0$$

when $t \to \infty$. But

$$P(I J_t | 1) = P(I) P(J)$$

which by the aid of the remark in the preceding section shows that $x(t)$ is metrically transitive. The maximum likelihood estimate has the simple form

$$m^* = \frac{1}{n+1} \sum_0^n x_\nu,$$

which can be considered as an integral with a weight function depending upon the realization. It is unbiased because

$$E m^* = \sum_0^\infty P_\nu E[m^* | \nu] = m \sum_0^\infty P_\nu = m.$$

The variance is easily calculated

$$E(m^* - m)^2 = \sum_0^\infty P_\nu E[(m^* - m)^2 | \nu] = \sum_0^\infty \frac{(\beta T)^\nu}{\nu!} e^{-\beta T} \frac{1}{\nu+1} = \frac{1 - e^{-\beta T}}{\beta T}.$$

And as

$$E\left(\frac{\partial \log f(\omega, m)}{\partial m}\right)^2 = \sum_0^\infty P_\nu E\left\{\left[\sum_0^\nu x_n - (\nu+1) m\right]^2 \Big| \nu\right\} =$$

$$= e^{-\beta T} \sum_0^\infty \frac{(\beta T)^\nu}{\nu!} (\nu + 1) = 1 + \beta T,$$

we get the following expression for the efficiency

$$e(m^*) = \frac{\beta T}{(1 + \beta T)(1 - e^{-\beta T})}.$$

For $T = 0$ we get the efficiency 1, and when T increases $e(m^*)$ decreases at first, but for large values of T $e(m^*)$ tends to 1 again. If we had used the best linear estimate of m, we should have got

$$m_L^* = \frac{x(0) + x(T) + \beta \int\limits_0^T x(t)\,dt}{2 + \beta T}$$

with the variance

$$D^2\, m_L^* = \frac{2}{2 + \beta T}.$$

We then have the efficiency

$$e(m_L^*) = \frac{1 + \frac{\beta T}{2}}{1 + \beta T}.$$

For $T = 0$ we have $e(m_L^*) = e(m^*) = 1$, and the two estimates must coincide. For $T = 0$ we get with probability one $n = 0$ and $m^* = x_0$, and of course we have $m_L^* = x_0$ from the expression for m_L^*. When T tends to infinity we get $e(m_L^*) \to \frac{1}{2}$.

The equidistributed estimate

$$m_E^* = \frac{1}{T} \int\limits_0^T x(t)\,dt$$

has the variance

$$D^2\, m_E^* \sim \frac{2}{\beta T}$$

asymptotically when T tends to infinity, and the asymptotic efficiency

$$\lim_{T \to \infty} e(m_E^*) = \tfrac{1}{2}.$$

By using some of these linear estimates we thus loose about 50 % of the efficiency if T is large.

5.12. Distribution of a type of estimates. When we consider stationary point-processes with adjoined stochastic variables, we may sometimes meet with estimates involving expressions of the type $\sum\limits_1^n x_i$. To study the asymptotic distribution of these when T tends to infinity we suppose, without any attempt to generality, that the x's are independent with mean value zero and standard deviation σ. We suppose further that $n \to \infty$ in probability when $T \to \infty$. Then

$$\frac{\sum\limits_1^n x_i}{\sigma \sqrt{n}}$$

is asymptotically normal (0, 1) when $T \to \infty$, because

$$P\left\{\frac{\sum\limits_1^n x_i}{\sigma\sqrt{n}} \leq a\right\} = \sum_1^\infty P_T(\nu)\, P\left\{\frac{\sum\limits_1^n x_i}{\sigma\sqrt{n}} \leq a \mid \nu = n\right\}$$

and

$$P\left\{\frac{\sum\limits_1^n x_i}{\sigma\sqrt{n}} \leq a \mid \nu = n\right\} \to \Phi(a); \quad \nu \to \infty;$$

where $\Phi(x)$ is the normal distribution function, because of the central limit law. For $\varepsilon > 0$ there is a number $\nu(\varepsilon)$ such that

$$\left| P\left\{\frac{\sum\limits_1^n x_i}{\sigma\sqrt{n}} \leq a \mid \nu = n\right\} - \Phi(a)\right| < \varepsilon \text{ for } \nu > \nu(\varepsilon).$$

Choose T so large that $\sum\limits_1^{\nu(\varepsilon)} P_T(\nu) < \varepsilon$. Then

$$\left| P\left\{\frac{\sum\limits_1^n x_i}{\sigma\sqrt{n}} \leq a\right\} - \Phi(a)\right| \leq \sum_1^\infty P_T(\nu) \left| P\left\{\frac{\sum\limits_1^n x_i}{\sigma\sqrt{n}} \leq a \mid n = \nu\right\} - \Phi(a)\right| \leq$$

$$\leq 2\,\varepsilon + \varepsilon \sum_{\nu(\varepsilon)+1}^\infty P_T(\nu) < 3\,\varepsilon$$

which is the stated result.

If $\dfrac{D\,n(T)}{E\,n(T)}$ tends to zero when T tends to infinity, the sum $\sum\limits_1^n x_i$ is asymptotically normal $\left\{0, \sigma\sqrt{E\,n(T)}\right\}$ because

$$\frac{\sum\limits_1^n x_i}{\sigma\sqrt{E\,n(T)}} = \frac{\sum\limits_1^n x_i}{\sigma\sqrt{n}} \sqrt{\frac{n}{E\,n(T)}}.$$

But the stochastic variable $\dfrac{n}{E\,n(T)}$ has mean value 1 and standard deviation $\dfrac{D\,n(T)}{E\,n(T)}$, and hence converges to 1 in probability when $T \to \infty$. With the aid of theorem 20.6 in CRAMÉR 4 the result immediately follows.

Finally, $\frac{\sum\limits_1^n x_i}{n}$ is asymptotically normal $\left\{0, \frac{\sigma}{\sqrt{E\, n(T)}}\right\}$ which is seen in the same manner because

$$\sqrt{E\, n(T)}\, \frac{\sum\limits_1^n x_i}{n\,\sigma} = \frac{\sum\limits_1^n x_i}{\sigma \sqrt{n}} \sqrt{\frac{E\, n(T)}{n}}.$$

5.13. Approximation of estimates. We shall now give a result concerning estimates analogous to 4.12. To obtain the regular case we have supposed that P_α is absolutely continuous with respect to P_0. We now demand that this holds uniformly for $\alpha \in A$. Then, if $\alpha^*(\omega)$ is an estimate with

$$E_\alpha\, \alpha^{*2}(\omega) = v(\alpha),$$

which we suppose to be a continuous function of α, it is possible to approximate α^* by estimates $\alpha^*(x_1, \ldots x_n)$ involving only a finite number of the coordinates. The approximation will be uniform for $\alpha \in A$.

Form the stochastic variable

$$t_N(\omega) = \begin{cases} \alpha^*(\omega) & \text{if } \quad |\alpha^*(\omega)| < N \\ 0 & \text{if } \quad |\alpha^*(\omega)| \geq N. \end{cases}$$

Then

$$E_\alpha (t_N - \alpha^*)^2 = \int\limits_\Omega (t_N - \alpha^*)^2\, d\, P_\alpha = \int\limits_{|\alpha^*| \geq N} \alpha^{*2}\, d\, P_\alpha \downarrow 0$$

when N tends to infinity. But

$$\left| \int\limits_{|\alpha^*| < N} \alpha^{*2} \{f(\omega, \alpha_0) - f(\omega, \alpha)\}\, d\, P_0(\omega) \right| \leq N^2 \sqrt{\int\limits_\Omega [f(\omega, \alpha_0) - f(\omega, \alpha)]^2\, d\, P_0(\omega)}$$

which according to the assumption on $f(\omega, \alpha)$ in 5.1 tends to zero when α tends to α_0. Thus

$$\int\limits_{|\alpha^*| \geq N} \alpha^{*2}(\omega)\, f(\omega, \alpha)\, d\, P_0(\omega) = v(\alpha) - \int\limits_{|\alpha^*| < N} \alpha^{*2}(\omega)\, f(\omega, \alpha)\, d\, P_0(\omega)$$

is a continuous function of α. Because of Dini's theorem the above convergence will be uniform, so that for every $\varepsilon > 0$ there is a $N_0 = N_0(\varepsilon)$ with

$$E_\alpha [t_N - \alpha^*]^2 < \varepsilon \quad \text{for} \quad N > N_0.$$

Consider now the stochastic variable

$$\alpha_N^*(x_1, \ldots x_n) = E_0 [t_N \mid x_1, \ldots x_n].$$

When n tends to infinity $a_N^*(x_1, \dots x_n)$ tends to $t_N(\omega)$ almost certainly. Introduce the set

$$\{|a_N^*(x_1, \dots x_n) - t_N(\omega)| < \varepsilon\}_\omega = E_n < \Omega.$$

We get

$$E_\alpha[a_N^*(x_1, \dots x_n) - t_N(\omega)]^2 = \int_\Omega [a_N^*(x_1, \dots x_n) - t_N(\omega)]^2 \, d P_\alpha(\omega) \leq$$
$$\leq \varepsilon^2 P_\alpha(E_n) + 4 N^2 P_\alpha(E_n^*).$$

But $P_0(E_n^*) \to 0$ when $n \to \infty$ and because of the uniform absolute continuity of P_α with respect to P_0 we obtain

$$E_\alpha[a_N^*(x_1, \dots x_n) - t_N(\omega)]^2 < \delta \quad \text{if} \quad n > n_0(N, \delta).$$

Using the triangular inequality we get the wanted result

$$E_\alpha[a_N^*(x_1, \dots x_n) - a^*(\omega)]^2 < \varepsilon \quad \text{for all } \alpha \in A$$

if N and n are chosen sufficiently large. We can then get an estimate depending on a finite number of coordinates which has mean value and variance arbitrarily close to those of $\alpha^*(\omega)$ uniformly for $\alpha \in A$.

If we form estimates involving a finite number n of coordinates and choose the best one $\alpha^*(x_1, \dots x_n)$ of these, the above shows that when we increase n sufficiently, $\alpha^*(x_1, \dots x_n)$ is practically as good an estimate as any estimate depending on all coordinates.

5.14. Estimation of functions. We have hitherto mainly considered the case when the probability distribution of the process is known but for a real parameter. The problem dealt with in 5.2–5 is of another type, because we deal there with processes about whose probability distribution nothing is known except the covariance function. Another type of problems of similar nature is obtained when *the distribution of the process depends upon an unknown function, which we want to estimate by the aid of our observations.* The two following cases are illustrative.

Let $x(t)$ be a real, stationary, normal and (D)-measurable process with mean value zero and correlation function $r(t)$ which is supposed to be continuous as usual. The process is observed during the time $(0, T)$ and we want to estimate $r(t)$. It is possible to give a consistent estimate, if the process is metrically transitive, i.e. if the spectrum is continuous (see 5.10.). We known (see HOPF 1, p. 54—55) that almost certainly for all t

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T x(s)\, x(s-t)\, ds = r(-t) = r(t).$$

As the process is real and the correlation function symmetrical, we have to consider only $t > 0$. But

$$\frac{1}{T} \int_0^T x(s)\, x(s-t)\, ds - \frac{1}{T} \int_t^T x(s)\, x(s-t)\, ds = \frac{1}{T} \int_0^t x(s)\, x(s-t)\, ds$$

which almost certainly tends to zero for all t when T tends to infinity. Hence the expression

$$r_T^*(t) = \begin{cases} \dfrac{1}{T} \displaystyle\int_t^T x(s)\, x(s-t)\, ds & \text{for} \quad 0 \leq t < T \\ 0 & \text{for} \quad t \geq T, \end{cases}$$

which depends only upon observations during the time $(0, T)$, is a consistent estimate of $r(t)$.

Consider now a process that is still stationary, (D)-integrable and metrically transitive. We want to find a consistent estimate for the distribution function

$$F(a) = P\{x(t) \leq a\}, \quad -\infty < a < \infty.$$

To this end we introduce the stochastic process

$$e_t(\omega) = \begin{cases} 1 & \text{if} \quad x(t, \omega) \leq a \\ 0 & \text{if} \quad x(t, \omega) > a. \end{cases}$$

For a fixed value of t this is a stochastic variable. $e_t(\omega)$ is measurable and integrable on the product space $T \times \Omega$ where T is an arbitrary finite interval. We thus have with probability one

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} e_t(\omega)\, dt = E\, e_t(\omega) = P\{x(t) \leq a\} = F(a).$$

But $\int_{-T}^{T} e_t(\omega)$ is the time belonging to the interval $(-T, T)$ when $x(t) \leq a$. Denoting

$$\frac{1}{2T} m\{x(t, \omega) \leq a;\ |t| < T\}_t = F_T^*(a, \omega),$$

which is possible almost certainly according to Fubini's theorem, we get

$$\lim_{T \to \infty} F_T^*(a, \omega) = F(a).$$

Let $\{a_\nu;\ \nu = 1, 2, \ldots\}$ be a sequence of real numbers that is everywhere dense on the real axis. Because of the denumerability we have

$$\lim_{T \to \infty} F_T^*(a_\nu, \omega) = F(a_\nu)$$

almost certainly for all ν. But $F_T^*(a, \omega)$ is a non-decreasing function of a. If a is a point of continuity of $F(x)$ and $a_\nu' \downarrow a$ and $a_\nu'' \uparrow a$ we have

$$F_T^*(a_\nu'', \omega) \leq F_T^*(a, \omega) \leq F_T^*(a_\nu', \omega)$$

and it follows that

$$\lim_{T \to \infty} F_T^*(a, \omega) = F(a)$$

almost certainly for all points of continuity of $F(x)$, i.e. F_T^* is a consistent estimate of F. In the same way we can construct consistent estimates for the multi-dimensional distribution functions.

It is easily seen that the estimate in the latter case is unbiased and that the estimate in the former case can be made unbiased by multiplication with the factor $\frac{T}{T-t}\cdot$ It seems desirable to define concepts like efficiency and to investigate properties of estimates of functions in such terms.

Before we leave the problem of estimation, we want to point out that the definition of a confidence region can now be translated almost word for word to the case of a stochastic process.

The problem of regression

6.1. Regression in function space. Besides the problem of testing and estimation we will just shortly deal with two other types of inference and show how they belong to the theory of regression which is since long familiar to the statistician. In the following we shall have to deal with conditional distributions, and assume, as some times before, that these can be defined with probability one in such a way that they are probability distributions.

We observe a stochastic process $x(t)$ during the time-interval T, and desire to make a statement about a stochastic variable y, when we know the simultaneous distribution of $\{y, x(t);\ t \in T\}$. Denoting the observed realization by ω, we have a conditional probability distribution $P\{y \mid \omega\}$ for y. We want to give a probable value of y knowing ω, and we can take some central value of the distribution $P\{y \mid \omega\}$. If y has a finite expectation, it seems reasonable to take the conditional expectation as an estimate of y

$$y^* = E[y \mid \omega].$$

To be able to proceed further we must specify the distribution. Suppose that the process and y are normally distributed with mean value zero and that $x(t)$ is continuous in the mean. Denote the Hilbert space generated by $x(t)$, $t \in T$, by $L_2(X)$ and form

$$y_1 = P_{L_2(X)}\, y,$$

and put

$$y = y_1 + z.$$

Then $z \perp x(t);\ t \in T$, and because of the normality of the distributions $x(t)$, $(t \in T)$, and z are independent stochastic variables. We have almost certainly

$$E[y \mid \omega] = E[y_1 \mid \omega] + E[z \mid \omega] = y_1(\omega) + E z = y_1(\omega).$$

But y_1 can also be considered as the point in $L_2(X)$ which makes $\|y - x\|$; $x \in L_2(X)$; as small as possible. If nothing is assumed about the distributions, y_1 seems still to be a reasonable estimate of y. This is nothing but a generalization of the fact that the regression of the multi-dimensional normal distribution is linear.

6.2. Prognosis as regression. Suppose that $x(t)$ has the properties demanded in the preceding section and is normally distributed. As y we take $x(c)$ with $c \notin T$. The question of how $x(c)$ shall be estimated by the aid of $x(t)$, $t \in T$, is known as the problem of prediction or prognosis. The result of 6.1 is immediately applicable to this problem. As several times before we represent the process by the following series which converges in the mean

$$x(t) = \sum_{1}^{\infty} z_{\nu} \frac{\varphi_{\nu}(t)}{\sqrt{\lambda_{\nu}}}; \quad t \in T;$$

where the involved quantities are defined in 1.3. It is easily seen that $L_2(X)$ coincides with the space Z spanned by the orthonormal vectors z_{ν}, $\nu = 1, 2, \ldots$, so that in order to obtain the prognosis of the process to the time c, we just have to form

$$x^{*}(c) = P_{Z}\, x(t) = \sum_{1}^{\infty} z_{\nu} E z_{\nu} x(c).$$

But

$$E z_{\nu} x(c) = \sqrt{\lambda_{\nu}}\, E x(c) \int_{T} x(t) \varphi_{\nu}(t)\, dt = \sqrt{\lambda_{\nu}} \int_{T} r(c, t) \varphi_{\nu}(t)\, dt.$$

For $s \in T$ we have

$$\lambda_{\nu} \int_{T} r(s, t) \varphi_{\nu}(t)\, dt = \varphi_{\nu}(s)$$

and so it is natural to put

$$E z_{\nu} x(c) = \frac{\varphi_{\nu}^{*}(c)}{\sqrt{\lambda_{\nu}}},$$

where $\varphi_{\nu}^{*}(c)$ are the continued eigen-functions of the process. Hence the best prognosis is given by

$$x^{*}(c) = \sum_{1}^{\infty} z_{\nu} \frac{\varphi_{\nu}^{*}(c)}{\sqrt{\lambda_{\nu}}}.$$

This representation is due to Karhunen, defining the best prognosis as the point in $L_2(X)$ which has the smallest distance to $x(c)$. For the important case of a stationary process observed during the interval $T = (-\infty, a)$, WIENER 1 has obtained a technique of finding the best linear prognosis.

6.3. An example. If the process is not normally distributed we can either use the best linear prognosis, or we can try to calculate the conditional expectation. Let us consider the simple process of 4.11, which is not normally distributed. Putting $T = (a, b)$ we get the conditional distribution function for $x(c)$

$$F(x \mid \omega) = e^{-\beta(b-c)} \varepsilon[x - x(b)] + [1 - e^{-\beta(b-c)}] \Phi(x).$$

The conditional expectation is then

$$E[x(c) \mid \omega] = x(b)\, e^{-\beta(b-c)},$$

which is the same result that would have been obtained by constructing the best linear prognosis.

6.4. Regions of prognosis. It may happen that we want to give, not a single point, but a *region* into which the values of the process at some future time-points may be expected to fall. This is quite analogous to the case of estimation by confidence regions (see 2.3).

Suppose that we have observed the process during the time (a, b), and denote the realization by $\omega_{a,b} \in \Omega_{a,b}$. We want to determine a region π depending upon $\omega_{a,b}$ in the space of all realizations $\omega_{c,d}$ during the time (c, d), $\pi < \Omega_{c,d}$, such that $\omega_{c,d}$ can reasonably be expected to fall in π, i. e. such that $P(\pi \mid \omega_{a,b})$ is large. In order to be able to choose between different regions, we introduce a measure m in $\Omega_{c,d}$ with the property that $\Omega_{c,d}$ is the denumerable sum of sets of finite m-measure. For a fixed $\omega_{a,b}$ we have the conditional probability

$$P(S \mid \omega_{a,b});\ S < \Omega_{c,d},$$

which is a probability distribution with probability one. According to the theorem of Lebesgue on decomposition of additive set functions we obtain

$$P(S \mid \omega_{a,b}) = P(X S \mid \omega_{a,b}) + \int_S f(\omega_{c,d} \mid \omega_{a,b})\, d\, m(\omega_{c,d}),$$

where $X = X(\omega_{a,b})$ is the singular part of the distribution. We want to find a set $\pi < \Omega_{c,d}$ with fixed $m(\pi)$ and with maximum $P(\pi \mid \omega_{a,b})$. This is formally the same problem that we have considered in 4.1 and we get

$$\pi = X(\omega_{a,b}) + \{f(\omega_{c,d} \mid \omega_{a,b}) \geq k\}_{\omega_{c,d}} < \Omega_{c,d},$$

where the constant k is determined to give $m(\pi)$ the required value. *π is called the best region of prognosis with respect to m, and is thus obtained by using a sort of maximum likelihood principle.*

It is evident that the obtained best region of prognosis will depend upon the choice of the measure m. We will just sketch two possible ways of choosing m, applied to a Markoff process.

Consider the following case, where $x(n)$ is a stationary, normal, Markoff process observed in the integral time-points $(\cdots -1, 0, 1, \cdots)$, with mean value zero and standard deviation 1. The correlation function is then

$$r(n) = e^{-\beta |n|},\ \beta > 0.$$

Here we have left out the trivial cases $\beta = 0$, $\beta = +\infty$. Let

$$(a, b) = (-N, -N + 1, \cdots 0)$$

and let the interval (c, d) be the point $t = \tau > 0$. Because of the Markoff property the conditional probability of $x(\tau)$ with respect to

$$x(-N), x(-N + 1), \ldots x(0)$$

depends only on $x(0)$. As the measure m we take the Lebesgue measure on the axis $-\infty < x(\tau) < \infty$. The singular part X does not appear and

$$f = c\, e^{-\frac{1}{2} \frac{[x(\tau) - e^{-\beta\tau} x(0)]^2}{1 - e^{-2\beta\tau}}}$$

Thus we get the best region of prognosis

$$-k < x(\tau) - e^{-\beta\tau} x(0) < k.$$

If instead the interval (c, d) is $(\tau, \tau + 1, \ldots \tau + \nu)$, we get in the same way a $(\nu + 1)$-dimensional ellipsoid in the Euclidean space with coordinates $x(\tau)$, $x(\tau + 1), \ldots x(\tau + \nu)$. Such a region of prognosis is not fit for applications, but we can instead confine ourselves to consider intervals

$$\{a_\mu < x(\tau + \mu) < b_\mu;\ \mu = 0, 1, \ldots \nu\}$$

and among them try to find one of maximum probability under the condition that its Lebesgue volume is fixed. We have the conditional frequency function

$$f = c\, e^{-\frac{1}{2} \sum\limits_0^{\nu} \frac{(y_{i+1} - \varrho_i y_i)^2}{1 - \varrho_i^2}},$$

where

$$\begin{cases} y_0 = x(0) \\ y_i = x(\tau + i - 1),\ i = 1, 2, \ldots \nu + 1. \end{cases}$$

and

$$\begin{cases} \varrho_0 = e^{-\beta\tau} \\ \varrho_i = e^{-\beta},\ i = 1, 2, \ldots \nu. \end{cases}$$

The problem is now reduced to finding the $(\nu + 1)$-dimensional interval with fixed volume and with maximum probability with respect to the given frequency function.

Another possibility is to take as m the absolute (non-conditional) probability P in $\Omega_{c,d}$. We still consider a process with a discrete time-parameter and suppose for simplicity that the distribution is of the continuous type. If then $S < \Omega_{c,d}$ we have

$$\begin{cases} P(S \mid \omega_{a,b}) = \int\limits_S f(\omega_{c,d} \mid \omega_{a,b})\, dv \\ P(S) = \int f(\omega_{c,d})\, dv \end{cases}$$

and we get the best region of prognosis π with respect to P

$$\pi = \left\{ \frac{f(\omega_{c,d} \mid \omega_{a,b})}{f(\omega_{c,d})} \geq k \right\} = \left\{ \frac{f(\omega_{a,b})\, f(\omega_{c,d} \mid \omega_{a,b})}{f(\omega_{a,b})\, f(\omega_{c,d})} \geq k \right\} =$$

$$= \left\{ \frac{f(\omega_{a,b};\ \omega_{c,d})}{f(\omega_{a,b})\, f(\omega_{c,d})} \geq k \right\} = \{ f(\omega_{a,b} \mid \omega_{c,d}) \geq k\, f(\omega_{a,b}) \} < \Omega_{c,d}$$

so that for a given $\omega_{a,b}$ we obtain $\pi(\omega_{a,b})$ by choosing points in $\Omega_{c,d}$ for which the conditional frequency function $f(\omega_{a,b} | \omega_{c,d})$ is large.

Take especially the Markoff process that we have just considered. We get with the same notation

$$\begin{cases} f(\omega_{c,d} | \omega_{a,b}) = k\, e^{-\frac{1}{2}\sum\limits_{0}^{\nu} \frac{(y_{i+1}-\varrho_i y_i)^2}{1-\varrho_i^2}} \\ f(\omega_{c,d}) = k_1\, e^{-\frac{1}{2} y^2{}_1 - \frac{1}{2}\sum\limits_{1}^{\nu} \frac{(y_{i+1}-\varrho_i y_i)^2}{1-\varrho_i^2}}, \end{cases}$$

and

$$\frac{f(\omega_{c,d} | \omega_{a,k})}{f(\omega_{c,d})} = k_2\, e^{-\frac{1}{2}\left\{\frac{(y_1-\varrho_0 y_0)^2}{1-\varrho^2{}_0} - y^2{}_1\right\}}.$$

Thus

$$\pi = \{|x(\tau) - e^{\beta\tau} x(0)| < k\}_{x(\tau), \ldots x(\tau+\nu)}$$

which is different from the best region of prognosis with respect to Lebesgue measure and does not contain any restriction on the values of $x(\tau+1), \ldots x(\tau+\nu)$.

6.5. Filtering as regression. Finally we shall apply the method of 6.1 on the problem of filtering of stationary processes. This problem has been treated in WIENER 1, where the filters considered are one-sided, i.e. depend only on the past. Though this is a very natural assumption in many cases, it is still plausible that in the filtering of statistical data we have usually no cause to use only the values of the process in the past. We consider the case when the realization is known in a long time-interval that can be considered infinite with regard to the effective breadth of the spectrum of the process.

Either by supposing the process to be normally distributed and using the conditional expectation, or by finding the best linear filter, we get formally the same result. Suppose that $y(t)$ is stationary with mean value zero and with the correlation function

$$r_y(t) = \int_{-\infty}^{\infty} e^{it\lambda} f_y(\lambda)\, d\lambda$$

where $f_y(\lambda)$ is a non-negative function integrable over $(-\infty, \infty)$. On $y(t)$ is superimposed a noise term $\delta(t)$ which is a stationary process with mean value zero and correlation function

$$r_\delta(t) = \int_{-\infty}^{\infty} e^{it\lambda} f_\delta(\lambda)\, d\lambda.$$

The observed process is then $x(t) = y(t) + \delta(t)$. We suppose, at first, that the noise is incoherent, i. e. that $\delta(t)$ and $y(t)$ are non-correlated. By the aid of $x(t)$, $-\infty < t < \infty$, we want to form an estimate of $y(T)$. We consider the linear combinations

$$z_n = \sum_{1}^{n} c_i^{(n)} x(t_i^{(n)}) = \int_{-\infty}^{\infty} \sum_{1}^{n} c_i^{(n)} e^{i t_i^{(n)} \lambda}\, dZ(\lambda) = \int_{-\infty}^{\infty} \gamma_n(\lambda)\, dZ(\lambda)$$

where $Z(\lambda)$ is the orthogonal process belonging to $x(t)$ (see 1.3). In order that a sequence z_n shall converge in the mean to an element $z \in L_2(X)$, it is necessary and sufficient that $\gamma_n(\lambda)$ converges in the mean to a function $\gamma(\lambda) \in L_2(F)$, where

$$F(\lambda) = E\,|Z(\lambda)|^2 = \int_{-\infty}^{\lambda} [f_y(\lambda) + f_\delta(\lambda)]\,d\lambda$$

and

$$z = \int_{-\infty}^{\infty} \gamma(\lambda)\,dZ(\lambda).$$

In accordance with the method of 6.1 we shall take $z = P_{L_2(X)}\,y(T)$ which makes $\|z - y(T)\|$ minimum subject to the condition $z \in L_2(X)$. But

$$z - y(T) = \int_{-\infty}^{\infty} \gamma(\lambda)\,dZ_y(\lambda) + \int_{-\infty}^{\infty} \gamma(\lambda)\,dZ_\delta(\lambda) - \int_{-\infty}^{\infty} e^{iT\lambda}\,dZ_y(\lambda)$$

and hence

$$\|z - y(T)\|^2 = \int_{-\infty}^{\infty} |\gamma(\lambda) - e^{iT\lambda}|^2 f_y(\lambda)\,d\lambda + \int_{-\infty}^{\infty} |\gamma(\lambda)|^2 f_\delta(\lambda)\,d\lambda.$$

But

$$\int_{-\infty}^{\infty} \{|\gamma(\lambda) - e^{iT\lambda}|^2 f_y(\lambda) + |\gamma(\lambda)|^2 f_\delta(\lambda)\}\,d\lambda$$

is minimized by

$$\gamma(\lambda) = \frac{f_y(\lambda)}{f_y(\lambda) + f_\delta(\lambda)}\,e^{iT\lambda}$$

because for each value of λ the integrand is minimized by this value of $\gamma(\lambda)$. Further this $\gamma(\lambda)$ belongs to $L_2(F)$, because $|\gamma(\lambda)| \leq 1$. The variance of the error of this filter is given by

$$\|z - y(T)\|^2 = \int_{-\infty}^{\infty} \frac{f_y(\lambda)\,f_\delta(\lambda)}{f_y(\lambda) + f_\delta(\lambda)}\,d\lambda.$$

The best filter is then

$$y^*_{\text{opt}}(T) = \int_{-\infty}^{\infty} \frac{f_y(\lambda)}{f_y(\lambda) + f_\delta(\lambda)}\,e^{iT\lambda}\,dZ(\lambda).$$

It may happen that this expression is too complicated and we can then try to approximate to $y^*_{\text{opt}}(T)$. This is easily seen to be the same as to approximate to the function $\dfrac{f_y(\lambda)}{f_y(\lambda) + f_\delta(\lambda)}$ with the quadratic metric corresponding to the weight function $f_y + f_\delta$. If e.g. we use an approximation of the form

$$\sum_{1}^{N} c_\nu\,e^{i a_\nu \lambda_\nu} \sim \frac{f_y(\lambda)}{f_y(\lambda) + f_\delta(\lambda)}$$

so that

$$\int_{-\infty}^{\infty} \left| \sum_{1}^{N} c_\nu e^{ia_\nu \lambda} - \frac{f_y(\lambda)}{f_y(\lambda) + f_\delta(\lambda)} \right|^2 [f_y(\lambda) + f_\delta(\lambda)]\, d\lambda < \varepsilon,$$

we get the approximate filter

$$y^*_{\mathrm{appr}}(T) = \sum_{1}^{N} c_\nu\, x(T + a_\nu)$$

with

$$\| y^*_{\mathrm{opt}}(T) - y^*_{\mathrm{appr}}(T) \|^2 < \varepsilon.$$

6.6. A more general problem of filtering. Let us consider the following case which is completely analogous to the usual problem of regression in a finite-dimensional space. $x(t)$ and $y(t)$ are stationary processes with the correlation functions

$$\begin{cases} r_x(t) = \int_{-\infty}^{\infty} e^{it\lambda} f_x(\lambda)\, d\lambda \\ r_y(t) = \int_{-\infty}^{\infty} e^{it\lambda} f_y(\lambda)\, d\lambda. \end{cases}$$

We suppose that they are stationarily correlated with the cross-correlation function

$$r_{yx}(t) = E\,\overline{y(s)}\, x(s+t) = \int_{-\infty}^{\infty} e^{it\lambda} f_{yx}(\lambda)\, d\lambda.$$

The only restriction made here is that the cross-correlation spectrum shall be absolutely continuous. We want to estimate the value $y(T)$ by a linear filter operating on $x(t)$. Putting

$$y^*(T) = \int_{-\infty}^{\infty} \gamma(\lambda)\, dZ_x(\lambda)$$

we have

$$\| y^*(T) - y(T) \|^2 = \| y^*(T) \|^2 + \| y(T) \|^2 - 2\, Re\, E\, y^*(T)\, \overline{y(T)} =$$

$$= \int_{-\infty}^{\infty} |\gamma(\lambda)|^2 f_x(\lambda)\, d\lambda + \int_{-\infty}^{\infty} f_y(\lambda)\, d\lambda - 2\, Re \int_{-\infty}^{\infty} \gamma(\lambda) e^{-iT\lambda} f_{yx}(\lambda)\, d\lambda.$$

We choose $\gamma(\lambda)$ so as to minimize the error $\| y^*(T) - y(T) \|$ and get

$$\gamma^*(\lambda) = \frac{\bar{f}_{yx}(\lambda)}{f_x(\lambda)} e^{iT\lambda}$$

because of

$$|\gamma|^2 f_x - 2\, Re\, \gamma\, e^{-iT\lambda} f_{yx} \geq -\frac{|f_{yx}|^2}{f_x},$$

which follows from

$$2\,Re\,\gamma e^{-iT\lambda} f_{yx} \leq 2\,|\gamma \sqrt{f_x}|\left|\frac{f_{yx}}{\sqrt{f_x}}\right| \leq |\gamma|^2 f_x + \frac{|f_{yx}|^2}{f_x}$$

and because of the choice

$$\gamma^*(\lambda) = \frac{\overline{f_{yx}(\lambda)}}{f_x(\lambda)}\, e^{iT\lambda}$$

leading to the equality

$$|\gamma^*(\lambda)|^2 f_x(\lambda) - 2\,Re\,\gamma^*(\lambda)\, e^{-iT\lambda} f_{yx}(\lambda) = -\frac{|f_{yx}(\lambda)|^2}{f_x(\lambda)}.$$

This filter function can be used because

$$\int_{-\infty}^{\infty} |\gamma^*(\lambda)|^2 f_x(\lambda)\, d\lambda = \int_{-\infty}^{\infty} \frac{|f_{yx}(\lambda)|^2}{f_x(\lambda)}\, d\lambda \leq \int_{-\infty}^{\infty} f_y(\lambda)\, d\lambda < \infty$$

using theorem 3 in Cramér 2. The error of the filter is

$$\|y^*(T) - y(T)\|^2 = \int_{-\infty}^{\infty} \frac{f_x(\lambda) f_y(\lambda) - |f_{yx}(\lambda)|^2}{f_x(\lambda)}\, d\lambda \geq 0.$$

The problem in 6.5 can be considered as a special item of this. The case of coherent noise can be treated in the same way. Let the cross-correlation be given by

$$r_{y\delta}(t) = \int_{-\infty}^{\infty} e^{it\lambda} f_{y\delta}(\lambda)\, d\lambda = E\, y(s)\, \overline{\delta(s+t)}.$$

We get the spectral intensities

$$\begin{cases} x(t) : f_y(\lambda) + f_\delta(\lambda) + 2\,Re\, f_{y\delta}(\lambda) \\ y(t) : f_y(\lambda) \\ x(t) \times y(t) : f_y(\lambda) + f_{y\delta}(\lambda) \end{cases}$$

and the filter function

$$\gamma^*(\lambda) = \frac{f_y(\lambda) + \overline{f_{y\delta}}(\lambda)}{f_y(\lambda) + f_\delta(\lambda) + 2\,Re\, f_{y\delta}(\lambda)}\, e^{iT\lambda}.$$

For $f_{y\delta} \equiv 0$ we have the previous result.

It is of some interest to note, that if T is considered as a parameter, the filters obtained are stationary transformations according to the definition of Karhunen (2).

REFERENCES.

W. Ambrose: (1) On measurable stochastic processes, Trans. Amer. Math. Soc., 1940.

S. Banach: (1) Théorie des opérations linéaires, Warsaw, 1933.

R. C. Cameron and **W. T. Martin:** (1) The behaviour of measure and measurability under change of scale in Wiener space, Bull. Amer. Math. Soc., 1947.

H. Cramér: (1) Random variables and probability distributions, Cambridge tracts in Math., Cambridge, 1937.

——: (2) On the theory of stationary random processes, Ann. of Math., 1940.

——: (3) On harmonic analysis in certain functional spaces, Arkiv f. Mat. Astr. Fys., 1942.

——: (4) Mathematical Methods of Statistics, Princeton Univ. Press, Princeton, 1946.

J. L. Doob: (1) Stochastic processes depending on a continuous parameter, Trans. Amer. Math. Soc., 1937.

——: (2) Stochastic processes with an integral-valued parameter, Trans. Amer. Math. Soc., 1938.

——: (3) The law of large numbers for continuous stochastic processes, Duke Math. Jour., 1940.

——: (4) The Brownian movement and stochastic equations, Ann. of Math., 1942.

——: (5) The elementary Gaussian processes, Ann. of Math. Stat., 1944.

J. L. Doob and **W. Ambrose:** (1) On the two formulations of the theory of stochastic processes depending upon a continuous parameter, Ann. of Math., 1940.

H. A. Einstein: (1) Der Geschiebetrieb als Wahrscheinlichkeitsproblem, Dissert., Zürich, 1937.

U. Grenander: (1) Stochastic processes and integral equations, Arkiv för Mat., 1949.

O. Hanner: (1) Deterministic and non-deterministic stationary stochastic processes, Arkiv för Mat., 1949.

E. Hopf: (1) Ergodentheorie, Ergebnisse der Math., Vol. 5, No 2, Berlin, 1937.

K. Îto: (1) On the ergodicity of a certain stationary process, Proc. Imp. Acad. Tokyo, Vol. XX, 1944.

James, Nichols, Phillips: (1) Theory of servomechanisms, New York, 1947.

M. Kac and **A. J. F. Siegert:** (1) An explicit representation of a stationary Gaussian process, Ann. of Math. Stat., 1947.

K. Karhunen: (1) Zur Spektraltheorie stochastischer Prozesse, Ann. Ac. Sci. Fennicae, A I, 34, Helsinki, 1946.

——: (2) Lineare Transformationen stationärer stochastischer Prozesse, X Skand. Mat. Kongr. København, 1946.

——. (3) Über lineare Methoden in der Wahrscheinlichkeitsrechnung, Ann. Ac. Sci. Fennicae, A I 37, Helsinki, 1947.

——: (4) Über die Struktur stationärer zufälliger Funktionen, Arkiv för Mat., 1949.

M. G. Kendall: (1) The advanced theory of statistics, I, II, London, 1943, 1946.

A. Khintchine: (1) Korrelationstheorie der stationären stochastischen Prozesse, Math. Ann., 1934.

A. Kolmogoroff: (1) Grundbegriffe der Wahrscheinlichkeitsrechnung, Ergebnisse der Math., Vol. 2, No 3, Berlin, 1933.

P. Lévy: (1) Théorie de l'addition des variables aléatoires, Paris, 1937.

O. Lundberg: (1) On random processes and their application to sickness and accident statistics, Thesis, Stockholm, 1940.

R. Paley and **N. Wiener:** (1) Fourier transforms in the complex domain, New York, 1934.

B. J. Pettis: (1) On integration in vector spaces, Trans. Amer. Math. Soc., 1938.

S. Saks: (1) Theory of the integral, Warsaw, 1937.

A. Wald: (1) Asymptotic properties of the maximum likelihood estimate of an unknown parameter of a discrete stochastic process, Ann. of Math. Stat., 1948.

N. Wiener: (1) Extrapolation, interpolation and smoothing of stationary time series, New York 1949.

J. Wolfowitz: (1) The efficiency of sequential estimates and Wald's equation for sequential processes, Ann. of Math. Stat., 1947.

Contents

Tryckt den 26 april 1950

Uppsala 1950. Almqvist & Wiksells Boktryckeri AB