STATISTICAL INFERENCE ON TIME SERIES BY HILBERT SPACE METHODS, I.

BY

EMANUEL PARZEN

TECHNICAL REPORT NO.23

PREPARED UNDER CONTRACT Nonr-225(21)

(NR-042-993)

FOR

OFFICE OF NAVAL RESEARCH

APPLIED MATHEMATICS AND STATISTICS LABORATORY

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

JANUARY 2,1959

CONTENTS

STATISTICAL INFERENCE ON TIME SERIES BY HILBERT SPACE METHODS , I

by

EMANUEL PARZEN

0. Summary.

This paper is the first of a series of projected papers on modern time series analysis, in which it is hoped to show how Hilbert space methods (which were introduced in the 1940's to clarify the probabilistic structure of time series) can be used to clarify, and to solve, various problems of statistical inference on time series. In this paper, among other things, we introduce a tool which plays a major role in our work, namely the representation of a stochastic process with finite second moments by means of a reproducing kernel Hilbert space.

The contents of the paper are as follows. In sections 1 and 2, we summarize those notions of Hilbert space theory, and of the theory of random functions, that we require. In section 3, we introduce our point of view towards the theory of Hilbert space representations of random functions by considering the special case of a random function defined on a finite interval. In section 4 we rederive from our point of view the well known theory of representation of a random function as a stochastic integral with respect to an orthogonal random set function. In section 5, we prove that any random function of second order admits a representation by a reproducing kernel Hilbert space. In section 6,

we review certain well known facts concerning projections in Hilbert spaces, and prove a useful convergence theorems for reproducing kernels. In section 7, we show how reproducing kernel Hilbert spaces may be used to give a general solution to a functional equation in Hilbert space which generalizes the notion of linear and integral equations. In section 8, the ideas developed in the paper are applied to the problem of locally minimum variance unbiased estimation. In section 9, we indicate how the previous results may be applied to Normal (or Gaussian) random functions, by giving a formula for the probability density functional of a Normal random function. In section 10, we discuss minimum variance linear unbiased estimation. In section 11, we discuss minimum variance unbiased prediction.

1. Random functions of second order.

A stochastic process (or a random process, or a random function) is best defined as a family $\{X(t), t \in T\}$ of random variables. The set T is called the index set of the process. No restriction is placed on the nature of T. However, two important cases are when $T = \{0, \pm 1, \pm 2, \ldots, \}$, the set of all integers (in which case the stochastic process is said to be a discrete parameter process) or when $T = \{-\infty < t < \infty\}$, the set of all real numbers (in which case the stochastic process is said to be a continuous parameter process).

Stochastic processes, and random functions, arise in many diverse fields of science, as the following examples indicate. In communication theory, any signal received by a radio receiver may be conceived of as a random function, since the received signal is not exactly the same as the transmitted signal, but rather differs from the transmitted signal by chance quantities, due, among other things, to the "noise" present in the electronic components of the receiver. In nuclear physics, the number $N(t)$ of electrons emitted by a radioactive source in the time interval from 0 to t is a random function of time. In technology, the diameter of the thread spun by a cotton spinning machine varies between points on the thread and constitutes a stochastic process.

In order to specify the joint probability law of a stochastic process $\{X(t),\ t \in T\}$, one needs to specify the joint probability laws of the n random variables $X(t_1),\ldots,X(t_n)$, for any integer n and any n points $t_1, t_2,\ldots, t_n$ in T. To specify the joint probability law of the n random variables $X(t_1),\ldots,X(t_n)$ one may specify either (i) the joint distribution function, for any n real numbers $x_1,\ldots,x_n$,

$$(1.1)\quad F_X(x_1,x_2,\ldots,x_n;\ t_1,t_2,\ldots,t_n) = P[X(t_1) \le x_1, X(t_2) \le x_2,\ldots,X(t_n) \le x_n]$$

or (ii) the joint characteristic function, for any n real numbers $u_1,\ldots,u_n$,

$$\varphi_X(u_1,\ldots,u_n\ ;\ t_1,\ldots,t_n) = E[\exp i\ (u_1\ X(t_1) + \ldots + u_n\ X(t_n))]$$

$$(1.2) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \exp i\ (u_1\ x_1 + \ldots + u_n x_n)\ dF_X(x_1,\ldots,x_n\ ;\ t_n,\ldots,t_n).$$

The question immediately arises of course, whether from a knowledge of these finite dimensional distributions one can answer all questions about the stochastic process which are of interest. The reader interested in examining what additional conditions need to be imposed on the stochastic process in order for this to be the case is referred to Doob (1953, Chapter II). In this paper we shall be concerned mainly with what can be said about a stochastic process which is assumed to possess finite means and second moments.

Consider a stochastic process $\{X(t), t \in T\}$, consisting of random variables defined on a probability space $(\Omega, \mathcal{A}, P)$. If each of the random variables $X(t)$ possesses a finite second moment, so that

$$E|X(t)|^2 = \int_\Omega |X(t)|^2 \, dP < \infty \text{ for every } t \text{ in } T \tag{1.3}$$

then $\{X(t), t \in T\}$ is called a random function, or a random function of second order.

Given a probability space $(\Omega, \mathcal{A}, P)$, let us define $L_2(\Omega, \mathcal{A}, P)$ to be the set of all random variables U whose second moment $E|U|^2 = \int_\Omega |U|^2 \, dP$ is finite. It may be verified that $L_2(\Omega, \mathcal{A}, P)$ is a linear space, in the sense of the following definition.

Definition of an abstract linear space: A set H, whose members $u, v, \ldots$ are usually called vectors, is a (real) linear space if

(a) for every pair of vectors u and v in H, there exists a unique vector, denoted by $u+v$, such that $u+v=v+u$ and $u+(v+w) = (u+v)+w$, for all u,v, and w in H;

(b) for every vector u in H, and every real number a, there exists a unique vector, denoted by au, such that $a(u+v) = au + av$, $(a+b)u = au + bu$, $(ab)u = a(bu)$, $1 \cdot u = u$, for all vectors u,v in H, and all real numbers a and b;

(c) there exists a zero element in H, denoted by 0, such that $u + 0 = u$, $0 \cdot u = 0$ for every u in H, where $0 \cdot u$ denotes the multiplication of the vector u by the scalar 0.

Now define the <u>scalar product</u> (U,V) between any two random variables U and V in $L_2(\Omega, \mathcal{A}, P)$ by

$$(U,V) = E[UV] = \int_\Omega UV\, dP \;. \tag{1.4}$$

It is easily verified that the space $L_2(\Omega, \mathcal{A}, P)$ is an inner product space, in the sense of the following definition.

<u>Definition of an abstract inner product space</u>: A set H is a (real) inner product space if it is a linear space, and if to every pair of points u,v in H there corresponds a real number, written (u,v) and called the inner product of u and v, such that, for all points u,v, and w in H, and every real number a, we have

(a) $(au,v) = a(u,v)$

(b) $(u+v,w) = (u,w) + (v,w)$

(c) $(v,u) = (u,v)$

(d) $(u,u) > 0$ if $u \neq 0$.

The square root of the scalar product $(u,u) = \|u\|^2$ of u with itself is called the norm $\|u\|$ of u. It may be shown to satisfy the triangle inequality, for any vectors u and v,

$$\|u+v\| \leq \|u\| + \|v\| . \tag{1.5}$$

In an abstract inner product space H, one may introduce the notion of the limit of a sequence of points $\{u_n\}$ as follows. The sequence of points $u_1, u_2, \ldots, u_n, \ldots$ is said to <u>converge</u> (<u>strongly</u>) to the point u, as $n \to \infty$, if $\lim_{n \to \infty} \|u_n - u\| = 0$; the sequence $\{u_n\}$ is then called a <u>convergent</u> sequence. It is easy to verify that the following property holds; if $\{u_n\}$ is a convergent sequence, then

$$\lim_{\substack{m \to \infty \\ n \to \infty}} \|u_m - u_n\| = 0 . \tag{1.6}$$

A sequence of points $\{u_n\}$ in H satisfying (1.6) is called a <u>Cauchy</u> sequence. The notions are now at hand to define a Hilbert space.

<u>Definition of an abstract Hilbert space</u>: A set H is a (real) Hilbert space if it is a (real) inner product space, and if it possesses the completeness property that every Cauchy sequence of points $\{u_n\}$ in H is a convergent sequence; that is, if the sequence $\{u_n\}$ satisfies

(1.6), then there is a point u in H such that $\lim_{n\to\infty} \|u_n - u\| = 0$. More concisely, we say that a Hilbert space is a complete inner product space.

It may be proved that the space $L_2(\Omega, \mathcal{A}, P)$ of all square integrable random variables on the probability space $(\Omega, \mathcal{A}, P)$ is a Hilbert space, with inner product given by equation (1.4).

Consider now a random function of second order $\{X(t), t \in T\}$ consisting of random variables defined on the probability space $(\Omega, \mathcal{A}, P)$. For each t in T, $X(t)$ can be regarded as a point in the Hilbert space $L_2(\Omega, \mathcal{A}, P)$. Hilbert space methods can then be employed to study the stochastic process. In particular, we may define the important concept of the Hilbert space spanned by a random function.

Given a random function $\{X(t), t \in T\}$, we define the linear manifold spanned by the random function $\{X(t), t \in T\}$, denoted $L(X(t), t \in T)$, to be the set of all random variables U which may be written in the form $U = \sum_{i=1}^{n} c_i X(t_i)$ for some integer n, real constants $c_1, \ldots, c_n$, and points $\{t_1, \ldots, t_n\}$ in T. In other words, $L(X(t), t \in T)$ consists of all finite linear combinations of the random variables $\{X(t), t \in T\}$. The linear manifold $L(X(t), t \in T)$ spanned by the random function $\{X(t), t \in T\}$ is a linear space.

It is clear that $L_2(\Omega, \mathcal{A}, P)$ contains every random variable which belongs to the linear manifold $L(X(t), t \in T)$ spanned by the family $\{X(t), t \in T\}$. If we define an inner product on $L(X(t), t \in T)$ by means of equation (1.4), then $L(X(t),\ t \in T)$ is an inner product subspace of $L_2(\Omega, \mathcal{A}, P)$. However $L(X(t),\ t \in T)$ does not necessarily possess the completeness property. Consequently, we enlarge the linear manifold $L(X(t),\ t \in T)$ as follows.

Given a random function of second order $\{X(t),\ t \in T\}$, define the Hilbert space spanned by it, denoted by $L_2(X(t),\ t \in T)$, to consist of all random variables in the linear manifold $L(X(t),\ t \in T)$, together with all random variables U such that there exists a sequence of random variables U_n in $L(X(t),\ t \in T)$ converging to U, in the sense that, as $n \to \infty$, $\|U_n - U\|^2 = E|U_n - U|^2 \to 0$.

It may be proved that $L_2(X(t),\ t \in T)$ is a Hilbert space. Indeed, one may define $L_2(X(t),\ t \in T)$ as the smallest subset of $L_2(\Omega, \mathcal{A}, P)$ containing the family of random variables $\{X(t),\ t \in T\}$ which possesses the properties of a Hilbert space.

The Hilbert space $L_2(X(t),\ t \in T)$ spanned by a random function of second order consists of all random variables which may be obtained by means of linear operations on the random variables $\{X(t),\ t \in T\}$. Thus $L_2(X(t),\ t \in T)$ constitutes the set of all possible linear functionals over the random variables $X(t),\ t \in T$.

An important preliminary step to the study of the problem of statistical inference on random functions of second order, and consequently one aim of this paper, is the study of the structure of the Hilbert space spanned by a random function of second order. An important role in this study is played by the notions of isomorphism and congruence which we now discuss for abstract Hilbert spaces.

Consider two abstract Hilbert spaces H_1 and H_2. Denote the inner product between two vectors u_1 and u_2 in H by $(u_1,u_2)_1$. Similarly, denote the inner product between two vectors v_1 and v_2 in H_2 by $(v_1,v_2)_2$. The Hilbert spaces H_1 and H_2 are said to be <u>isomorphic</u> if there exists a mapping ψ from H_1 onto H_2 which is one-to-one, so that the inverse ψ^{-1} exists as a mapping from H_2 onto H_1, satisfying the following properties, for any vectors u_1 and u_2 in H_1 and real number a:

$$(1.7) \qquad \psi(u_1 + u_2) = \psi(u_1) + \psi(u_2), \; \psi(au) = a\psi(u) \; .$$

A function ψ possessing these properties will be said to be an <u>isomorphism between</u> H_1 <u>and</u> H_2.

The spaces H_1 and H_2 are said to be congruent if there exists an isomorphism ψ between H_1 and H_2 which preserves inner products; that is,

$$(1.8) \qquad (u_1,u_2)_1 = (\psi(u_1), \psi(u_2))_2 \; .$$

An isomorphism ψ which satisfies equation (1.8) is said to be a congruence. A congruence not only maps linear combinations of vectors in one space into a corresponding linear combination of vectors into the other spaces (as does an isomorphism) but also maps limits into limits in the sense that, if ψ is a congruence from H_1 onto H_2, then

$$(1.9) \qquad u = \lim_n u_n \text{ if and only if } \psi(u) = \lim_n \psi(u_n) .$$

Next consider a Hilbert space H, let T be an index set, and let $\{u(t), t \in T\}$ be a family of members of H. We define the linear manifold spanned by the family $\{u(t), t \in T\}$, denoted $L(u(t), t \in T)$, to be the set consisting of all vectors u in H which may be represented in the form $u = \sum_{i=1}^{n} c_i u(t_i)$ for some integer n, some constants $c_1, \ldots, c_n$, and some points $t_1, \ldots, t_n$ in T. We define the Hilbert space spanned by the family $\{u(t), t \in T\}$, denoted $L^*(u(t), t \in T)$, to be the set of vectors which either belong to the linear manifold $L(u(t), t \in T)$ or may be represented as a limit of vectors in $L(u(t), t \in T)$.

It may happen that the Hilbert space $L^*(u(t), t \in T)$ spanned by the family of vectors $\{u(t), t \in T\}$ coincides with H. We then say that $\{u(t), t \in T\}$ <u>spans</u> H.

The following lemma (whose proof will be given in section 6) will be used often.

<u>Lemma 1a</u>: The family $\{u(t), t \in T\}$ spans H if and only if the vector $g = 0$ is the only vector in H satisfying $(g,u(t)) = 0$ for every t in T.

A family of vectors $\{u(t), t \in T\}$ in a Hilbert space H is said to be a basis for H if the family spans H, but no subset of the family spans H. It may be proved that if each of two families of vectors $B = \{u(t), t \in T\}$ and $B' = \{u'(t'), t' \in T'\}$ is a basis for H, then they have the same number of vectors (that is, B and B' may be put into one-to-one correspondence with each other).

The dimension of a Hilbert space is defined to be the number of vectors in any family of vectors $\{u(t), t \in T\}$ which is a basis for H. It is a theorem that any two abstract Hilbert spaces H_1 and H_2 which have the same dimension are congruent, and, indeed, there are a multitude of congruences between H_1 and H_2. We now prove a theorem which constitutes the crux of the proof of this assertion.

The Basic Congruence Theorem: Let H_1 and H_2 be two abstract Hilbert spaces. Let T be an index set. Let $\{u(t), t \in T\}$ be a family of vectors which span H_1. Similarly, let $\{v(t), t \in T\}$ be a family of vectors which span H_2. Suppose that, for every s and t in T,

$$(1.10) \qquad (u(s),u(t))_1 = (v(s),v(t))_2 .$$

Then the spaces H_1 and H_2 are congruent, and one can define a congruence ψ from H_1 onto H_2 which has the property that

$$(1.11) \qquad \psi(u(t)) = v(t) \quad \text{for } t \text{ in } T .$$

Proof: Define the function ψ from H_1 to H_2 as follows. For each vector in the family $\{u(t), t \in T\}$, define $\psi(u(t)) = v(t)$. For each vector u in the linear manifold $L(u(t), t \in T)$, define

$$\psi(u) = \Sigma\, c_i v(t_i) \quad \text{if} \quad u = \Sigma\, c_i u(t_i) \ . \tag{1.12}$$

We need to prove that the mapping ψ is well defined; that is, it needs to be shown that two different representations of a vector, $u = \Sigma\, c_i u(t_i) = \Sigma\, c_i' u(t_i')$, lead to the same value $\psi(u) = \Sigma c_i v(t_i) = \Sigma\, c_i' v(t_i')$. To prove this it suffices to prove that

$$\Sigma\, c_i u(t_i) = 0 \quad \text{if and only if} \quad \Sigma\, c_i v(t_i) = 0 \tag{1.13}$$

which follows from the fact that

$$0 = \|\Sigma c_i u(t_i)\|_1^2 = \Sigma c_i c_j (u(t_i), u(t_j))_1 = \Sigma c_i c_j (v(t_i), v(t_j))_2 = \|\Sigma c_i v(t_i)\|_2^2 \ .$$

From the last equation we see that ψ is a congruence from $L(u(t), t \in T)$ onto $L(v(t), t \in T)$. Consequently, it follows for any sequence $\{u_n\}$ in $L(u(t), t \in T)$ that $\{u_n\}$ is a Cauchy sequence in H_1 if, and only if $\{\psi(u_n)\}$ is a Cauchy sequence in H_2 , and $\lim_n u_n = 0$ if, and only if $\lim_n \psi(u_n) = 0$. Therefore, for $u = \lim_n u_n$, define

$\psi(u) = \lim_n \psi(u_n)$. In this way ψ is defined for every u in H_1 . One may verify that ψ is a congruence from H_1 onto H_2 .

A family of vectors $\{u(t), t \in T\}$ in a Hilbert space H is said to be orthonormal if, for every s and t in T,

$$(1.14) \qquad (u(t),u(t)) = 1, \; (u(s),u(t)) = 0 \text{ if } s \neq t .$$

By the well known Gramm-Schmidt orthogonalization method, it may be proved that any Hilbert space possesses an orthonormal basis. From this fact and the Basic Congruence Theorem one sees that any two Hilbert spaces of the same dimension are congruent. More precisely, let H_1 and H_2 be two Hilbert spaces of the same dimension. Then there exists an index set T, and families of vectors $\{u(t), t \in T\}$ and $\{v(t), t \in T\}$ which are respectively orthonormal bases for H_1 and H_2 . Then equation (1.8) holds, from which it follows that H_1 and H_2 are congruent.

2. Continuity, differentiability, and integrability of random functions.

A basic role in the study of a random function of second order is played by its covariance kernel.

Definition 2A: The covariance kernel R of the random function of second order $\{X(t), t \in T\}$ is defined to be the function on the product space $T \times T$, with value, at any s and t in T , given by

$$R(s,t) = E[X(s)X(t)] \tag{2.1}$$

In treating random functions, it is often customary to admit complex valued random variables. Then one writes

$$R(s,t) = E[X(s)\overline{X}(t)] \tag{2.2}$$

where we write $\overline{X}(t)$ to denote the complex conjugate of $X(t)$. We will use equation (2.2) in cases where it is appropriate without further comment. However, for the most part we will be dealing only with real valued random variables.

We shall see that the continuity, differentiability and integrability properties of the covariance kernel lead to corresponding properties for the random function. First, however, let us show the equivalence between covariance kernels and non-negative kernels, a class of kernels that plays an important role in functional analysis.

Definition 2B: Let T be an index set, and let R be a real-valued function of two variables defined on $T \times T$. The function (or kernel) R is called a non-negative kernel if for any integer n, any n points $\{t_1,\ldots,t_n\}$ in T, and any set of n real numbers $\{a_1,\ldots,a_n\}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \, R(t_i,t_j) \geq 0 \,. \tag{2.3}$$

The kernel R is said to be symmetric if, for all s and t in T,

$$R(s,t) = R(t,s) \quad (2.4)$$

It is to be noted that if (2.3) were required to hold for all complex numbers, then (2.3) would imply (2.4).

Theorem 2A: R is the covariance kernel of a random function if, and only if, R is a symmetric non-negative kernel.

Proof: If R is the covariance kernel of a random function $\{X(t), t \in T\}$, then it is a symmetric non-negative kernel, since $R(s,t) = E[X(s)X(t)] = R(t,s)$ and

$$\sum_{i,j=1}^{n} a_i a_j R(t_i,t_j) = E\left|\sum_{i=1}^{n} a_i X(t_i)\right|^2 \geq 0 .$$

Conversely, suppose R is a non-negative kernel. For every integer n, and any n points $\{t_1,\ldots,t_n\}$ in T, define, for any n real numbers $u_1,\ldots,u_n$,

$$\varphi_R(u_1,\ldots,u_n ; t_1,\ldots,t_n = \exp - \frac{1}{2} \sum_{i,j=1}^{n} u_i u_j R(t_i,t_j) .$$

In words, $\varphi_R(u_1,\ldots,u_n ; t_1,\ldots,t_n)$ is the characteristic function of n jointly Normally distributed random variables with means 0 and covariance matrix $\{R(t_i,t_j)\}$. For each t in T, let R_t be a real line, with Borel field $\mathcal{B}_t$. Let $(\Omega = \prod_{t\in T} R_t, \mathcal{A} = \prod_{t\in T} \mathcal{B}_t)$ be their infinite product space. Let P be the probability measure on $(\Omega, \mathcal{A})$ whose restriction to any finite dimensional subspace $R_{t_1} \times R_{t_2} \times \ldots \times R_{t_n}$ of Ω coincides with the Normal probability distribution corresponding to the characteristic function $\varphi_R(u_1, \ldots, u_n \,;\, t_1, \ldots, t_n)$. The existence of P is guaranteed by the celebrated Kolmogorov extension theorem for probability measures induced on an infinite dimensional space by a consistent family of finite dimensional probability measures (see Kolmogorov 1950 p. 29). Define the random function $\{X(t), t \in T\}$ on $(\Omega, \mathcal{A}, P)$ by, for ω in Ω, $X(t,\omega)$ is equal to the t^{th} coordinate of ω. It may be verified that the random function $\{X(t), t \in T\}$ is of second order, and has covariance kernel R. The proof of Theorem 2A is complete.

We next discuss the continuity of the covariance kernel R in the case that the index set T is a separable metric space. We could be concrete, and assume only that T is a finite interval on the real line, but this assumption is not broad enough to cover all the applications.

<u>Definition of an abstract metric space</u>: A set T of points s, t, ... is said to be a metric space if there is defined a distance d(s,t) between any two points s and t in T satisfying the conditions that

(i) $d(s,t) = d(t,s) \geq 0$,

(ii) $d(s,t) = 0$ if, and only if, $s = t$,

(iii) $d(s,t) \leq d(s,u) + d(t,u)$.

In a metric space T the notion of limit is defined in the following way; a sequence $\{t_n\}$ of points in T is said to converge to the point t, written $t_n \to t$ as $n \to \infty$, if $d(t_n,t) \to 0$ as $n \to \infty$.

Definition of a separable metric space: A metric space T is said to be separable if there exists a subset T^* of T which has a countable number of members such that for every point t in T and $\epsilon > 0$ there exists a point t' in T^* such that $|t' - t| < \epsilon$. The set T^* is said to be dense in T.

Example: Any interval a to b on the real line is a metric space, with distance $d(s,t) = |s-t|$. It is a separable metric space, since the set T^* of all rational numbers in the interval is dense in the interval.

The concepts are now at hand to define the notion of a continuous random function.

Definition 2C: A random function $\{X(t), t \in T\}$ whose index set T is a metric space is said to be continuous in quadratic mean on T (or strongly continuous) if, for every t in T,

$$\lim_{s \to t} E|X(s) - X(t)|^2 = 0 \tag{2.5}$$

where the limit in (2.5) is taken over all points s in T.

Necessary and sufficient conditions that a random function be continuous in quadratic mean are given by the following theorem.

Theorem 2B: Let T be a metric space, and let $\{X(t), t \in T\}$ be a random function of second order with covariance kernel R. The following three statements are equivalent.

(1) $\{X(t), t \in T\}$ is continuous in quadratic mean on T.

(2) R is continuous on $T \times T$; that is, for any s and t in T, $R(s',t') \to R(s,t)$ as $s' \to s$ and $t' \to t$.

(3) For every t in T, the functions $R(\cdot,t)$ and $r(\cdot)$ are continuous on T, where $R(\cdot,t)$ is the function defined on T with value at s in T equal to $R(s,t)$, and $r(\cdot)$ is the function defined on T with value at s in T equal to $R(s,s)$.

<u>Proof:</u> That (1) implies (2) follows from the facts that

$$R(s',t')-R(s,t) = E[X(s')X(t')-X(s)X(t)]$$
$$= E[\{X(s')-X(s)\}X(t')] + E[\{X(t')-X(t)\}X(s)] ,$$
$$|E[\{X(s')-X(s)\}X(t')]|^2 \leq E|X(s')-X(s)|^2 E|X(t')|^2 ,$$
$$|E[\{X(t')-X(t)\}X(s)]|^2 \leq E|X(t')-X(t)|^2 E|X(s)|^2 .$$

Next, it is immediate that (2) implies (3). Finally, to show that (3) implies (1), note that $E|X(s) - X(t)|^2 = R(t,t) - 2R(s,t) + R(s,s)$. Now, under (3), for fixed t, $R(s,s) \to R(t,t)$ and $R(s,t) \to R(t,t)$, as $s \to t$. The proof of Theorem 2B is now complete.

In treating problems of statistical inference on a random function of second order, it is often convenient to assume that the Hilbert space spanned by the random function is separable. A sufficient condition that $L_2(X(t), t \in T)$ be separable is given by the following theorem.

Theorem 2C: Let T be a separable metric space, and let $\{X(t), t \in T\}$ be continuous in quadratic mean. Then $L_2(X(t), t \in T)$ is a separable Hilbert space.

Proof: Let T^* be a countable subset of T which is dense in T. Let L^* consist of all random variables V which may be written in the form

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

for some integer n, rationals $c_1, \ldots, c_n$, and points $t_1, \ldots, t_n$ in T^*. Clearly L^* is a countable subset of $L_2(X(t), t \in T)$. Next L^* is dense in $L_2(X(t), t \in T)$, since to any $\epsilon > 0$ and U in $L_2(X(t), t \in T)$, there is a V in L^* such that $E|U-V|^2 < \epsilon$.

Strong and weak continuity: Many of the conclusions that hold for strongly continuous random functions hold also for weakly continuous random functions.

Definition 2D: A random function $\{X(t), t \in T\}$, whose index set T is a metric space, is said to be weakly continuous on T if for every t in T and U in $L_2(X(t), t \in T)$

$$\lim_{s \to t} E[UX(s)] = E[UX(t)] \tag{2.6}$$

Theorem 2D: Let T be a separable metric space, and let $\{X(t), t \in T\}$ be weakly continuous. Then $L_2(X(t), t \in T)$ is a separable Hilbert space.

Proof: Let T^* be a countable dense subset of T. We show that the random variables $\{X(t), t \in T^*\}$ span $L_2(X(t), t \in T)$ by showing that if a random variable U in $L_2(X(t), t \in T)$ satisfies the condition $E[UX(t)] = 0$ for every t in T^*, then $U = 0$, which is true, since the function $E[UX(t)]$, being continuous on T, vanishes on T if it vanishes on T^*.

The notion of weak continuity is best understood by comparing it with the notion of weak convergence. A sequence of vectors f_n in a Hilbert space H is said to converge weakly to the vector f in H, denoted $f_n \xrightarrow{W} f$, if for every g in H

$$\lim_{n\to\infty} (g, f_n) = (g, f) . \tag{2.7}$$

We next give a criterion in terms of its covariance kernel that a random function be weakly continuous. We shall use the characterization of weak convergence given by the following lemma, which we state without proof.

Lemma 2a: A sequence f_n converges weakly to some vector f if, and only if the following two conditions hold: (i) for some constant M,

$$\|f_n\| \leq M \quad \text{for all } n , \tag{2.8}$$

and (ii) for some family of vectors $\{k(t), t \in T\}$ which span $L^*(f_n, n = 1,2,\ldots)$

$$(2.9) \qquad \lim_{n\to\infty} (f_n, k(t))$$

exists for every $t \in T$.

Theorem 2E: Let T be a metric space. The random function $\{X(t), t \in T\}$ with covariance kernel R is weakly continuous if, and only if, (i) for every t in T, $R(\cdot,t)$ is continuous on T, and (ii) for every t in T, there is an open sphere $S(t)$ containing t and a constant M (depending on t) such that $R(t',t') \leq M$ for t' in $S(t)$.

Proof: By the definition of weak continuity it follows that $\{X(t), t \in T\}$ is weakly continuous if, and only if, for every s in T and sequence s_n converging to s, $X(s_n)$ converges weakly to $X(s)$ as vectors in $L_2(X(t), t \in T)$. By lemma 2a, letting $k(t) = X(t)$, this holds if, and only if, for every t in T

$$(2.10) \qquad R(s_n,t) = (X(t), X(s_n)) \to (X(t), X(s)) = R(s,t)$$

and, for some M,

$$(2.11) \qquad R(s_n,s_n) = \|X(s_n)\|^2 \leq M \quad \text{for all } n .$$

Now (2.10) is equivalent to (i), and (2.11) is equivalent to (ii) . The proof of theorem 2E is complete.

It may clarify the reader's understanding of the relation between theorem 2B and theorem 2E to point out the following general lemma connecting strong and weak convergence.

Lemma 2b: The sequence f_n converges strongly to f if, and only if, $f_n \overset{w}{\to} f$ and $\|f_n\| \to \|f\|$.

The following characterization of strong convergence is very useful.

Lemma 2c: The sequence f_n converges strongly to some vector f if, and only if, the double limit

$$\text{(2.12)} \qquad \lim_{m,n \to \infty} (f_m, f_n) \quad \text{exists.}$$

Stochastic derivatives: The tools are now at hand to define the strong and weak derivatives of the random function $\{X(t), t \in T\}$ in the case that T is a Euclidean space. In particular, let us consider the case that T is an interval on the real line.

In view of lemma 2c, it follows that $\lim_{h \to 0} \frac{X(t+h)-X(t)}{h}$ exists as a strong limit if, and only if,

$$\lim_{h,h' \to 0} E\left[\frac{X(t+h)-X(t)}{h}, \frac{X(t+h')-X(t)}{h'}\right] \quad \text{exists}$$

if, and only if,

$$\lim_{h,h'\to o} \frac{R(t+h,t+h')-R(t,t+h')-R(t+h,t)+R(t,t)}{hh'} \quad \text{exists.}$$

We thus see that the derivative

$$X'(t) = \lim_{h\to o} \frac{X(t+h)-X(t)}{h} \tag{2.13}$$

exists as a strong limit if, and only if the symmetric second partial derivative

$$\frac{\partial^2}{\partial t\,\partial t'} R(t,t') \tag{2.14}$$

exists and is finite on the line $t' = t$. It then follows that

$$E[X'(t_1)X'(t_2)] = \frac{\partial^2}{\partial t_1 \partial t_2} R(t_1,t_2) \tag{2.15}$$

$$E[X'(t_1)X(t_2)] = \frac{\partial}{\partial t_1} R(t_1,t_2) \tag{2.16}$$

We will usually consider stochastic derivatives defined as strong limits. However, let us mention conditions for the stochastic derivative to be defined as a weak limit. It follows by lemma 2a that $X'(t) = \lim_{h\to o} \frac{X(t+h)-X(t)}{h}$ exists as a weak limit if, and only if,

(i) for every t'

$$\lim_{h \to 0} E\left[\frac{X(t+h)-X(t)}{h} \cdot X(t')\right] = \lim_{h \to 0} \frac{R(t+h,t')-R(t,t')}{h} = \frac{\partial}{\partial t} R(t,t')$$

exists and (ii) there is a constant M such that, for all h,

$$E\left|\frac{X(t+h)-X(t)}{h}\right|^2 = \frac{R(t+h,t+h)+R(t,t)-2R(t,t+h)}{h^2} \leq M .$$

<u>Stochastic</u> <u>integrals</u>: Let $\{X(t), t \in T\}$ be a strongly continuous random function on an interval T. Let g be a continuous function on T such that the double integral

(2.17) $$\int_T \int_T g(t_1) g(t_2) R(t_1,t_2) dt_1 \, dt_2$$

exists as a Riemann integral. Similarly, let V be a function (of bounded variation) on T such that the double integral

(2.18) $$\int_T \int_T R(t_1,t_2) \, dV(t_1) \, dV(t_2)$$

exists as a Riemann - Stieltjes integral. Then we may define stochastic integrals

(2.19) $$\int_T g(t) X(t) dt = \lim \sum_{k=1}^{n} X(t_k) g(t_k)(t_k - t_{k-1})$$

(2.20) $$\int_T X(t) \, dV(t) = \lim \sum_{k=1}^{n} X(t_k)\{V(t_k)-V(t_{k-1})\}$$

where the limits in (2.19) and (2.20) are to be taken as strong limits over all partitions $a = t_o < t_1 < \ldots < t_n = b$ of the interval $T = [a,b]$, as $\max_{1 \leq k \leq n} |t_k - t_{k-1}|$ goes to 0. We leave it to the reader to verify that, in view of Lemma 2c, (2,19) exists if and only if (2.17) exists, and (2.20) exists if and only if (2.18) exists (see Loève, 1955, p. 472).

3. Representations of a random function defined on a finite interval.

The method we use to study random functions of second order consists in examining various concrete Hilbert spaces which are congruent to the Hilbert space spanned by the random function.

Definition 3A: A Hilbert space H is said to be a representation of a random function $\{X(t), t \in T\}$ if H is congruent to $L_2(X(t), t \in T)$.

In view of the Basic Congruence Theorem, we are justified in making the following definition.

Definition 3B: A family of vectors $\{f(t), t \in T\}$ in a Hilbert space H is said to be a representation of a random function $\{X(t), t \in T\}$ if, for every s and t in T,

$$(f(s), f(t))_H = R(s,t) = E[X(s)X(t)] \quad (3.1)$$

We will see in the sequel the usefulness of various concrete representations of a random function. In the next three sectons we examine the conditions under which various representations exist. To clarify the ideas involved, we consider in this section representations of a random function $\{X(t), t \in T\}$ of second order defined on a closed finite interval T of the real line. It will be clear however that the considerations of this section continue to hold if T is only assumed to be a closed bounded subset of a Euclidean space, since Mercer's theorem may be extended to this case (see Zaanen, p. 534).

We assume the random function $X(t)$ to be continuous in quadratic mean. Its covariance kernel R is then a symmetric non-negative kernel continuous on $T \times T$. For such a function, there is a series expansion given by Mercer's theorem, which we quote here without proof.

Mercer's Theorem: Let R be a continuous symmetric non-negative kernel on $T \times T$, where T is a closed finite interval. Let $\{\varphi_n(t), n = 1,2,\ldots\}$ be the sequence of normalized eigenfunctions of the kernel R, and $\{\lambda_n, n = 1,2,\ldots\}$ be the sequence of corresponding (non-negative) eigenvalues; that is, for all integers m and n,

(3.2) $$\int_T R(s,t)\, \varphi_n(s)ds = \lambda_n\, \varphi_n(t), t \in T$$

(3.3) $$\int_T \varphi_m(t)\, \varphi_n(t)dt = \delta_{m,n}$$

where $\delta_{m,n}$ is the Kronecker delta, equal to 1 or 0 according as $m = n$ or $m \neq n$. Then

$$R(s,t) = \sum_{n=1}^{\infty} \lambda_n \varphi_n(s) \varphi_n(t) \tag{3.4}$$

where the series converges absolutely and uniformly on $T \times T$.

By comparing equations (3.4) and (3.1), one notices that one may obtain as follows a Hilbert space H which is a representation of $\{X(t), t \in T\}$. Let H be the space of all real valued sequences $\{a_n, n = 1,2,\ldots\}$ such that

$$\sum_{n=1}^{\infty} \lambda_n a_n^2 < \infty . \tag{3.5}$$

The inner product (a,b) between two sequences $a = \{a_n, n = 1,2,\ldots\}$ and $b = \{b_n, n = 1,2,\ldots\}$ is defined by

$$(a,b) = \sum_{n=1}^{\infty} \lambda_n a_n b_n . \tag{3.6}$$

From (3.4) it follows that, for each t in T , the sequence

$$\varphi_{.}(t) = \{\varphi_n(t), n = 1,2,\ldots\} \tag{3.7}$$

belongs to H . Further

$$R(s,t) = (\varphi_{.}(s), \varphi_{.}(t)) . \tag{3.8}$$

Comparing (3.1) and (3.8) we see that the Hilbert space $L_2(X(t), t \in T)$

is congruent to the Hilbert space $L_2(\varphi_.(t), t \in T)$. Thus by defining H in the way that we have, we have obtained a Hilbert space of sequences which is a representation of the random function $\{X(t), t \in T\}$.

Now from (3.8) it follows that there is a congruence ψ from $L_2(\varphi_.(t), t \in T)$ onto $L_2(X(t), t \in T)$ such that

$$\psi(\varphi_.(t)) = X(t) \ . \tag{3.9}$$

It is of some interest to actually be able to compute (in a sense to be made clear) for any sequence a the random variable $\psi(a)$. We now discuss how this may be done.

We first show that ψ can be regarded as being defined on H . If the family $\{\varphi_.(t), t \in T\}$ spans H , then ψ is a congruence from H onto $L_2(X(t), t \in T)$ satisfying (3.9) . If $L_2\{\varphi_.(t), t \in T\}$ is a proper subspace of H , then under the assumption that $L_2(\Omega, \mathcal{A}, P)$ is infinite dimensional, we may extend ψ so that it satisfies (3.9) and is a congruence from H onto a Hilbert subspace of $L_2(\Omega, \mathcal{A}, P)$ which contains $L_2(X(t), t \in T)$

For $m = 1,2,\ldots,$ let $\epsilon_m = \{\epsilon_m(\nu), \nu = 1,2,\ldots\}$ be the sequence in H which has 1 as its m - th co-ordinate, and is 0 elsewhere; in symbols, $\epsilon_m(\nu) = \delta_{m,\nu}$ for $\nu = 1,2,\ldots$. Define the random variable

$$\xi_m = \psi(\epsilon_m) \ . \tag{3.10}$$

The sequence $\{\xi_m, m = 1,2,\ldots\}$ is an orthogonal sequence of random variables, with variances respectively equal to λ_m, since

$$(3.11) \qquad E[\xi_m \xi_n] = \sum_{\nu=1}^{\infty} \lambda_\nu \epsilon_m(\nu) \epsilon_n(\nu) = \lambda_m \delta_{m,n} .$$

Further, each random variable ξ_m belongs to $L_2(\Omega, \mathcal{A}, P)$. Let us form the Hilbert space $L_2(\xi_m, m = 1,2,\ldots)$ spanned by the sequence. From (3.11) it is clear that $L_2(\xi_m, m = 1,2,\ldots)$ consists of all random variables of the form $\sum_{n=1}^{\infty} a_n \xi_n$, where $\{a_n, n = 1,2,\ldots\}$ is a real sequence satisfying (3.5). We thus see that $L_2(\xi_m, m = 1,2,\ldots)$ and H are congruent under the congruence $I: H \to L_2(\xi_m, m = 1,2,\ldots)$, defined for any sequence $a = (a_1, a_2 \ldots)$ by

$$(3.12) \qquad I(a) = \sum_{n=1}^{\infty} a_n \xi_n .$$

Indeed from a rigorous point of view the infinite series in (3.12) is best defined as the random variable in $L_2(\xi_m, m = 1,2,\ldots)$ which corresponds to the sequence a in $H = L_2(\epsilon_m, m = 1,2,\ldots)$ under the congruence I from H onto $L_2(\xi_m, m = 1,2,\ldots)$ which satisfies

$$(3.13) \qquad I(\epsilon_m) = \xi_n .$$

That the congruence I exists follows by the Basic Congruence Theorem.

Comparing (3.10) and (3.13) we see that I and ψ coincide on H . Consequently, we obtain the conclusion that for any sequence a in H , the random varible $\psi(a)$ may be expressed in terms of the orthogonal sequence $\{\xi_m, m = 1,2,\ldots\}$ by

$$(3.14) \qquad \psi(a) = I(a) = \sum_{n=1}^{\infty} a_n \, \xi_n \quad .$$

In particular, we obtain an orthogonal decomposition of the random function $\{X(t), t \in T\}$, since

$$(3.15) \qquad X(t) = \psi(\varphi_{\cdot}(t)) = \sum_{n=1}^{\infty} \varphi_n(t) \; \xi_n \; .$$

We might regard (3.14) as an explicit formula for computing $\psi(a)$ if we had an explicit formula for computing ξ_m in terms of $X(t)$. To obtain such an explicit formula, we now show that ξ_m can be defined as a stochastic integral with respect to the random function $\{X(t), t \in T\}$.

Since, for $m = 1,2,\ldots,$ the eigenfunction φ_m , being a solution of the integral equation (3.2), is continuous on T , we may define the random variable

$$(3.16) \qquad \xi_m = \int_T \varphi_m(t) \; X(t)dt$$

as a Riemann stochastic integral. Its mean square is given by

$$(3.17)\qquad E|\xi_m|^2 = \int_T \int_T \varphi_m(t_1)\ \varphi_m(t_2)\ R(t_1,t_2)dt_1dt_2 = \lambda_m$$

where we have used (3.2) and (3.3) . Similarly one may show that $E[\xi_m\ \xi_n] = 0$ for $m \neq n$.

From the representation of $X(t)$ given by (3.15) it follows that ξ_n as defined by (3.10) coincides with ξ_n given in (3.16) , since multipling both sides of (3.15) by $\varphi_m(t)$ and integrating over T , we obtain

$$\begin{aligned}\int_T \varphi_m(t)X(t)dt &= \int_T \varphi_m(t)\Big(\sum_{n=1}^{\infty} \varphi_n(t)\xi_n\Big)dt \\ &= \sum_{n=1}^{\infty} \xi_n \int_T \varphi_n(t)\ \varphi_n(t)dt \qquad (3.18) \\ &= \xi_m .\end{aligned}$$

We leave it to the reader to justify for himself the interchange of integration and summation in (3.18) .

In view of the representation of $X(t)$ given by (3.15) it is formally appealing to define the stochastic integral for a wide class of functions g by defining

$$(3.19)\qquad \int_T X(t)g(t)dt = \sum_{n=1}^{\infty} \xi_n \int_T g(t)\varphi_n(t)dt$$

where it is assumed that g is such that all the integrals on the right hand side of (3.19) are finite. In view of (3.11), the infinite series on the right hand side of (3.19) converges if and only if the Fourier coeficients

$$g_n = \int_T g(t)\varphi_n(t)dt \tag{3.20}$$

satisfy the condition that

$$\sum_{n=1}^{\infty} \lambda_n g_n^2 < \infty \ . \tag{3.21}$$

It is clear that any function g continuous on T satisfies (3.21), since the series in (3.21) is equal to the integral in (2.17) for any function g for which the processes of summation and integration involved in (3.21) may be interchanged. However, the class of functions satisfying (3.21) does not in general include all functions g square summable with respect to Lebesque measure over T, since $\int_T g^2(t) < \infty$ is equivalent to $\sum_{n=1}^{\infty} g_n^2 < \infty$, which cannot be equivalent to (3.21), since the eigenvalues λ_n tend to ∞ with increasing n .

We can define the stochastic integral $\int_T g(t)X(t)dt$ for a wider class of functions than those for which the Fourier coefficients in (3.20) are well defined and satisfy (3.21) . Let us consider a function g on T which may be represented in the form

$$g(t) = \sum_{n=1}^{\infty} \lambda_n a_n \varphi_n(t) \tag{3.22}$$

for some sequence $\{a_n, n = 1,2,\ldots\}$ satisfying (3.5), and lying in the Hilbert subspace $L_2(\varphi(t), t \in T)$ of H. The representation of g in the form (3.22) is easily seen to be unique. For such a function g, define the stochastic integral by

$$\int_T g(t)X(t)dt = \sum_{n=1}^{\infty} a_n \xi_n \quad . \tag{3.23}$$

Let $H(R)$ denote the class of functions $g(\cdot)$ which may be represented in the form (3.23) in terms of the eigenvalues λ_n and eigenfunctions φ_n of the covariance kernel R. We now show that $H(R)$ forms a Hilbert space with certain interesting properties. Define the inner product between two functions $g(t) = \sum \lambda_n a_n \varphi_n(t)$, $h(t) = \sum \lambda_n b_n \varphi_n(t)$, where $\sum \lambda_n a_n^2 < \infty$, $\sum \lambda_n b_n^2 < \infty$ by

$$(g,h) = \sum_{n=1}^{\infty} \lambda_n a_n b_n \quad . \tag{3.24}$$

It is clear that (g,h) has all the usual properties of an inner product so that $H(R)$ is an inner product space. To show that it is a Hilbert space, one needs to show that it is complete, which is clear since if $g_m(t) = \sum \lambda_n a_n^{(m)} \varphi_n(t)$ is a Cauchy sequence then each sequence $\{a_n^{(m)}, m = 1,2,\ldots\}$ converges to a limit a_n, from which one can construct the limit $g(t) = \sum \lambda_n a_n \varphi_n(t)$ of the Cauchy sequence $g_n(t)$.

The Hilbert space H(R) has two striking properties. For each t in T , let us define $R(\cdot,t)$ to mean the function on T with value at s in T equal to $R(s,t)$. From (3.3) one sees that, for each t in T , $R(\cdot,t)$ belongs to H(R), since we may write

$$R(s,t) = \sum_{n=1}^{\infty} \lambda_n a_n \varphi_n(s), \; a_n = \varphi_n(t) \; . \tag{3.25}$$

Further, one sees that for any function g in H(R) the inner product of g with $R(\cdot,t)$ is equal to $g(t)$; more precisely for $g(\cdot)$ given by (3.22),

$$(g, R(\cdot,T)) = \sum_{n=1}^{\infty} \lambda_n a_n \varphi_n(t) = g(t) \; . \tag{3.26}$$

Because of (3.26), the Hilbert space H(R) is called a reproducing kernel Hilbert space with reproducing kernel R . Equation (3.26) is called the reproducing property of the kernel R .

The Hilbert space H(R) is a representation of the random function $\{X(t), t \in T\}$ with covariance kernel R , since from (3.26) it follows that

$$(R(\cdot,s), R(\cdot,t)) = R(s,t) \tag{3.27}$$

so that one may define a congruence J from $H(R)$ onto $L_2(X(t), t \in T)$ such that

$$(3.28) \qquad J(R(\cdot,t)) = X(t) \ .$$

It should be noted that the family $\{R(\cdot,t), t \in T\}$ spans $H(R)$ since the only function orthogonal to each of the $R(\cdot,t)$ is the zero function; for every t in T, $0 = (g,R(\cdot,t))$ implies $g(t) = 0$, in view of (3.26).

To obtain an explisit representation for J, let us first note that the congruence J may be characterized as the unique mapping from $H(R)$ into $L_2(X(t), t \in T)$ satisfying the condition that, for every g in $H(R)$ and every t in T

$$(3.29) \qquad E[J(g)X(t)] = (g,R(\cdot,t)) = g(t) \ .$$

It is clear that J satisfies (3.29). Consequently, any mapping J' satisfying (3.29) coincides with J; that is, for every g in $H(R)$, $J'(g) = J(g)$, since for any t in T $(J'(g) - J(g), X(t)) = 0$.

Using (3.29), we may prove that the congruence J may be explicitly represented as follows: for g given by (3.22)

$$(3.30) \qquad J(g) = \sum_{n=1}^{\infty} a_n \xi_n \ .$$

To prove (3.30), consider the linear mapping J', from $H(R)$ into

$L_2(X(t), t \in T)$, defined by $J'(g) = \sum_{n=1}^{\infty} a_n \xi_n$. To prove that J' coincides with J , it suffices to prove that for any g in $H(R)$ and for all t in T

$$E[J'(g)X(t)] = E[J(g)X(t)] = g(t) \quad . \tag{3.31}$$

To prove that (3.31) holds it suffices to prove that for every sequence a in $L_2(\varphi(t), t \in T)$

$$E[(\sum_{n=1}^{\infty} a_n \xi_n) X(t)] = g(t) \tag{3.32}$$

which is true since the expectation in (3.32) is equal to $\sum_{n=1}^{\infty} \lambda_n a_n \varphi_n(t)$.

<u>Summary:</u> The considerations of this section may be summarized in the following theorem.

<u>Theorem 3A:</u> Let $\{X(t), t \in T\}$ be a continuous (in quadratic mean) random function defined on a closed finite interval T . Then the following conclusions hold:

(i) The covariance kernel R possesses the expansion (3.4) ,

(ii) There exists an orthogonal sequence of random variables ξ_n in terms of which one may give by (3.15) an orthogonal decomposition for the random function $X(t)$,

(iii) There exists a Hilbert space $L_2(\varphi_.(t), t \in T)$ of sequences (a sub-space of a suitable Hilbert space H) which is a representation of the random function,

(iv) There exists a reproducing kernel Hilbert space $H(R)$, of functions on T, which is a representation of the random function.

In section 4 we show how corresponding results may be proved for a random function whose covariance kernel admits a representation generalizing that given by (3.4). In section 5 we show that result (iv) may be generalized to any random function, without assuming a representation for its covariance kernel.

4. Representations with respect to orthogonal random set functions.

Let $\{X(t), t \in T\}$ be a random function of second order, with covariance kernel R. It will often be the case that R admits a representation as follows: for every s and t in T

$$R(s,t) = \int_Q f(s)f(t)d\mu \tag{4.1}$$

where $(Q, \mathcal{B}, \mu)$ is a measure space (that is Q is an abstract set (which often will be taken to be the real line), $\mathcal{B}$ is a σ-field of subsets of Q (which often will be taken to be the σ-field of Borel sets), and μ is a measure on the measurable space $(Q, \mathcal{B})$) and $\{f(t), t \in T\}$ is a family of functions, belonging to the Hilbert space $L_2(Q, \mathcal{B}\mu)$ of all $\mathcal{B}$-measurable real valued functions f defined on Q satisfying

$$(f,f)_\mu = \int_Q f^2 d\mu < \infty . \tag{4.2}$$

It should be noted that the representation given by (3.4) is of the form of (4.1), if we let $Q = \{1,2,\ldots\}$, $\mathcal{B}$ be the family of all subsets of Q, μ be the measure on $\mathcal{B}$ satisfying $\mu(\{n\}) = \lambda_n$, and $f(t) = \varphi_{.}(t)$.

In order to generalize the orthogonal decomposition (3.15), we now introduce the notion of an <u>orthogonal random set function</u>.

<u>Definition 4 A</u>: Let $(Q,\mathcal{B},\mu)$ be a measure space, and, for every B in $\mathcal{B}$, let $Z(B)$ be a random variable in $L_2(\Omega,\mathcal{A},P)$. The family of random variables $\{Z(B), B \in \mathcal{B}\}$ is call an <u>orthogonal random set function with covariance kernel</u> μ if, for any two sets B_1 and B_2 in $\mathcal{B}$,

$$E[Z(B_1)Z(B_2)] = \mu(B_1B_2) , \tag{4.3}$$

where, as usual, B_1B_2 denotes the intersection of B_1 and B_2.

From (4.3) it follows that $Z(B)$ is finitely additive, in the sense that for any n disjoint sets $B_1,B_2,\ldots,B_n$, it holds with probability one that

$$Z(B_1 \cup B_2 \cup \ldots \cup B_n) = Z(B_1) + Z(B_2) + \ldots + Z(B_n) . \tag{4.4}$$

It suffices to prove (4.4) for $n = 2$. If $B_1B_2 = \emptyset$, then

$$\begin{aligned} E|Z(B_1)+Z(B_2)-Z(B_1\cup B_2)|^2 &= E|Z(B_1)|^2+E|Z(B_2)|^2+E|Z(B_1\cup B_2)|^2 \\ &+2E[Z(B_1)Z(B_2)]-2E[Z(B_1)Z(B_1\cup B_2)]-2E[Z(B_2)Z(B_1\cup B_2)] \\ &=\mu(B_1)+\mu(B_2)+\mu(B_1\cup B_2)+0-2\mu(B_1)-2\mu(B_2)=0 . \end{aligned}$$

However, in order to define the integral $\int_Q f\,dZ$ of a sure function f with respect to the orthogonal random set function $\{Z(B), B \in \mathcal{B}\}$, we do not use the fact that $Z(B)$ is finitely additive, but rather proceed as follows.

The Hilbert space $L_2(Z(B), B \in \mathcal{B})$, of random variables spanned by an orthogonal random set function, may be defined, as was the Hilbert space spanned by a random function, to be the smallest Hilbert subspace of $L_2(\Omega, \mathcal{A}, P)$ containing all random variables U of the form $U = \sum_{i=1}^{n} c_i Z(B_i)$ for some integer n, subfamily $\{B_1, \ldots, B_n\} \subset \mathcal{B}$, and real constants $c_1, \ldots, c_n$. On the other hand, $L_2(Q, \mathcal{B}, \mu)$ may be described as the Hilbert space spanned under the norm (4.2) by the family of indicator functions $\{I_B, B \in \mathcal{B}\}$, where for any B in $\mathcal{B}$ we define the indicator function I_B of B by $I_B(q) = 1$ or 0 according as $q \in B$ or $q \notin B$. Now for any B_1, B_2 in $\mathcal{B}$,

$$(4.5) \qquad (I_{B_1}, I_{B_2})_\mu = \mu(B_1 B_2) = E[Z(B_1)Z(B_2)] \ .$$

Therefore, by the Basic Congruence Theorem, there is a congruence ψ from $L_2(Q, \mathcal{B}, \mu)$ onto $L_2(Z(B), B \in \mathcal{B})$ such that, for any $B \in \mathcal{B}$,

$$(4.6) \qquad \psi(I_B) = Z(B) \ .$$

This fact justifies the following definition of the stochastic integral.

Definition 4B: Let $(Q, \mathcal{B}, \mu)$ be a measure space and let $\{Z(B), B \in \mathcal{B}\}$ be an orthogonal random set function with covariance kernel μ. For any function f in $L_2(Q, \mathcal{B}, \mu)$ one defines the stochastic integral of f with respect to $\{Z(B), B \in \mathcal{B}\}$, denoted $\int_Q f\, dZ$, by

$$\int_Q f\, dZ = \psi(f) , \tag{4.7}$$

where ψ is the congruence from $L_2(Q, \mathcal{B}, \mu)$ onto $L_2(Z(B), B \in \mathcal{B})$ determined by equation (4.6). The stochastic integral has the property that, for any functions f_1 and f_2 in $L_2(Q, \mathcal{B}, \mu)$,

$$\begin{aligned} &E[(\int_Q f_1\, dZ)(\int_Q f_2\, dZ)] = \int_Q f_1 f_2\, d\mu \\ &E\left|\int_Q f_1\, dZ - \int_Q f_2\, dZ\right|^2 = \int_Q |f_1 - f_2|^2\, d\mu . \end{aligned} \tag{4.8}$$

We now turn to the problem of determining conditions on a random function $\{X(t), t \in T\}$ under which there exists an orthogonal random function $\{Z(B), B \in \mathcal{B}\}$ such that one may write

$$X(t) = \int_Q f(t)\, dZ . \tag{4.9}$$

<u>It should be noted that (4.9) can only be expected to hold with probability one</u>. A similar remark should be made for all equations below of the form of (4.9).

From (4.8) one sees that if (4.9) holds, then (4.2) holds. That conversely (4.2) implies (4.9) was originally proved by Karhunen in his 1947 thesis.

Theorem 4A: Let $\{X(t), t \in T\}$ be a random function of second order, with covariance kernel R, such that, for every $t \in T, X(t)$ is a random variable defined on the probability space $(\Omega, \mathcal{A}, P)$. Let $(Q, \mathcal{B}, \mu)$ be a measure space such that the dimension of $L_2(\Omega, \mathcal{A}, P)$ is greater than or equal to the dimension of $L_2(Q, \mathcal{B}, \mu)$. If there exists a family of functions $\{f(t), t \in T\}$ in $L_2(Q, \mathcal{B}, \mu)$ such that (4.1) holds then there exists an orthogonal random set function $\{Z(B), B \in \mathcal{B}\}$ with covariance kernel μ such that for every t in T there is a function $f(t)$ in $L_2(Q, \mathcal{B}, \mu)$ of which $X(t)$ is the stochastic integral, in the sense that

$$(4.10) \qquad X(t) = \int_Q f(t)\, dZ .$$

It further follows that

$$(4.11) \qquad L_2(X(t), t \in T) \subseteq L_2(Z(B), B \in \mathcal{B})$$

with equality in (4.11) if, and only if, the functions $\{f(t), t \in T\}$ span $L_2(Q, \mathcal{B}, \mu)$.

Proof: From the fact that (4.1) holds, it follows by the Basic Congruence Theorem that there is a congruence ψ from $L_2(f(t), t \in T)$

onto $L_2(X(t), t \in T)$ such that $\psi(f(t) = X(t)$. If it is not already the case that $L_2(f(t), t \in T) = L_2(Q, \mathcal{B}, \mu)$, extend (how?) the congruence ψ so that it maps $L_2(Q, \mathcal{B}, \mu)$ onto a Hilbert subspace of $L_2(\Omega, \mathcal{A}, P)$, containing $L_2(X(t), t \in T)$, and still possessing the property that $\psi(f(t)) = X(t)$ for t in T . Now, for $B \in \mathcal{B}$, define $Z(B) = \psi(I_B)$. Clearly $E[Z(I_{B_1})Z(I_{B_2})] = \mu(B_1 B_2)$, so that the family of random variables $\{Z(B), B \in \mathcal{B}\}$ is an orthogonal random set function with covariance kernel μ . By the definition of the stochastic integral, equation (4.10) is merely another way of writing the fact that $X(t) = \psi(f(t$ The proof of Theorem 4A is complete.

In most applications of Theorem 4A, it will be the case that $L_2(\Omega, \mathcal{A}, P)$ has dimension greater than or equal to that of $L_2(Q, \mathcal{B}, \mu)$. This is the case, for example, if the family $\{f(t), t \in T\}$ spans $L_2(Q, \mathcal{B}, \mu)$. However, Theorem 4A remains valid, if (4.2) holds, even if the dimension of $L_2(\Omega, \mathcal{A}, P)$ is less than that of $L_2(Q, \mathcal{B}, \mu)$ if one interprets Theorem 4A in the following extended sense; there exists (1) a product probability space $(\Omega^*, \mathcal{A}^*, P^*)$ of the form $\Omega^* = \Omega \times \Omega'$, $\mathcal{A}^* = \mathcal{A} \times \mathcal{A}'$, and $P^* = P \times P'$, for some probability space $(\Omega', \mathcal{A}', P')$, such that $L_2(\Omega^*, \mathcal{A}^*, P^*)$ has the same dimension as $L_2(Q, \mathcal{B}, \mu)$, and (2) an orthogonal random set function $\{Z(B), B \in \mathcal{B}\}$ with covariance kernel μ , consisting of random variables $Z(B)$ defined on $(\Omega^*, \mathcal{A}^*, P^*)$, such that, at any point (ω, ω') in Ω^* , the random variable $\int_Q f(t)\, dZ$ has value equal to the value of $X(t)$ at ω . To see this, merely apply Theorem 4A to the random function $\{X^*(t), t \in T\}$ defined on

$(\Omega^*, \mathcal{Q}^*, P^*)$ by $X^*(t,\omega^*) = X(t,\omega)$ if $\omega^* = (\omega,\omega')$.

An alternative formulation of Theorem 4A which we leave as an exercise for the reader is the following.

Theorem 4B: Let $\{X(t), t \in T\}$ be a random function of second order with covariance kernel R . Let $(Q, \mathcal{B})$ be a measurable space, and $\{f(t), t \in T\}$ be a family of $\mathcal{B}$-measurable functions. In order that there exists a measure μ on $\mathcal{B}$ and an orthogonal random set function $\{Z(B), B \in \mathcal{B}\}$ with covariance kernel μ such that (4.10) holds (possibly in the extended sense mentioned in the preceding paragraph) it is sufficient that there is a measure μ on $\mathcal{B}$ such that (4.1) holds.

The usefulness of an orthogonal decomposition, of the form of (4.10), of a random function depends largely on the tractability of the functions $f(t)$ which occur in the decomposition. For example, if the functions $f(t)$ are the complex exponential functions, so that $f(t,q) = e^{itq}$, then (4.10) is the very useful spectral decomposition of a random function.

It is clear that a random function $\{X(t), t \in T\}$ possesses a spectral decomposition,

$$X(t) = \int_Q e^{itq} \, dZ(q) \tag{4.12}$$

where Q is some interval on the real line if, and only if, its covariance kernel R satisfies for all s and t in T

$$R(s,t) = \int_Q e^{iq(s-t)}\, d\mu(q) \tag{4.13}$$

for some measure μ . From (4.13) one sees that a necessary condition for the representation (4.12) to hold is that $R(s,t)$ be a function only of $(s-t)$; a random function whose covariance kernel satisfies this condition is said to be <u>stationary in the wide sense</u>. Using the classical theorems of Herglotz and Bochner (see Loève, 1955, p. 207) it may be shown that in the case that $T = \{0, \pm 1, \pm 2, \ldots\}$ or $T = \{-\infty < t < \infty\}$, the condition that the random function be stationary in the wide sense is sufficient, as well as necessary, for the random function to have a spectral decomposition with respect to an orthogonal random set function.

On the other hand, if we are not fussy about the functions $f(t)$ which appear in the orthogonal decomposition (4.10), than a multitude of such decompositions always exist.

<u>Theorem 4C</u>: Let $\{X(t), t \in T\}$ be a random function with covariance kernel R . Let $(Q, \mathcal{B}, \mu)$ be a measure space such that $L_2(X(t), t \in T)$ and $L_2(Q, \mathcal{B}, \mu)$ are congruent. Then, in a multitude of ways, there exists an orthogonal random set function $\{Z(B), B \in \mathcal{B}\}$ with covariance kernel μ such that (i) the random function $X(t)$ may be represented, in the form of (4.10), as a stochastic integral with respect to $Z(B)$, and (ii) the covariance kernel R may be represented in the form (4.1) with respect to μ .

Proof: Since $L_2(X(t), t \in T)$ and $L_2(Q, \mathcal{B}, \mu)$ are congruent, there exists a multitude of congruences between them. Let ψ be a congruence from $L_2(Q, \mathcal{B}, \mu)$ onto $L_2(X(t), t \in T)$. Define, for $B \in \mathcal{B}, Z(B) = \psi(I_B)$. Then $\{Z(B), B \in \mathcal{B}\}$ is an orthogonal random set function with covariance kernel μ. Further, since $\{I_B, B \in \mathcal{B}\}$ spans $L_2(Q, \mathcal{B}, \mu)$, the family $\{Z(B), B \in \mathcal{B}\}$ spans $L_2(X(t), t \in T)$. Consequently, $L_2(X(t), t \in T) = L_2(Z(B), B \in \mathcal{B})$, and equation (4.10) follows, where $f(t)$ is the vector in $L_2(Q, \mathcal{B}, \mu)$ such that $\psi(f(t)) = X(t)$.

We have so far generalized results (i)-(iii) of Theorem 3A. We next generalize result (iv) by obtaining the reproducing kernel Hilbert space of functions defined on T which is a representation of the random function $\{X(t), t \in T\}$ whose covariance kernel has the representation (4.2).

Theorem 4D: Let (Q, B, μ) be a measure space, and let $\{f(t), t \in T\}$ be functions in $L_2(Q, \mathcal{B}, \mu)$ such that (4.1) holds. Let $H(R)$ be the space of functions g, defined on T, which may be represented in the form

$$g(t) = \int_Q g^* f(t) d\mu \tag{4.14}$$

for some (necessarily unique) function g^* in the Hilbert subspace $L_2(f(t), t \in T)$ of $L_2(Q, \mathcal{B}, \mu)$ spanned by the family of functions

$\{f(t), t \in T\}$. The norm of g is given by

$$(4.15) \qquad \|g\|^2 = \int_Q |g^*|^2 \, d\mu \ .$$

Then H(R) is a reproducing kernel Hilbert space (in the sense that (4.19) and (4.20) hold). Further there is a congruence J from H(R) onto $L_2(X(t), t \in T)$ such that

$$(4.16) \qquad X(t) = J(R(\cdot,t)) \ ,$$

for every g in H(R) and t in T

$$(4.17) \qquad g(t) = (g, R(\cdot,t)) = (J(g), X(t));$$

and in terms of the orthogonal decomposition (4.10)

$$(4.18) \qquad J(g) = \int_Q g^* \, dZ \ .$$

Proof: One may easily verify that H(R) is a Hilbert space possessing the properties, that, for every t in T and g in H(R)

$$(4.19) \qquad R(\cdot,t) \in H(R)$$

$$(4.20) \qquad (g, R(\cdot,t)) = g(t) \ .$$

Consequently $H(R)$ is a reproducing kernel Hilbert space with reproducing kernel R. That $H(R)$ is congruent to $L_2(X(t), t \in T)$ follows from the fact that the functions $\{R(\cdot,t), t \in T\}$ span $H(R)$, and in view of (4.20), $R(s,t) = (R(\cdot,s), R(\cdot,t))$. It remains only to prove that the mapping $J' : H(R) \to L_2(X(t), t \in T)$ defined (for the function g of the form (4.14)) by

$$(4.21) \qquad J'(g) = \int_Q g^* \, dZ$$

coincides with the congruence J from $H(R)$ onto $L_2(X(t), t \in T)$ satisfying (4.16). To prove that $J'(g) = J(g)$ for every g in $H(R)$, it suffices to prove that

$$(4.22) \qquad E[J'(g)X(t)] = E[J(g)X(t)] = g(t)$$

for every g in $H(R)$ and t in T. But

$$(4.23) \quad E[J'(g)X(t)] = E[\int_Q g^* dZ \int_Q f(t) dZ] = \int_Q g^* f(t) d\mu = g(t) .$$

The proof of Theorem 4D is now complete.

5. Representations by reproducing kernel spaces.

A Hilbert space H is said to be a reproducing kernel Hilbert space, with reproducing kernel K, if the members of H are functions on some set T , and if there is a kernel K on $T \times T$ having the following two properties; for every t in T (where $K(\cdot,t)$ is the function defined on T , with value at s in T equal to $K(s,t)$):

$$(5.1) \qquad K(\cdot,t) \in H$$

$$(5.2) \qquad (g,K(\cdot,t)) = g(t)$$

for every g in H .

In the foregoing we have considered mainly representations of random functions by means of L_2-spaces (Hilbert spaces of square integrable functions on some measure space) . Now, in view of (5.2), a reproducing kernel Hilbert space has more structure than an L_2-space. Consequently, we shall make much use of the fact that random functions can be represented by reproducing kernel Hilbert spaces. That this is the case is a consequence of the following basic theorem of the theory of reproducing kernels (see Aronszajn (1943, 1950) for a complete exposition of the theory of reproducing kernel Hilbert spaces) .

Theorem 5A (E.H.Moore): A symmetric non-negative kernel K generates a unique Hilbert space, which we denote by $H(K)$, of which K is the reproducing kernel.

Proof: At the end of the section we sketch the proof of Theorem 5A given by Aronszajn. We here give a proof involving random functions. In view of Theorem 2A let $\{X(t), t \in T\}$ be a random function of which K is the covariance kernel. Let $H(K)$ be the family of functions g on T of the form

$$g(t) = E[UX(t)] \tag{5.3}$$

for some (necessarily unique) random variable U in $L_2(X(t), t \in T)$. Define the norm of g by

$$\|g\|^2 = E|U|^2 \quad . \tag{5.4}$$

It is easily verified that $H(K)$ is a Hilbert space of which K is the reproducing kernel. That $H(K)$ is unique follows by the following theorems.

Theorem 5B: If K is a reproducing kernel for the Hilbert space H, then the family of functions $\{K(\cdot,t), t \in T\}$ span H.

Proof: To prove the theorem it suffices to prove that the only vector g in H orthogonal to each function $K(\cdot,t)$ is the zero function; but this is obvious, since by the reproducing property $(g,K(\cdot,t)) = 0$ for every t in T implies $g(t) = 0$ for all t.

Theorem 5C: If K is the reproducing kernel of two Hilbert spaces H and H' then H and H' are identical.

Proof: By the Basic Congruence Theorem, there exists a congruence ψ between H and H' under which $K(\cdot,t)$ in H corresponds to $K(\cdot,t)$ in H'. We now show that if g in H corresponds under ψ to g' in H', then $g(t) = g'(t)$ for every t in T. By the reproducing property,

$$(5.5) \qquad g(t) = (g,K(\cdot,t))_H = (g',K(\cdot,t))_{H'} = g'(t) \ .$$

The proof of Theorem 5C is now complete.

By the Basic Congruence Theorem we now obtain the result, due to Loève (1948,pp.338-341) that any random function of second order possesses a representation by a reproducing kernel Hilbert space.

Theorem 5D: Let $\{X(t), t \in T\}$ be a random function with covariance kernel R. Then $L_2(X(t), t \in T)$ is congruent to the reproducing kernel Hilbert space $H(R)$. Further, any linear map J from $H(R)$ into $L_2(X(t), t \in T)$ which has the property that for any g in $H(R)$ and any t in T

$$(5.6) \qquad E[J(g)X(t)] = g(t)$$

is the congruence from $H(R)$ onto $L_2(X(t), t \in T)$ which maps

$R(\cdot,t)$ into $X(t)$.

It is to be again emphasized that the congruence J enjoys many of the properties of an integration operator.

We shall show in section 7 how one uses Theorem 5D. In the remainder of this section we discuss some of the properties of $H(R)$.

The space $H(R)$ will generally consist of continuous functions, in view of the following theorem:

Theorem 5E: Let T be a metric space, and let $\{X(t)t \in T\}$ be a random function with covariance kernel R . Then $\{X(t),t \in T\}$ is weakly continuous if, and only if, every function g in $H(R)$ is continuous on T .

Proof: Theorem 5E is an immediate consequence of the definitions of the notions involved, and the fact that for any g in $H(R)$, and s and t in T,

$$g(s)-g(t) = (g,R(\cdot,s)-R(\cdot,t)) = E[J(g)(X(s)-X(t))] .$$

Indeed $H(R)$ will often consist of differentiable functions, in view of the following theorem whose proof we leave to the reader.

Theorem 5F: Let T be an interval on the real line, and let $\{X(t),t \in T\}$ be a random function with covariance kernel R. Then, for every t in T, $X'(t)$ exists as a weak derivative if, and only if, every function g in $H(R)$ is differentiable on T.

Another property of reproducing kernel Hilbert spaces which is noteworthy is the following.

Theorem 5G: Let H be a reproducing kernel Hilbert space consisting of functions defined on T. If a sequence of functions f_n in H converge weakly (or strongly) to a function f, then the functions converge pointwise (that is, $f_n(t) \to f(t)$ for every t in T).

We conclude this section by sketching the idea of Aronszajn's proof of Theorem 5A. Let F be the linear manifold of functions on T spanned by the family $\{K(\cdot,t), t \in T\}$. In symbols,

$$F = \{f \text{ on } T: f = \sum_{j=1}^{n} a_j K(\cdot,t_j)\} . \tag{5.7}$$

On F, define an inner product (f,g) by

$$(f,g) = \sum_{j=1}^{n} \sum_{k=1}^{n'} K(t_k', t_j) a_j a_k' \tag{5.8}$$

where

$$f = \sum_{j=1}^{n} a_j K(\cdot,t_j), g = \sum_{k=1}^{n'} a_k' K(\cdot,t_k') . \tag{5.9}$$

To see that the inner product (5.8) is well defined, in the sense that its value does not depend on the particular representations

(5.9) used of the functions f and g, one need only note that one may write

$$(5.10) \qquad (f,g) = \sum_{j=1}^{n} a_j g(t_j) = \sum_{k=1}^{n'} a_k' f(t_k') \quad .$$

It is clear that the reproducing property (5.2) holds for functions g in F. Consequently if $\{f_n, n = 1,2,\ldots\}$ is a Cauchy sequence of functions in F, so that $\|f_n - f_m\| \to 0$, then, for every t in T, $\lim_{n \to \infty} f_n(t)$ exists, since

$$|f_n(t)-f_m(t)|^2 \leq |(f_n-f_m, K(\cdot,t))|^2 \leq \|f_n-f_m\| \; \|K(\cdot,t)\| \to 0 \ .$$

If _we define_ $H(K)$ _to consist of all functions which are either in_ $\underline{F}$ _or are pointwise limits of Cauchy sequences in_ $\underline{F}$, then one may verify that $H(K)$ is a Hilbert space with reproducing kernel K.

6. Projection and least squares linear prediction.

A basic problem of statistical inference, which as we shall see is also basic to the study of the structure of a random function, is that of _least squares linear prediction_. In the problem of prediction, one considers a random variable Z, and a random function $\{X(t), t \in T\}$, and one seeks that random variable in $L_2(X(t), t \in T)$ whose mean square distance from Z is smallest. In other words, if one desires to predict the value of Z on the basis of having observed the values of the family of random variables $\{X(t), t \in T\}$, one method might be to taken that linear functional in the observations whose mean square error as a predictor is least. In this section we discuss certain well known results which prove the existence of, and characterize, the best predictor.

Given a random function $\{X(t), t \in T\}$ and a random variable Z, one proves the existence of a random variable in $L_2(X(t), t \in T)$ which is closest to Z, among all random variables in $L_2(X(t), t \in T)$, by proving more generally the following theorem.

Theorem 6A: Let H be an abstract Hilbert space, let M be a Hilbert subspace of H, and let v be a vector in H. Let, for any vector v in H,

$$d[v|M] = \underset{\text{over all } u \text{ in } M}{\text{infimum}} \|u-v\| . \tag{6.1}$$

Then there exists a unique vector in M, denoted by $E^*[v|M]$, which satisfies each of the following equivalent conditions:

$$\|E^*[v|M]-v\| = d[v|M] = \min_{u \in M} \|u-v\| \tag{6.2}$$

$$(E^*[v|M]-v, u) = 0 \qquad \text{for every } u \text{ in } M \tag{6.3}$$

$$(E^*[v|M], u) = (v, u) \qquad \text{for every } u \text{ in } M . \tag{6.4}$$

Proof: We prove first that (6.2),(6.3), and (6.4) are equivalent. It is clear that (6.3) and (6.4) are equivalent. That (6.3) implies (6.2) follows by the fact that for any u in M not equal to v

$$(6.5) \quad \|u-v\|^2=\|u-E^*[v|M]\|^2+\|v-E^*[v|M]\|^2+2(u-E^*[v|M],v-E^*[v|M]) > \|v-E^*[v|M]\|^2 .$$

Conversely, to prove that (6.2) implies (6.3), we use the fact that for any real number a and vector u in M (so that $au+E^*[v|M]$ is in M)

$$(6.6) \quad d^2[v|M] \leq \|E^*[v|M] -v+au\|^2=\|E^*[v|M]-v\|^2+a^2\|u\|^2+2a(u,E^*[v|M]-v) .$$

Consequently, $0 \leq a^2\|u\|^2+ 2a(u,E^*[v|M]-v)$. Letting $a=b(u,E^*[v|M]-v)$, where b is an arbitrary real number, it follows that, for every real number b ,

$$(6.7) \quad b^2\|u\|^2|(u,E^*[v|M]-v)|^2+ 2b|(u,E^*[v|M]-v)|^2 \geq 0 .$$

Now $b^2\|u\|^2+ 2b$ is negative for b in the range $-(1/2\|u\|^2) < b < 0$. Consequently, from (6.7) one infers (6.3) .

To complete the proof of Theorem 6A, we need to prove that there exists a unique vector in M whose distance from v is equal to $d[v|M]$, which we denote by d . We first prove that there exists such a vector. Let $\{u_n\}$ be a sequence of vectors in M such that $\|v-u_n\| \to d$ as $n \to \infty$. We prove that $\{u_n\}$ is a Cauchy sequence, from which it follows that there is a vector u_o in M such that $\|u_n-u_o\| \to 0$ and therefore $\|v-u_o\| = d$. It follows from the parallelogram law (which states for any vectors x and y in H

$$\|x-y\|^2 + \|x+y\|^2 = 2\|x\|^2 + 2\|y\|^2)$$

that, for every n and m,

$$\|u_n-u_m\|^2 = 2\|u_n-v\|^2 + 2\|u_m-v\|^2-4\| \tfrac{1}{2} (u_n+ u_m)-v\|^2 .$$

Since $\frac{1}{2}(u_n+u_m)$ belongs to M, it follows that $\|\frac{1}{2}(u_n+u_m)-v\|^2 \geq d^2$, and hence that

$$\|u_n-u_m\|^2 \leq 2\|u_n-v\|^2+ 2\|u_m-v\|^2-4 d^2 .$$

As n and m tend to ∞ independently, the right side of the last written inequality tends to $2d^2 + 2d^2 - 4d^2 = 0$. Consequently, $\{u_n\}$ is a Cauchy sequence, and we have proved that there is a vector u_o in M such that $\|v-u_o\| = d$.

We next prove that the vector in M, with distance from v equal to d, is unique. Suppose there were two such vectors u_1 and

u_2 . By the parallelogram law,

$$\|u_1-u_2\|^2 = 2\|u_1-v\|^2 + 2\|u_2-v\|^2-4\| \frac{1}{2} (u_1+ u_2)-v\|^2 \quad .$$

Under the assumptions on u_1 and u_2 it follows that $\|u_1-u_2\|^2 \leq 0$, which implies that $\|u_1-u_2\| = 0$ and $u_1 = u_2$. The proof of Theorem 6A is now complete.

Definition 6A: The vector $E^*[v|M]$ is called the projection of the vector v on the subspace M . If M is the Hilbert space $L^*(k(t),t \in T)$ spanned by a family $\{k(t),t \in T\}$ of vectors in H , we write $E^*[v|k(t),t \in T]$ for $E^*[v|L^*(k(t),t \in T)]$.

In probability theory the conditional expectation $E[U|\mathscr{M}]$ of a random variable U relative to a σ-field $\mathscr{M}$ is defined to be the unique vector which is measurable with respect to the σ-field $\mathscr{M}$, and satisfies the condition

$$E[E[U|\mathscr{M}]V] = E[UV]$$

for all bounded random variables V measurable with respect to $\mathscr{M}$. Our choice of the notation $E^*[u|M]$ to denote the projection of the vector u on the Hilbert subspace M is motivated by the fact that projection has many of the properties of conditional expectation. It is for this reason that Doob (1953,p.155) calls projection "wide-sense conditional expectation."

In the case that we are dealing with projections in a Hilbert space consisting of square integrable functions on a probability space, there is a necessary and sufficient condition due to Bahadur (1956) that a projection be true conditional expectation.

<u>Lemma 6a</u>: Let $(\Omega,\mathcal{A},P)$ be a probability space, and let M be a Hilbert subspace of $L_2(\Omega,\mathcal{A},P)$. Let $\mathcal{M}$ be the smallest σ-field such that each function in M is $\mathcal{M}$-measurable. A necessary and sufficient condition that for every U in $L_2(\Omega,\mathcal{A},P)$

$$(6.8) \qquad E^*[U|M] = E[U|\mathcal{M}]$$

with probability one is that (i) $1 \in M$, where 1 is the function in $L_2(\Omega,\mathcal{A},P)$ identically equal to 1, and (ii) for every U in $L_2(\Omega,\mathcal{A},P)$

$$(6.9) \qquad U \geq 0 \quad \text{implies} \quad E^*[U|M] = 0 \ .$$

It should be noted that the projection $E^*[v|k(t),t \in T]$ can be characterized as the unique vector in $L^*(k(t),t \in T)$ satisfying the condition

$$(6.10) \qquad (E^*[v|k(t),t \in T],k(t)) = (v,k(t)) \quad \text{for every} \quad t \in T \ .$$

From (6.10) we obtain a proof of Lemma 1a, since $H = L^*(k(t), t \in T)$ if, and only if, for every vector v in H, $v = E^*[v|k(t), t \in T]$ which is so if, and only if, $g = 0$ is the only vector in H satisfying $(g, k(t)) = 0$ for every $t \in T$. From (6.10) we also obtain the following useful result.

Lemma 6a: Let H_1 and H_2 be two abstract Hilbert spaces, and let ψ be a congruence from H_1 onto H_2. Then, for any vector v in H_1 and family of vectors $\{k(t), t \in T\}$ in H_1

$$\psi(E^*[v|k(t), t \in T]) = E^*[\psi(v)|\psi(k(t)), t \in T] .$$

Proof: Verify that, for any t in T,

$$(\psi(E^*[v|k(t), t \in T]), \psi(k(t))) = (E^*[v|k(t), t \in T], k(t))$$

$$= (v, k(t)) = (\psi(v), \psi(k(t))) .$$

In the study of the properties of projection, a basic role is played by (6.4). For example, from (6.4) we may deduce the following smoothing property, which states that the projection of a projection can in certain cases be expressed as a projection.

Theorem 6B: Let M_1 and M_2 be Hilbert subspaces of an abstract Hilbert space H such that M_1 is a subspace of M_2; then for any vector v in H,

$$E^*[E^*[v|M_2]|M_1] = E^*[E^*[v|M_1]|M_2] = E^*[v|M_1] \quad .$$

Proof: By (6.4), it suffices to prove that, for every u in M_1, $(E^*[E^*[v|M_1]|M_2],u) = (v,u)$ which is clear, since one has

$$\begin{aligned}(E^*[E^*[v|M_1]|M_2],u) &= (E^*[v|M_1],u) \quad \text{for every } u \text{ in } M_2 \\ &= (v,u) \quad \text{for every } u \text{ in } M_1 \subset M_2 \quad .\end{aligned}$$

Similarly, it may be proved that $E^*[E^*[v|M_2]|M_1] = E^*[v|M_1]$.

We next use (6.4) and the Basic Congruence Theorem to obtain the following importent result characterizing the reproducing kernel Hilbert space corresponding to the restriction of a reproducing kernel.

Theorem 6C: Let T be an abstract set, and let $T' \subset T$. Let K_T be a symmetric non-negative kernel defined on $T \times T$, and let $K_{T'}$ be the restriction of K_T to $T' \times T'$. Then $H(K_{T'})$ consists of all functions f' which are restrictions to T' of functions f belonging to $H(K_T)$. Further, for any f' in $H(K_{T'})$, and for any function f in $H(K_T)$ such that

$$(6.11) \qquad f'(t) = f(t) \quad \text{for } t \text{ in } T' ,$$

it follows that

$$(6.12) \qquad \|f'\|_{H(K_{T'})} = \|E^*[f|K_T(\cdot,t),t\epsilon T']\|_{H(K_T)} \leq \|f\|_{H(K_T)} .$$

<u>Proof</u>: Since for s and t in T'

$$(6.13) \qquad (K_T(\cdot,s),K_T(\cdot,t)) = K_T(s,t) = K_{T'}(s,t)$$

it follows that there is a congruence ψ from $H(K_{T'})$ onto $L^*(K_T(\cdot,t),t \in T')$ such that

$$(6.14) \qquad \psi(K_{T'}(\cdot,t)) = K_T(\cdot,t) \quad \text{for } t \text{ in } T' .$$

Consequently, for any f' in $H(K_{T'})$, and t in T', letting $f = \psi(f')$,

$$(6.15) \qquad f'(t) = (f',K_{T'}(\cdot,t)) = (f,K_T(\cdot,t)) = f(t) .$$

From (6.15) we see that any f' in $H(K_{T'})$ is a restriction to T' of a function f in $H(K_T)$. Further, by (6.4) and (6.15) it follows that for any f' in $H(K_{T'})$ and for any f in $H(K_T)$ such that (6.11) holds,

$$(6.16) \qquad \psi(f') = E[f|K_T(\cdot,t),t \in T'] .$$

From (6.16) one sees that (6.12) holds.

We next consider some useful convergence theorems for sequences of projections, predictions, and reproducing kernels.

Let us first consider a sequence of predictions. Suppose one has observed an infinite sequence of random variables $\{X(t), t = N, N-1, \ldots, -1, -2, \ldots\}$, for some integer N. To predict the value of a random variable Z, one may use all or part of the observations. Now suppose one were to form a sequence of predictions of Z which, at each stage, utilized more of the information available; that is, one considers the sequence of predictors, for $n = 1, 2, \ldots, E^*[Z|X(t), N-n \leq t \leq N]$. It seems reasonable to expect that the predictors over the finite past tend to the predictor over the infinite past; that is,

$$(6.18) \qquad E^*[Z|X(t), t \leq N] = \lim_{n \to \infty} E^*[Z|X(t), N-n \leq t \leq N] \quad .$$

That (6.18) holds follows by the following theorem (due to Doob, 1953, p. 164) on the convergence of a sequence of vectors Z_n satisfying (6.19). Such a sequence is called a <u>martingale</u> in the wide sense.

<u>Theorem 6D:</u> Let $\{M_n, n = 1, 2, \ldots\}$ be a sequence of Hilbert subspaces of H which are either (i) monotone non-decreasing; that is, $M_n \subset M_{n+1}$, or (ii) monotone non-increasing; that is $M_n \supset M_{n+1}$.

Define M_∞ to be, in case (i), the Hilbert subspace of H spanned by the union $\bigcup_n M_n$, and, in case (ii), the intersection $\bigcap_{n=1}^{\infty} M_n$. Let $Z_1, Z_2, \ldots$ be a sequence of vectors in H such that for every integer m and n

$$E^*[Z_n|M_m] = Z_m \quad \text{if} \quad m \leq n \ . \tag{6.19}$$

Then there is a unique vector X in M_∞ such that

$$Z_n = E^*[X|M_n] \quad \text{for every} \quad n \tag{6.20}$$

$$\lim_{n \to \infty} \|Z_n - X\| = 0 \tag{6.21}$$

if and only if

$$\lim_n \|Z_n\|^2 < \infty \ . \tag{6.22}$$

If Z is a vector in H such that

$$Z_n = E^*[Z|M_n] \ , \quad n = 1,2,\ldots \tag{6.23}$$

then

$$X = E^*[Z|M_\infty] \ . \tag{6.24}$$

Proof: We consider only the case where the spaces M_n are non-decreasing. Define $Y_1 = Z_1$, and, for $n = 2,3,\ldots$, $Y_n = Z_n - Z_{n-1}$. From (6.19), it follows that, for every $n = 1,2,\ldots, Y_n$ belongs to M_n and is orthogonal to M_{n-1} . Consequently, $\{Y_n, n = 1,2,\ldots\}$ is a sequence of orthogonal vectors such that

$$\sum_{n=1}^{\infty} \|Y_n\|^2 = \lim_{m \to \infty} \sum_{n=1}^{m} \|Y_n\|^2 = \lim_{m \to \infty} \| \sum_{n=1}^{m} Y_n\|^2 = \lim_{m \to \infty} \|Z_m\|^2 < \infty .$$

Consequently, there exists a vector, denoted by X , say, such that

$$X = \sum_{n=1}^{\infty} Y_n = \lim_{m \to \infty} \sum_{n=1}^{m} Y_n = \lim_{n \to \infty} Z_n .$$

We leave it to the reader to complete the proof of Theorem 6D .

We next state a useful theorem on the convergence of a sequence of reproducing kernels.

Theorem 6E: Let T be either a countable set or a separable metric space. Let K be a symmetric non-negative kernel defined on $T \times T$, which is weakly continuous on T if T is metric. Let, for $n = 1,2,\ldots,$

$$T_n = \{t_1, t_2, \ldots, t_{N(n)}\} \tag{6.25}$$

be a sequence of subsets of T which are monotone increasing (that is, $T_n \subset T_{n+1}$ for any integer n), and such that the union $T' = \bigcup_{n=1}^{\infty} T_n$ is (i) equal to T, if T is a countable set, and (ii) dense in T if T is a separable metric space. Let C be the class of functions f, defined on T, such that, for any $n = 1,2,\ldots$, there exists a solution $\{c_1(f),\ldots,c_{N(n)}(f)\}$ of the set of linear equations

$$(6.26) \qquad f(t_j) = \sum_{k=1}^{N(n)} c_k(f)K(t_j,t_k), \; j = 1,2,\ldots,N(n) \; .$$

For any functions f and g in C, define

$$(6.27) \qquad (f,g)_n = \sum_{k=1}^{N(n)} c_k(f)g(t_k) = \sum_{k=1}^{N(n)} c_k(g)f(t_k) \quad .$$

Then, for any function f in C, the sequence $(f,f)_n$ is monotone increasing; that is, for any integer n,

$$(6.28) \qquad (f,f)_n \leq (f,f)_{n+1} \quad .$$

Further (assuming that f is continuous in the case that T is metric),

$$(6.29) \qquad f \in C \text{ and } \lim_n (f,f)_n < \infty \text{ if, and only if, } f \in H(K) \; .$$

If f and g both belong to $H(K)$, then

$$(6.30) \qquad (f,g)_{H(K)} = \lim_n (f,g)_n \quad .$$

Proof: Let $H(K)$ be the reproducing kernel Hilbert space whose kernel is K . Let $M_n = L^*(K(\cdot,t), t \in T_n)$. Let f be a function on T . For $n = 1,2,\ldots$, define

$$F_n = \sum_{k=1}^{N(n)} c_k(f)K(\cdot,t_k)$$

which is a vector in $H(K)$ with norm

$$\|F_n\|_{H(K)} = (f,f)_n \ .$$

It is clear that the sequence $\{F_n, n = 1,2,\ldots\}$ is a martingale in the wide sense (that is, $E^*[F_n|M_m] = F_m$ for $m \leq n$) . Consequently, (6.28) holds. If $f \in H(K)$, then $F_n = E^*[f|M_n]$ so that $\lim_n (f,f)_n \leq \|f\|_{H(K)} < \infty$. On the other hand, if $\lim_n (f,f)_n < \infty$, then by Theorem 6D (and the fact that under the assumptions made in Theorem 6E, $M_\infty = H(K)$), there exists a unique vector F in $H(K)$ such that $F_n = E^*[F|M_n]$ for $n = 1,2,\ldots$. We show that $f \in H(K)$, by showing that for every t in T' (whence for every t in T) $f(t) = F(t)$. Now if $t \in T'$ then there is some n such that $t \in T_n$; therefore $F(t) = (F_n, K(\cdot,t)) = f(t)$.

The proof of Theorem 6E is now complete.

7. Generalized linear and integral equations.

A system of m linear equations in variables $x_1, x_2, \ldots, x_n$, such as

$$
\begin{aligned}
k_{11}\, x_1 + k_{12}\, x_2 + \ldots + k_{1n}\, x_n &= f_1 \\
k_{21}\, x_1 + k_{22}\, x_2 + \ldots + k_{2n}\, x_n &= f_2 \\
\ldots\ldots\ldots\ldots\ldots\ldots& \\
k_{m1}\, x_1 + k_{m2}\, x_2 + \ldots + k_{mn}\, x_n &= f_m
\end{aligned} \tag{7.1}
$$

where the rectangular matrix $\{k_{ij}\}$ and the quantities $f_1, f_2, \ldots, f_n$ are known, may be formulated as a Hilbert space problem in the following way. Define the row vectors

$$
\begin{aligned}
x &= (x_1, x_2, \ldots, x_n , \\
k(t) &= (k_{t1}, k_{t2}, \ldots, k_{tn}), \; t = 1,2,\ldots,m \quad .
\end{aligned} \tag{7.2}
$$

The inner product between two such vectors $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$ is defined as usual by

$$
(a,b) = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n \quad . \tag{7.3}
$$

Then the linear equations (7.1) may be reformulated as follows: given the family of vectors $\{k(t), t = 1,2,\ldots,m\}$ and the function f defined on $T = \{1,2,\ldots,m\}$, find the vector x such that

$$(x, k(t)) = f(t) \quad \text{for } t \text{ in } T. \tag{7.4}$$

Now it is well known that (7.4) may have no solutions, one solution, or an infinity of solutions. Let us now show how one may use the concepts of reproducing kernel Hilbert spaces to state conditions under which a solution to (7.4) exists and to give a formula for the solution.

We first restrict the notion of solution so as to render the solution of (7.4) *unique, if it exists.* A vector x will be called a solution of (7.4) if it belongs to the linear manifold $L(k(t), t \in T)$ spanned by the family $\{k(t), t \in T\}$. In other words, we are seeking the vector x which among all vectors satisfying (7.4) has minimum norm.

Next, substituting a trial vector

$$x = \sum_{s=1}^{m} c_s k(s) \tag{7.5}$$

into (7.5), we obtain that the m coefficients c_s must satisfy the m linear equations

(7.6) $$\sum_{s=1}^{m} c_s \, (k(s),k(t)) = f(t), t \in T \quad .$$

Let us define K to be the covarience kernel of the family of vectors $\{k(t), t \in T\}$; for s and t in T ,

(7.7) $$K(s,t) = (k(s),k(t)) \quad .$$

Then (7.6) may be written

(7.8) $$\sum_{s=1}^{m} c_s K(s,t) = f(t), t \in T \quad .$$

Now the kernel K is a symmetric non-negative kernel. Consequently let $H(K)$ be its reproducing kernel space. In terms of $H(K)$ we may state the following theorem.

Theorem 7A: A necessary and sufficient condition that the system of equations (7.4) possess a (necessarily unique) solution x in $L(k(t), t \in T)$ is that $f \in H(K)$. Then the solution x may be given explicitly as

(7.9) $$x = \psi(f)$$

where ψ is the congruence from $H(K)$ onto $L(k(t), t \in T)$ such that

(7.10) $$\psi(K(\cdot,t)) = k(t)$$

In particular if $\{k(t), t \in T\}$ is a linearly independent family of vectors, so that the matrix $\{K(s,t), s,t \in T\}$ is non-singular and possesses an inverse $\{K^{-1}(s,t), s,t \in T)$, then the solution x of (7.4) is given by

$$(7.11) \qquad x = \sum_{s,t\in T} f(t)K^{-1}(s,t)k(s)$$

since the solution c_s of (7.8) is given by

$$(7.12) \qquad c_s = \sum_{t\in T} f(t)K^{-1}(s,t) \quad .$$

The norm of x is given by

$$(7.13) \qquad \|x\| = \|f\|_{H(K)} = \sum_{s,t\in T} f(t)K^{-1}(s,t)f(s)$$

Proof: The assertion that a solution x exists if, and only if, $f \in H(K)$ is an equivalent way of asserting the fact that (7.4) possesses a solution x given by (7.5) if and only if the equations (7.8) have a solution. Applying the congruence ψ to both sides of (7.10) we see that the solution x given by (7.5) may be written in the form (7.10) .

Theorem 7A represents a way of defining the notion of pseudo - inverse A^{-1} of a rectangular $m \times n$ metrix $A = \{k_{ij}\}$. If we define the operator A^{-1} to be ψ, then the system of linear equations $Ax = f$ has a formal solution $f = A^{-1}x$. In particular, if $K^{-1}(s,t)$ exists, then equation (7.11) may be written in the form

$$(7.14) \qquad x_j = \sum_{s,t=1}^{m} f_t K^{-1}(t,s)k_{sj}$$

so that we may define A^{-1} to be the $n \times m$ matrix with elements given by, for $j = 1,\ldots,n$ and $t = 1,\ldots,m$

$$(7.15) \qquad (A^{-1})_{jt} = \sum_{s=1}^{m} K^{-1}(t,s)k_{sj} .$$

In matrix notation we may write

$$(7.16) \qquad A^{-1} = ((AA^T)^{-1}A)^T$$

where A^T is the transpose of A . Note that if A is a square matrix, then (7.16) reduces to the usual notion of inverse. To illustrate the use of these results let us consider the following simple example.

Example: Let us find the vector $x = (x_1,x_2,x_3)$ of minimum norm satisfying the system of linear equations

$$(7.17) \qquad \begin{aligned} x_1 + x_2 + x_3 + x_4 &= y_1 \\ x_1 + 2x_2 + 3x_3 + 4x_4 &= y_2 \\ x_1 - x_2 &= y_3 \end{aligned} .$$

Let $k(1) = (1,1,1,1)$, $k(2) = (1,2,3,4)$, $k(3) = (1,-1,0,0)$. Then the matrix $\{K(s,t)\}$, defined by (7.7), is given by

$$\{K(s,t)\} = \begin{pmatrix} 4 & 10 & 0 \\ 10 & 30 & -1 \\ 0 & -1 & 2 \end{pmatrix} .$$

Its inverse is given by

$$\{K^{-1}(s,t)\} = \frac{1}{36} \begin{pmatrix} 59 & -20 & -10 \\ -20 & 8 & 4 \\ -10 & 4 & 20 \end{pmatrix} .$$

If we define

$$A = \{k_{ij}\} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & -1 & 0 & 0 \end{pmatrix}$$

then by (7.15) the transpose of its pseudo - inverse is given by

$$(A^{-1})^T = \frac{1}{36} \begin{pmatrix} 29 & 29 & -1 & -21 \\ -8 & -8 & 4 & 12 \\ 14 & -22 & 2 & 6 \end{pmatrix} .$$

The vector x of minimum norm satisfying the linear equations (7.17) is now given by $x = A^{-1}y$.

We next state a theorem which extends Theorem 7A to the case where the vectors x and $k(t)$ in (7.4) are vectors in an arbitrary Hilbert space H .

<u>Theorem 7B</u>: Let H be a Hilbert space, let $\{k(t), t \in T\}$ be a family of vectors in H, and let f be a function on the index set T . A necessary and sufficient condition that there exist a (necessarily unique) vector x in the Hilbert subspace $L^*(k(t), t \in T)$ satisfying the equations

$$(7.18) \qquad (x, k(t)) = f(t), t \in T$$

is that f belong to $H(K)$, the reproducing kernel Hilbert space corresponding to the covariance kernel K defined for s and t in T by

$$(7.19) \qquad K(s,t) = (k(s), k(t)) \quad .$$

The solution to (7.18) is then given by

$$(7.20) \qquad x = \psi(f)$$

where ψ is the congruence from $H(K)$ onto $L^*(k(t), t \in T)$ such that

$$\psi(K(\cdot,t)) = k(t) \quad . \tag{7.21}$$

The norm of x is given by

$$\|x\| = \|f\|_{H(K)} \quad . \tag{7.22}$$

<u>Proof</u>: Let ψ be the congruence from $H(K)$ onto $L^*(k(t), t \in T)$ satisfying (7.20). The existence of ψ is guaranteed by the Basic Congruence Theorem. If $f \in H(K)$, then a solution to (7.18) exists, since $x = \psi(f)$ satisfies (7.18). If a solution x to (7.18) exists, let f' be the function in $H(K)$ such that $x = \psi(f')$. Then

$$f'(t) = (f', K(\cdot,t)) = (x, k(t)) = f(t)$$

so that f' and f coincide, and f belongs to $H(K)$. The proof of Theorem 7B is now complete.

One may consider (7.18) a generalized integral equation, since in particular it includes the case where we seek the solution (in the Hilbert space $L_2(a,b)$ consisting of all functions square integrable with respect to Lebesgue measure over the interval a to b) of the integral equation of the first kind

$$\int_a^b x(s)k(s,t)dt = f(t), \qquad a \leq t \leq b \quad .$$

The problem of finding the projection $E^*[u|k(t), t \in T]$ of a vector u in a Hilbert space H onto the subspace $L^*(k(t), t \in T)$ is also of the form of (7.18), with $f(t) = (u,k(t))$. In particular, let us apply Theorem 7B to obtain an explicit formula for the projection $E^*[u|v_1,\ldots,v_n]$ of a vector in an arbitrary Hilbert space H onto the subspace $L(v_1,\ldots,v_n)$ spanned by n linearly independent vectors $v_1,\ldots,v_n$. The projection is the unique vector x satisfying

$$(x,v_t) = (u,v_t) \qquad t = 1,\ldots,n \quad . \tag{7.23}$$

We thus see that

$$E^*[u|v_1,\ldots,v_n] = \psi(f) \tag{7.24}$$

where f is the function on $T = \{1,2,\ldots,n\}$ defined by $f(t) = (u,v_t)$ and ψ is the congruence satisfying (7.21), where

$$K(s,t) = (v_s,v_t) \quad . \tag{7.25}$$

Explicitly we have

$$E^*[u|v_1,\ldots,v_n] = \sum_{s,t=1} (u,v_s)K^{-1}(s,t)v_t \tag{7.26}$$

$$(7.27)\quad d^2[u|v_1,\ldots,v_n]=\|u-E^*[u|v_1,\ldots,v_n]\|^2=\|u\|^2-\|E^*[u|v_1,\ldots,v_n]\|^2$$

$$(7.28)\qquad \|E^*[u|v_1,\ldots,v_n]\|^2 = \sum_{s,t=1}^{n} (u,v_s)K^{-1}(s,t)(u,v_t) \quad .$$

Equation (7.26) is well known (see, for example, Doob(1953),p.151).

Equation (7.28) has an interpretation from which one may obtain inequalities which generalize the Cramér-Rao inequalities in the theory of statistical estimation.

Lemma 7a: Let $v_1,\ldots,v_n$ be linearly independent vectors in a Hilbert space H . Let K be defined by (7.25) . Then for any vector u in H

$$(7.29)\qquad \|u\|^2 \geq \sum_{s,t=1}^{n} (u,v_s)K^{-1}(s,t)(u,v_t) \quad .$$

Equality holds in (7.29) if and only if, u is a linear combination of $v_1,\ldots v_n$.

Let us show how the multi-parameter Cramér-Rao inequality follows immediately from Lemma 7a . Let $\theta = (\theta_1,\ldots,\theta_n)$ be a parameter whose domain Λ is a subset of Euclidean n space. Let $(P_\theta, \theta \in \Lambda)$ be a family of probability measures on some probability space $(\Omega,\mathcal{A})$ which are dominated by a measure μ , so that for each θ in Λ the Radon-Nikodym derivative

$$f(\theta) = f(\theta_1,\ldots,\theta_n) = \frac{dP_\theta}{d\mu}$$

exists. Assume that, for each Θ in Λ, $\int_\Omega f^2(\Theta)d\mu < \infty$. Assuming all the derivatives below to be defined as limits in quadratic mean, the random variables

$$V_j(\Theta) = \frac{\partial}{\partial\Theta_j} \log f(\Theta) = \frac{1}{f(\Theta)} \frac{\partial}{\partial\Theta_j} f(\Theta_1,\ldots,\Theta_n)$$

have, for $j = 1,2,\ldots,n$, means

$$E_\Theta[V_j(\Theta)] = \int_\Omega V_j(\Theta)f(\Theta)d\mu = \int_\Omega \frac{\partial}{\partial\Theta_j} f(\Theta)d\mu = \frac{\partial}{\partial\Theta_j} \int_\Omega f(\Theta)d\mu = 0$$

and covariances

$$K_{ij}=E_\Theta[V_i(\Theta)V_j(\Theta)]=E_\Theta[\frac{\partial}{\partial\Theta_i} \log f(\Theta) \frac{\partial}{\partial\Theta_j} \log f(\Theta)]= -E_\Theta[\frac{\partial^2}{\partial\Theta_i\partial\Theta_j} \log f(\Theta)] .$$

For any random variable U and function g on Λ, letting

$$E_\Theta[\{U-g(\Theta)\}V_j] = E_\Theta[UV_j] = \frac{\partial}{\partial\Theta_j} E_\Theta[U] \equiv b_j$$

it follows by Lemma 7a that

$$E_\Theta|U-g(\Theta)|^2 \geq \sum_{i,j=1}^{n} b_i K_{ij}^{-1} b_j$$

which is the Cramér-Rao inequality.

In order for Theorem 7B to be a useful tool in solving generalized integral equations of the form of (7.18), we mast possess criteria for showing that a given function f belongs to $H(K)$, and for calculating $\psi(f)$. We now give examples of several such conditions. <u>They are all based on the idea that if a function f can be represented in terms of linear operations on the family $\{K(\cdot,t), t \in T\}$, including the operations of addition, differentiation, and integration, then f belongs to $H(K)$ and $\psi(f)$ may be expressed in terms of the family $\{k(t), t \in T\}$ by means of the same linear operations as are used to represent f in terms of $\{K(\cdot,t), t \in T\}$</u>.

We leave it to the reader to verify the following theorem. In particular, the reader should note how, by equations (7.36),(7.40), and (7.48), the calculation of $\psi(f)$ is reduced in certain circumstances to the solution of generalized Wiener-Hopf integral equations.

<u>Theorem 7C</u>: Let H be a Hilbert space, $\{k(t), t \in T\}$ be a family of vectors in H with covariance kernel K defined by (7.19), and let T be an interval on the real line. Let f be on a function on T.

(i) If there exists a finite subset T_n of T and constants $\{c(s), s \in T_n\}$ such that

$$f = \sum_{s \in T_n} c(s) K(\cdot, s) \tag{7.28}$$

then $f \in H(K)$ and

$$(7.29) \qquad \psi(f) = \sum_{s \in T_n} c(s)k(s)$$

$$(7.30) \qquad \|f\|^2 = \sum_{s,t \in T_n} c(s)c(t)K(s,t) = \sum_{t \in T_n} c(t)f(t)$$

(ii) If there exists a countable subset T_∞ of T and constants $\{c(s), s \in T_\infty\}$ such that

$$(7.31) \qquad \sum_{s,t \in T_\infty} c(s)c(t)K(s,t) \qquad \text{is finite}$$

and if the function f satisfies, for every t in T

$$(7.32) \qquad f(t) = \sum_{s \in T_\infty} c(s)K(t,s) ,$$

where the convergence of the series in (7.32) is implied by the convergence of the series in (7.31), then $f \in H(K)$ and

$$(7.33) \qquad \psi(f) = \sum_{s \in T_\infty} c(s)k(t)$$

where the series in (7.33) converges as a strong limit in H. Further

$$(7.34) \qquad \|f\|^2 = \sum_{s,t \in T_\infty} c(s)c(t)K(s,t) = \sum_{t \in T_\infty} c(t)f(t)$$

(iii) If there exists a continuous function g on T such that as a Riemann integral

$$\int_T \int_T g(s)g(t)K(s,t)ds\,dt \qquad \text{is finite} \tag{7.35}$$

and if the function f satisfies, for every t in T,

$$f(t) = \int_T g(s)K(t,s)ds \tag{7.36}$$

where the finiteness of the integral in (7.36) is implied by the finiteness of the integral in (7.35), then $f \in H(K)$ and

$$\psi(f) = \int_T g(s)k(s)ds \tag{7.37}$$

where the integral in (7.37) is a strongly convergent stochastic Riemann integral. Further

$$\|f\|^2 = \int_T \int_T g(s)g(t)K(s,t)ds\,dt = \int_T g(t)f(t)dt \quad . \tag{7.38}$$

(iv) If there exists a function of bounded variation V on T such that as a Riemann-Stieltjes integral

$$\int_T \int_T K(s,t)dV(s)dV(t) \tag{7.39}$$

is finite and if the function f satisfies, for every t in T,

$$f(t) = \int_T K(t,s)dV(s) \tag{7.40}$$

where the finiteness of the integral in (7.40) is implied by the finiteness of the integral in (7.39), then $f \in H(K)$ and

$$\psi(f) = \int_T k(s)dV(s) \tag{7.41}$$

where the integral in (7.41) is a strongly convergent stochastic Riemann-Stieltjes integral. Further

$$\|f\|^2 = \int_T \int_T K(s,t)dV(s)dV(t) = \int_T f(t)dV(t) \tag{7.42}$$

(v) If m is an integer such that the mixed symmetric partial derivatives

$$\frac{\partial^{2j}}{\partial s^j \partial t^j} K(s,t), \tag{7.43}$$

are finite for $j = 1,2,\ldots,m$ and if, for some t_o in T, the function f satisfies, for every t in T

$$f(t) = \sum_{j=0}^{m} a_j \frac{\partial^j}{\partial t^j} K(t,t_o) \tag{7.44}$$

then $f \in H(K)$ and

$$\psi(f) = \sum_{j=0}^{m} a_j \frac{\partial^j}{\partial t_j} k(t_o) \tag{7.45}$$

where the stochastic derivatives in (7.45) exist as strong limits in H. Further

$$(7.46) \qquad \|f\|^2 = \sum_{i,j=0}^{m} a_i a_j \frac{\partial^{i+j}}{\partial s^i \partial t^j} K(t_o, t_o)$$

(vi) If m is an integer such that (7.43) holds, and if $V_o, \ldots, V_m$ are functions of bounded variation on T such that

$$(7.47) \qquad \sum_{i,j=1}^{m} \int_T \int_T \frac{\partial^{i=j}}{\partial s^i \partial t^j} K(s,t) dV_i(s) dV_j(t)$$

is finite and if, for every t in T,

$$(7.48) \qquad f(t) = \sum_{j=0}^{m} \int_T \frac{\partial}{\partial s^j} K(t,s) dV_j(s)$$

then $f \in H(K)$ and

$$(7.49) \qquad \psi(f) = \sum_{j=0}^{m} \int_T X^{(j)}(s) dV_j(s) \quad .$$

Further, $\|f\|^2$ is equal to the double integral in (7.47) .

To illustrate the importance of Theorem 7C, let us apply it the problem of best linear prediction. Let $\{X(t), t \in T\}$ be a random function with covarience kernel R. Let Z be a random variable for which

$$\rho_Z(t) = E[ZX(t)], t \in T \tag{7.50}$$

is known. The best predictor $E^*[Z|X(t), t \in T]$ is the unique solution U in $L_2(X(t), t \in T)$ of the equation

$$E[UX(t)] = \rho_Z(t), t \in T \quad . \tag{7.51}$$

By Theorem 7C, if one can express ρ_Z in terms of linear operations on the family of functions $\{R(\cdot, t), t \in T\}$, then $E^*[Z|X(t), t \in T]$ may be expressed in terms of corresponding linear operations on $\{X(t), t \in T\}$. For example, let T be the interval from 0 to T; if there exists a solution $g(s)$ to

$$\rho_Z(t) = \int_0^T g(s)R(s,t)ds, \qquad 0 \leq t \leq T \tag{7.52}$$

such that the double integral corresponding to (7.35) exists and is finite, then

$$E^*[Z|X(t), 0 \leq t \leq T] = \int_0^T g(s)X(s)ds \quad . \tag{7.53}$$

On the other hand, if there exist solutions $g_0, \ldots, g_r$ to

$$\rho_Z(t) = \sum_{j=0}^{r} \int_0^T g_j(s) \frac{\partial}{\partial t^j} R(s,t)ds \ , \tag{7.54}$$

and a condition analogous to (7.47) holds, then

$$E^*[Z|X(t),\ 0 \leq t \leq T] = \sum_{j=0}^{r} \int_0^T g_j(s) \frac{\partial}{\partial s^j} X(s)ds \quad . \tag{7.55}$$

8. Minimum variance unbiased estimation.

Let X be a random element taking values in some space S. Assume that the probability law P of X is known only to be a member of a set $\mathcal{P}$ of probability measures over S. We assume that the set $\mathcal{P}$ has been indexed (or parametrized) by a parameter θ varying in a set Λ, so that we write $\mathcal{P} = \{P_\theta, \theta \in \Lambda\}$. We assume that $\mathcal{P}$ is dominated by a probability measure μ, by which we mean that each measure P_θ in $\mathcal{P}$ possesses a Radon-Nikodym derivative

$$k(\theta) = \frac{dP_\theta}{d\mu} \quad . \tag{8.1}$$

In the applications we will usually take for μ a probability measure P_{θ_0} belonging to $\mathcal{P}$. The Radon-Nikodym derivative $k(\theta)$ is non-negative and integrable with respect to μ. We assume further that, for each θ in Λ, $k(\theta)$ is square integrable; in symbols,

$$k(\theta) \in L_2(\mu), \theta \in \Lambda \tag{8.2}$$

where $L_2(\mu)$ is the Hilbert space of random variables U on S such that

$$(U,U) = \int_S U^2 \, d\mu < \infty \quad . \tag{8.3}$$

We are interested in estimating various functions f defined on the parameter space Λ. A function f, defined on Λ, is said to be <u>estimable</u> if there exists a random variable U in $L_2(\mu)$ such that

$$E_\theta[U] = (U,k(\theta)) = f(\theta), \theta \in \Lambda. \tag{8.4}$$

The subscript θ on E indicates that the expectation is to be taken under the assumption that P_θ is the true probability measure on S. A random variable U in $L_2(\mu)$ satisfying (8.4) is said to be an <u>unbiased</u> estimate of f.

A random variable U is said to be a μ-<u>efficient unbiased</u> estimate of f if it has minimum norm among all unbiased estimates. If μ is the probability measure P_{θ_o}, then the unbiased estimate U with least variance

$$E_{\theta_o}|U-f(\theta_o)|^2 = E_{\theta_o}[U^2]-f^2(\theta_o) \; ,$$

is the μ-efficient estimate. The estimate U is then called the locally best at θ_o unbiased estimate of f. An unbiased estimate is said to be a <u>uniformly minimum variance</u> estimate of f if it is locally best at all parameter points in Λ.

Using the notions discussed in preceding sections we are immediately able to characterize estimable functions f and μ-efficient unbiased estimates U. Define the covariance kernel

$$K(\theta_1,\theta_2) = (k(\theta_1),k(\theta_2)) , \tag{8.5}$$

and let $H(K)$ be the corresponding reproducing kernel Hilbert space. The following facts are all of some interest.

<u>Theorem 8A:</u> (1) A function f on Λ is an estimable function if, and only if, $f \in H(K)$.

(2) The μ-efficient unbiased estimate U of an estimable function f is $\psi(f)$, where ψ is the congruence from $H(K)$ onto $L_2(k(\theta),\theta \in \Lambda)$ satisfying

$$\psi(K(\cdot,\theta)) = k(\theta) \tag{8.6}$$

(3) any random variable U in $L_2(k(\theta),\theta \in \Lambda)$ is the μ-efficient unbiased estimate of the function f in $H(K)$ given by (8.4) .

(4) If V in $L_2(\mu)$ is an unbiased estimate of f, then the projection $E^*[V|k(\theta),\theta \in \Lambda]$ of V into $L_2(k(\theta),\theta \in \Lambda)$ is the u-efficient unbiased estimate of f .

(5) The norm $\|U\|$ of the μ-efficient unbiased estimate U of f is equal to $\|f\|_{H(K)}$. A multitude of lower bounds may be given for $\|U\|$; in particular, for any n linearly independent random variables $V_1, V_2, \ldots, V_n$ in $L_2(k(\theta), \theta \in \Lambda)$

$$\|U\| \geq \sum_{i,j=1}^{n} (U, V_i) v_{ij}^{-1} (U, V_j) \tag{8.7}$$

where

$$\{v_{ij}^{-1}\} = \{v_{ij}\}^{-1}, v_{ij} = (V_i, V_j) \quad .$$

It was pointed out in section 7 how (8.7) includes all the usual "information" inequalities such as the Cramér-Rao and Bhattacharya inequalities (for a statement of these inequalities, see C.R. Rao, 1952), p.143).

We are mainly interested in using Theorem 8A as a tool to find μ-efficient estimates. Before giving examples of how one uses Theorem 8A, let us note a condition for the existence of uniformly minimum variance estimates in the important special case that μ is the probability measure P_{θ_o} , where $\theta_o \in \Lambda$. Then $k(\theta_o) = 1$ belongs to $L_2(k(\theta), \theta \in \Lambda)$. Consequently

$$\psi(1) = 1 \quad . \tag{8.8}$$

Because of the non-negativity of the $k(\theta)$, it follows that for any random variable V in $L_2(\mu)$, $V \geq 0$ implies $E^*[V|k(\theta),\theta \in \Lambda] \geq 0$. Therefore, by Lemma 6a,

$$E^*_{\theta_o}[V|k(\theta),\theta \in \Lambda] = E_{\theta_o}[V|k(\theta),\theta \in \Lambda] \quad . \tag{8.9}$$

Thus the projection of V onto the space spanned by $\{k(\theta),\theta \in \Lambda\}$ coincides with the conditional expectation of V given the sigma field generated by $\{k(\theta),\theta \in \Lambda\}$. In view of these facts we may state the following theorem.

<u>Theorem 8B</u>: Let $\{P_\theta,\theta \in \Lambda\}$ be a family of probability measures such that for θ and θ_o in Λ the Radon-Nikodym derivative

$$k_{\theta_o}(\theta) = \frac{dP_\theta}{dP_{\theta_o}}$$

exists, and is square integrable

$$E_{\theta_o}[k^2_{\theta_o}(\theta)] < \infty \quad .$$

Let K_{θ_o} be the kernel on $\Lambda \times \Lambda$ defined by

$$K_{\theta_o}(\theta_1,\theta_2) = E_{\theta_o}[k_{\theta_o}(\theta_1)k_{\theta_o}(\theta_2)] \quad .$$

Then an estimable function f on Λ, belonging to $H(K_{\theta_o})$ for each θ_o in Λ, possesses a uniformly minimum variance estimate if for every V in $L_2(X(t), t \in T)$ there exists, for each θ_o in Λ, a determination of $E_{\theta_o}[V|k_{\theta_o}(\theta) \in \Lambda]$ which is functionally independent of θ_o .

<u>Example 8A</u>: Let us first consider the simple case where the random element X is a random variable assumed to be Normally distributed with unknown mean θ , and variance 1. Assume that $-\infty < \theta < \infty$. If μ is the measure corresponding to mean θ_o, then

$$k(x,\theta) = e^{(\theta - \theta_o)x} e^{-\frac{1}{2}(\theta^2 - \theta_o^2)} .$$

Using the known value of the moment generating function of a Normally distributed random variable, we obtain

$$K(\theta_1,\theta_2) = e^{-\frac{1}{2}(\theta_1^2 + \theta_2^2 - 2\theta_o^2)} E_{\theta_o}[e^{X(\theta_1 + \theta_2 - 2\theta_o)}]$$

$$= \exp\{-\tfrac{1}{2}(\theta_1^2 + \theta_2^2 - 2\theta_o^2) + \theta_o(\theta_1 + \theta_2 - 2\theta_o) + \tfrac{1}{2}(\theta_1 + \theta_2 - 2\theta_o)^2\}$$

$$= \exp -\tfrac{1}{2}\{(\theta_1 - \theta_o)^2 + (\theta_2 - \theta_o)^2 - (\theta_1 + \theta_2 - 2\theta_o)^2\} .$$

We are interested in estimating the identity function $f(\theta) = \theta$. To show that f is estimable, and to obtain an explicit formula for the θ_o -efficient unbiased estimate of f, it suffices to represent f by means of linear operations on the kernel K. Now

$$(8.10) \qquad \frac{\partial}{\partial\theta_1} K(\theta_1,\theta) = K(\theta_1,\theta)\{(\theta_1 + \theta - 2\theta_o)-(\theta_1 - \theta_o)\}$$

therefore, setting $\theta_1 = \theta_o$

$$(8.11) \qquad \theta = \frac{\partial}{\partial\theta_1} K(\theta_1,\theta)\Big|_{\theta_1 = \theta_o} + \theta_o \quad .$$

From (8.11) it follows that the identity function $f(\theta) = \theta$ belongs to $H(K)$, and is therefore estimable and the θ_o -efficient unbiased estimate of θ is

$$(8.12) \qquad \frac{\partial}{\partial\theta} k(\theta_o) + \theta_o = (x - \theta_o) + \theta_o = x \quad .$$

We thus obtain the well known result that x is, for every θ_o, a θ_o -efficient unbiased estimate of θ ; in other words, x is a uniformly minimum variance estimate of the mean θ of a Normally distributed random variable with unknown mean and known variance.

<u>Example 8B</u>: We next consider the case of a random variable X whose probability law P_θ is of Koopman - Darmois exponential type, by which is meant that, with respect to some measure μ , P_θ possesses a probability density

$$p(x,\theta) = \beta(\theta)e^{x\theta}, \beta(\theta) = \{\int e^{x\theta} d\mu\}^{-1} \quad . \tag{8.13}$$

Define

$$k(x,\theta) = \frac{p(x,\theta)}{p(x,\theta_o)} = \frac{\beta(\theta)}{\beta(\theta_o)} e^{x(\theta - \theta_o)}$$

$$K(\theta_1,\theta_2) = E_{\theta_o}[k(\theta_1)k(\theta_2)] = \frac{\beta(\theta_1)\beta(\theta_2)}{\beta(\theta_o)\beta(\theta_1+\theta_2-\theta_o)} \quad . \tag{8.14}$$

Forming the derivative

$$\frac{\partial}{\partial\theta_1} K(\theta_1,\theta) = \frac{\beta(\theta_1+\theta-\theta_o)\beta'(\theta_1)\beta(\theta)-\beta(\theta_1)\beta(\theta)\beta'(\theta_1+\theta-\theta_o)}{\beta(\theta_o)\beta^2(\theta_1+\theta-\theta_o)}$$

and evaluating it at $\theta_1 = \theta_o$, we obtain

$$\frac{\partial}{\partial\theta_1} K(\theta_o,\theta) = \frac{\beta'(\theta_o)}{\beta(\theta_o)} - \frac{\beta'(\theta)}{\beta(\theta)} \quad . \tag{8.15}$$

Define now as the function we desire to estimate

$$f(\theta) = -\frac{\beta'(\theta)}{\beta(\theta)} = E_\theta[x] = \beta(\theta)\int x e^{\theta x} d\mu \quad . \tag{8.16}$$

From (8.15) it follows that $f(\theta)$ is estimable, and its θ_o - efficient unbiased estimate is given by

$$\frac{\partial}{\partial\theta} k(\theta_o) + \frac{\beta'(\theta_o)}{\beta(\theta_o)} = x \quad . \tag{8.17}$$

We thus obtain the well known result that x is, for every θ_o, a θ_o -efficient unbiased estimate of $f(\theta)$.

Example 8C: We next apply Theorem 8A to obtain results which are important for our treatment of statistical inference on Normal random functions. Let us consider the case where the observed random element X is a finite collection of random variables $X_1, X_2, \ldots, X_n$ which are assumed to be jointly Normally distributed with means and covariances

$$(8.18) \qquad m(t) = E[X_t], \; R(s,t) = \mathrm{Cov}[X_s, X_t] \quad .$$

We assume the covariances to be known but the means to be unknown. We may then parametrize the set $\mathcal{P}$ of possible probability measures for $X = (X_1, \ldots, X_n)$ by the mean m . Let P_m be the measure corresponding to the assumption that X is normal with mean m and covariance kernel R .

Let us assume for the present that the inverse matrix $\{R^{-1}(s,t)\}$ exists. Let us define, for $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$,

$$(8.19) \qquad (a,b)_n = \sum_{s,t=1}^{n} a_s \, R^{-1}(s,t) b_t, \|a\|_n = (a,a)_n \quad .$$

One may then prove that the Radon-Nikodym derivative of P_m with respect to P_{m_o} (for any fixed mean m_o) is given by

$$(8.20) \qquad k(m) = \frac{dP_m}{dP_{m_o}} = e^{(X-m_o,m-m_o)-\frac{1}{2}\|m-m_o\|_n^2} \quad .$$

Equation (8.20) follows from the well known formula for the multivariate normal probability density function. Next using the fact that $(X-m_1,m_1+m_2-2m_o)_n$ is a normally distributed random variable with mean 0 and variance $\|m_1+m_2-2m_o\|_n^2$, we obtain that

$$K(m_1,m_2) = (k(m_1),k(m_2)) = E_{m_o}[k(m_1)k(m_2)]$$

$$(8.21) \qquad = \exp - \tfrac{1}{2}\{\|m_1-m_o\|_n^2+\|m_2-m_o\|_n^2-\|m_1+m_2-2m_o\|_n^2\} \quad .$$

$$= e^{(m_1-m_o,m_2-m_o)} \quad .$$

Now assume that the mean value function $m(t)$ is assumed to be an unknown linear combination of known functions $\varphi_1(t),\ldots,\varphi_q(t)$ so that

$$(8.22) \qquad m(t) = \beta_1\varphi_1(t) + \beta_2\varphi_2(t) + \ldots + \beta_q\varphi_q(t)$$

where the parameters $\beta_1,\beta_2,\ldots,\beta_q$ are unknown and are to be estimated. More generally, let us estimate the linear function

$$(8.23) \qquad f(\beta) = c_1\beta_1 + c_2\beta_2 + \ldots + c_q\beta_q$$

where the constants $c_1, c_2, \ldots, c_q$ are pre-assigned

Let $m_i(t) = \beta_1^{(i)}\varphi_1(t) + \ldots + \beta_q^{(i)}\varphi_q(t)_1$ and $\beta^{(i)} = (\beta_1^{(i)}, \ldots, \beta_q^{(i)})$. We shall find the $\beta^{(o)}$-efficient unbiased estimate of $f(\beta)$ by first expressing $f(\beta)$ as a linear combination of derivatines of

$$K(\beta^{(1)}, \beta^{(2)}) = e^{-(\beta^{(1)}-\beta^{(o)})M(\beta^{(2)}-\beta^{(o)})'} \tag{8.24}$$

where the matrix $M = \{M_{ij}\}$ is defined by

$$M_{ij} = (\varphi_i, \varphi_j)_n \qquad \text{for } i,j = 1,\ldots,q \quad . \tag{8.25}$$

Differentiating $K(\beta^{(1)}, \beta^{(2)})$, we obtain

$$\frac{\partial}{\partial \beta_\alpha^{(1)}} K(\beta^{(1)}, \beta) = K(\beta^{(1)}, \beta)\left\{\sum_{j=1} M_{\alpha j}(\beta_j - \beta_j^{(o)})\right\} .$$

Consequently,

$$\begin{pmatrix} \frac{\partial}{\partial \beta_1} & K(\beta^{(o)}, \beta) \\ \vdots & \\ \frac{\partial}{\partial \beta_q} & K(\beta^{(o)}, \beta) \end{pmatrix} = M(\beta - \beta^{(o)}) \quad . \tag{8.26}$$

Assuming for the present that the functions $\varphi_1(t), \ldots, \varphi_q(t)$ are linearly independent in the sense that the inverse matrix

$M^{-1} = \{M_{ij}^{-1}\}$ exists, we may express $\beta_1, \dots, \beta_q$ in terms of derivatives of K by

$$(8.27) \quad \beta = \beta^{(o)} + M^{-1} \begin{pmatrix} \frac{\partial}{\partial \beta_1} & K(\beta^o, \beta) \\ \vdots & \\ \frac{\partial}{\partial \beta_q} & K(\beta^o, \beta) \end{pmatrix} .$$

Consequently, the $\beta^{(o)}$ -efficient unbiased estimate of β , which we denote by $\hat{\beta}$, is given by

$$(8.28) \quad \hat{\beta} = \beta^{(o)} + M^{-1} \begin{pmatrix} \frac{\partial}{\partial \beta_1} & k(\beta^{(o)}) \\ \vdots & \\ \frac{\partial}{\partial \beta_q} & k(\beta^{(o)}) \end{pmatrix} = M^{-1} \begin{pmatrix} (X, \varphi_1)_n \\ \vdots \\ (X, \varphi_q)_n \end{pmatrix}$$

since

$$(8.29) \qquad \frac{\partial}{\partial \beta_\alpha} k(\beta^{(o)}) = (X - m_o, \varphi_\alpha)_n = (X, \varphi_\alpha)_n - \sum_{j=1}^{q} \beta_j^{(o)} M_{j\alpha} .$$

Next, the $\beta^{(o)}$ -efficient estimate $\widehat{f(\beta)}$ of an estimable function $f(\beta) = c_1\beta_1 + \dots + c_q\beta_q$ is given by

$$(8.30) \qquad \widehat{f(\beta)} = \sum c_i \hat{\beta}_i = \sum_{i,j=1}^{q} c_i M_{ij}^{-1} (X, \varphi_j)_n .$$

We again find that the $\beta^{(o)}$ -efficient unbiased estimates are independent of $\beta^{(o)}$ and are uniformly minimum variance unbiased estimates. The variance of $f(\beta)$ may be calculated to be, for all β ,

$$\text{Var}[f(\hat{\beta})] = \sum_{i,j=1}^{q} c_i M_{ij}^{-1} c_j \tag{8.31}$$

where we have used the fact that

$$\text{Cov}[(X,\varphi_i)_n,(X,\varphi_j)_n] = (\varphi_i,\varphi_j)_n = M_{ij} \quad . \tag{8.32}$$

An unusual number of reproducing kernel Hilbert spaces are involved in the foregoing discussion, corresponding to the covariance kernels K, R, and M . In terms of these spaces we may formulate the foregoing results as follows: for any vector $c = (c_1,\ldots,c_q)$ in $H(M)$, the parametric function $c'\beta = c_1\beta_1 + \ldots + c_q\beta_q$ possesses a uniformly minimum variance estimate $\widehat{c'\beta}$ given by

$$\widehat{c'\beta} = (c_i,(X,\varphi)_{H(R)})_{H(M)} \tag{8.33}$$

where we define

$$(X,\varphi)_{H(R)} = ((X,\varphi_1)_{H(R)},\ldots,(X,\varphi_q)_{H(R)}) \quad . \tag{8.34}$$

It is illuminating to compare (8.28) with the usual way in which one writes the solution to the problem of estimating the parameter $\beta = (\beta_1,\ldots,\beta_q)$ in the case that one writes the mean $E[X]=(m_1,\ldots,m_n)$ of the random vector $X = (X_1,\ldots,X_n)$ in the form

$$(8.35) \qquad m = A\,\beta' , \qquad A = \begin{pmatrix} \varphi_1(1) & \varphi_2(1) & \ldots & \varphi_q(1) \\ \varphi_1(2) & \varphi_2(2) & \ldots & \varphi_q(2) \\ \vdots & & & \\ \varphi_1(n) & \varphi_2(n) & \ldots & \varphi_q(n) \end{pmatrix} .$$

Then the minimum variance unbiased of β may be written (assuming non-singularity of all the matrices involved)

$$(8.36) \qquad \hat{\beta} = (A^T R^{-1} A)^{-1} (A^T R^{-1} X)$$

where R is the covariance matrix of X, and the superscript T denotes transpose.

9. <u>The probability density functional of a Normal random function.</u>

Given a random function $\{X(t), t \in T\}$, by the <u>proper covariance kernel</u> R we mean the function defined for s,t in T by

$$(9.1) \qquad R(s,t) = \mathrm{Cov}[X(s),X(t)] \quad .$$

Let us now consider a normal random function $\{X(t), t \in T\}$ whose proper covariance kernel R is known, but whose mean value function m, defined for t in T by

$$(9.2) \qquad m(t) = E[X(t)]$$

is not assumed to be known but is only assumed to belong to a known class $\mathcal{M}$ of possible value functions. One importent example of a class $\mathcal{M}$ is the class of all linear combinations of a finite number of known functions $\varphi_1, \ldots, \varphi_q$, so that each m in $\mathcal{M}$ is of the form

$$(9.3) \qquad m(t) = \beta_1\varphi_1(t) + \ldots + \beta_q\varphi_q(t)$$

for some real constants $\beta_1, \ldots, \beta_q$. Another example of $\mathcal{M}$ arises when m is assumed to be of the form

$$(9.4) \qquad m(t) = g(t-\beta)$$

where $g(t)$ is a known function (say, $g(t) = \cos\ t$), and the time delay β is unknown.

In order to use the considerations of the foregoing section to form minimum variance unbiased estimates, we obtain in this section an explicit formula for the probability density functional of a normal random function. We consider a stochastic process $\{X(t), t \in T\}$

which is separable in the sense defined by Doob (1953, Chapter 2). We assume further that the probabilizable space on which the random variables are defined is $(\Omega, \mathscr{C})$, where Ω is the function space consisting of all real valued functions defined on T, and $\mathscr{C}$ is the sigma-field of cylinder sets in Ω, defined as the smallest σ-field $\mathscr{C}$ sets in Ω containing, for any t in T and real number a, the set $\{\omega:\omega(t) < a\}$.

Let P_1 and P_2 be probability measures on $(\Omega, \mathscr{C})$ induced by normal random functions with the same proper covariance kernel R , and mean value functions equal to m_1 and m_2 respectively. More presisely, for $i = 1,2$, P_i is defined so that $P_i[\{\omega:\omega(t_1)<a_1,\ldots,\omega(t_n)<a_n\}]$ is equal to the probability that n normally distributed random variables $X(t_1),\ldots,X(t_n)$ with means $m_i(t_1),\ldots,m_i(t_n)$ and covariances $R(t_j,t_k)$ satisfy the inequalities $X(t_1)< a_1,\ldots,X(t_n) < a_n$.

Both P_1 and P_2 are probability measures on the probabilizable space $(\Omega, \mathscr{C})$. Consequently, by the Lebesgue decomposition theorem for measures, it follows that there is a set N of P_1-measure 0, and a non-negative function, denoted $\frac{dP_2}{dP_1}$, which is integrable over Ω with respect to P_1, such that for every set B in $\mathscr{C}$

$$(9.5) \qquad P_2(B) = \int_B \left(\frac{dP_2}{dP_1}\right) dP_1 + P_2(BN) \quad .$$

If the set N also has P_2-measure 0, then P_2 is absolutely continuous with respect to P_1, and $\frac{dP_2}{dP_1}$ is called the Radon-Nikodym derivative, or probability density function, of P_2 with respect to P_1. If $\frac{dP_2}{dP_1} > 0$ with P_1-probability 1(as well as $P_2(N) = 0$), then P_1 is absolutely continuous with respect to P_2 and

$$\frac{dP_1}{dP_2} = 1 \Big/ \left(\frac{dP_2}{dP_1}\right) \quad . \tag{9.6}$$

Two measures, P_1 and P_2, which are absolutely continuous with respect to one another, are called <u>equivalent</u>.

On the other hand, if $\frac{dP_2}{dP_1} = 0$ with P_1-probability 1, so that $1 = P_2(\Omega) = P_2(N)$ for some fixed set N satisfying $P_1(N) = 0$, then P_1 and P_2 are said to be <u>orthogonal</u> to each other.

Let $\{T_n, n = 1,2,\ldots\}$ be a sequence of finite subsets of T which are monotone increasing (that is, $T_n \subset T_{n+1}$) and such that $\bigcup_{n=1}^{\infty} T_n = T$ in the case that T is countable, and $\bigcup_{n=1}^{\infty} T_n$ is dense in T in the case that T is separable metric. For each integer n, let $\mathcal{B}_n$ be the smallest σ-field in $\mathcal{C}$ containing, for any t in T_n and real number a, the set $\{\omega \in \Omega : \omega(t) < a\}$. The sequence of σ-fields $\mathcal{B}_n$ is then monotone non-decreasing. For each integer n, let $P_1^{(n)}$ and $P_2^{(n)}$ be the restrictions of P_1 and P_2, respectively, to the σ-field $\mathcal{B}_n$.

Assume that the index set T is either countable or a separable metric space. Assume, for each n, that the Radon-Nikodym derivative $\frac{dP_2^{(n)}}{dP_1^{(n)}}$ exists. By the theory of martingales, it follows, letting $\frac{dP_2}{dP_1}$ be defined by (9.5), that $\frac{dP_2^{(n)}}{dP_1^{(n)}}$ converges to $\frac{dP_2}{dP_1}$, with probability one with respect to P_1-measure. If, further,

$$(9.7) \qquad \lim_{n \to \infty} \int_\Omega \left(\frac{dP_2^{(n)}}{dP_1^{(n)}} \right)^2 dP_1 < \infty$$

then P_2 is absolutely continuous with respect to P_1, with Radon-Nikodym derivative

$$(9.8) \qquad \frac{dP_2}{dP_1} = \lim_{n \to \infty} \frac{dP_2^{(n)}}{dP_1^{(n)}}$$

where the limit in (9.8) exists both as a limit in quadratic mean and as a limit with P_1-probability one. On the other hand, by a theorem of Kraft (1955), it follows that P_1 and P_2 are orthogonal and $\frac{dP_2}{dP_1} = 0$ if, and only if,

$$(9.9) \qquad \lim_{n \to \infty} \int_\Omega \left(\frac{dP_2^{(n)}}{dP_1^{(n)}} \right)^{1/2} dP_1 = 0 \quad .$$

<u>We now apply these considerations to obtain an explicit formula</u>

for the probability density function of P_2 with respect to P_1, where P_1 and P_2 are the measures induced by Normal random functions $\{X(t), t \in T\}$ with common proper covariance kernel R and respective means m_1 and m_2. We assume that the index set T is either countable or a separable metric space; in the latter case, we assume that R is weakly continuous (in the sense that it satisfies the conditions given in Theorem 2E). We assume further that the covariance kernel R has the property that it is non-singular on every finite subset of T. Then $P_2^{(n)}$ is absolutely continuous with respect to $P_1^{(n)}$, with Radon-Nikodym derivative given by

$$\frac{dP_2^{(n)}}{dP_1^{(n)}} = \exp\left\{(X-m_1, m_2-m_1)_n - \frac{1}{2}\|m_2-m_1\|_n^2\right\} \tag{9.10}$$

where we define, for any two functions f and g on T_n,

$$(f,g)_n = \sum_{s,t \in T_n} f(s)R^{-1}(s,t)g(t), \quad \|f\|_n^2 = (f,f) \quad . \tag{9.11}$$

Further using the fact that $E[e^Z] = e^{\frac{1}{2}\mathrm{Var}[Z]}$ if Z is normal with mean 0,

$$\int_\Omega \left(\frac{dP_2^{(n)}}{dP_1^{(n)}}\right)^2 dP_1 = e^{\|m_2-m_1\|_n^2} \tag{9.12}$$

$$(9.13) \qquad \int_\Omega \left(\frac{dP_2^{(n)}}{dP_1^{(n)}} \right)^{1/2} dP_1 = e^{-\frac{1}{8}\|m_2-m_1\|_n^2} .$$

Now $\|m_2-m_1\|_n$ is monotone increasing as n tends to ∞ . Further, by Theorem 6C,

$$(9.14) \qquad \lim_{n\to\infty} \|m_2-m_1\|_n^2 < \infty \quad \text{if, and only if,} \quad m_2-m_1 \in H(R) .$$

In view of the considerations developed we may now state the following theorem.

Theorem 9A: Under the assumptions stated prior to (9.10), the measures P_1 and P_2 are either equivalent or orthogonal. If m_2-m_1 does not belong to $H(R)$, then P_1 and P_2 are orthogonal. If m_2-m_1 does belong to $H(R)$, then P_2 and P_1 are equivalent, and

$$(9.15) \qquad \frac{dP_2}{dP_1} = \exp \{< X-m_1, m_2-m_1 > - \tfrac{1}{2} \|m_2-m_1\|_{H(R)},\}$$

where $< X-m_1, m_2-m_1 >$ is defined by

$$(9.16) \qquad < X-m_1, m_2-m_1 > = \lim_n (X-m_1, m_2-m_1)_n ,$$

the limit in (9.16) holding both, as a limit in quadratic mean, and as a limit with P_1 -probability one.

To complete the proof of Theorem 9A, we need only prove that the limit in (9.16) exists. We prove this by proving more generally the convergence of expressions of the form $(Y,g)_n$, where $\{Y(t), t \in T\}$ is a Normal random function with mean 0 and covariance kernel R, and g is a function in H(R).

Theorem 9B: Let $\{Y(t), t \in T\}$ be a random function with mean 0, whose covariance kernel R and index set T satisfy the same conditions as in Theorem 9A. Let ψ be the congruence from H(R) onto $L_2(Y(t), t \in T)$ such that

$$(9.17) \qquad \psi(R(\cdot,t)) = Y(t), \quad t \in T .$$

Then for any function g in H(R)

$$(9.18) \qquad (Y,g)_n = \psi(E^*[g|R(\cdot,t), t \in T_n])$$

$$(9.19) \qquad \lim_n (Y,g)_n = \psi(g)$$

where the limit in (9.19) exists as a limit in quadratic mean. Further, for $n = 1,2,\ldots$

$$(9.20) \qquad (Y,g)_n = E^*[(Y,g)_{n+1}|Y(t), t \in T_n]$$

so that $(Y,g)_n$ is a martingale in the wide sense. Further, if $\{Y(t), t \in T\}$ is Normal, then from (9.20) it follows that $(Y,g)_n$

is a martingale in the strict sense,

$$(9.21) \qquad (Y,g)_n = E[(Y,g)_{n+1} | Y(t), t \in T_n] \quad ,$$

and the limit in (9.19) exists with probability one.

<u>Proof</u>: That (9.18) holds follows from the fact that, by (7.26),

$$E^*[g | R(\cdot,t), t \in T_n] = \sum_{s,t \in T_n} g(s) R^{-1}(s,t) R(\cdot,t) \quad .$$

Taking the limit of both sides of (9.18) we obtain (9.19). We obtain (9.20) from (9.18) by noting that for f in $H(R)$ and $n = 1,2,\ldots$

$$(9.22) \qquad \psi(E^*[f | R(\cdot,t), t \in T_n]) = E^*[\psi(f) | Y(t), t \in T_n] \quad .$$

Letting $f = E^*[g | R(\cdot,t), t \in T_{n+1}]$, we obtain (9.20). That (9.20) implies (9.21) follows by the following formula (see Anderson, 1958,p.28):

$$(9.23) \qquad E[Z | Y(t), t \in T_n] = \sum_{s,t \in T_n} \rho_Z(s) R^{-1}(s,t) Y(t)$$

where Z and $\{Y(t), t \in T_n\}$ are jointly Normally distributed random variables with means 0 and covariances

$$\rho_Z(t) = E[ZY(t)], R(s,t) = E[Y(s)Y(t)] \quad . \tag{9.24}$$

The proof of Theorem 9B is complete.

The tools are now at hand to discuss minimum variance unbiased estimation of the mean value function of a Normal random function $\{X(t), t \in T\}$, whose proper covariance kernel R is known, but whose mean value function m is an unknown member of a known subset of $\mathcal{M}$. Let P_m be the measure corresponding to mean m, and let P_o be the measure corresponding to the mean value function identically equal to 0. Now for any m in $\mathcal{M}$.

$$\frac{dP_m}{dP_o} = e^{< X,m > - \frac{1}{2} \|m\|^2} \quad . \tag{9.25}$$

Therefore for any fixed m_o

$$k(m) = \frac{dP_m}{dP_{m_o}} = \frac{e^{< X,m > - \frac{1}{2} \|m\|^2}}{e^{< X,m_o > - \frac{1}{2} \|m_o\|^2}} \tag{9.26}$$

$$= e^{< X,m-m_o > - \frac{1}{2} \|m\|^2 + \frac{1}{2} \|m_o\|^2} \quad .$$

By the considerations of section 8, in order to find the m_o -minimum variance unbiased estimate of a function f defined on $\mathcal{M}$ it suffices to express f in terms of linear operations on the kernel

$$K(m_1,m_2) = E_{m_o}[k(m_1)k(m_2)] = e^{(m_1-m_o,m_2-m_o)} \quad . \tag{9.27}$$

To prove (9.27) we use the formula $E(e^Z) = e^{E(Z) + \frac{1}{2}\mathrm{Var}(Z)}$ if Z is Normally distributed, and the fact that, for any g in $H(R)$,

$$E_m[< X,g >] = (m,g), \mathrm{Var}[< X,g >] = (g,g) \quad . \tag{9.28}$$

To prove (9.28) use the fact that

$$< X,g > = \lim_{n \to \infty} (X,g)_n \quad . \tag{9.29}$$

Of course (9.29) requires the assumptions on T and R made previously in this section. In the next section we shall show that (9.28) is valid without any assumptions on T and R .

We are particularly interested in obtaining the minimum variance estimate of functions f which are <u>linearly estimatable</u>. A function f on $\mathcal{M}$ is said to be linearly estimable if there exists a random variable U in $L_2(X(t),t \in T)$ such that

$$E_m[U] = f(m), \qquad m \in \mathcal{M}. \tag{9.30}$$

Now, because $L_2(X(t),t \in T)$ and $H(R)$ are congruent, to any U in $L_2(X(t),t \in T)$ there is a g in $H(R)$ such that $U = < X,g >$.

Consequently, we may define a function f on $\mathscr{M}$ to be linearly estimable if there exists a function g in $H(R)$ such that

$$(9.31) \qquad E_m[< X,g >] = (m,g) = f(m) \ , \qquad m \in \mathscr{M}.$$

In particular, $m(t_o) = (m,R(\cdot,t_o))$ is linearly estimable. We now show that <u>the uniformly minimum variance estimate</u> $\hat{f}$ <u>of</u> f <u>is given by</u>

$$(9.32) \qquad \hat{f} = < X,E^*[g|\bar{\mathscr{M}}] >$$

where $\bar{\mathscr{M}}$ is the smallest Hilbert sub-space of $H(R)$ containing $\mathscr{M}$, and g is any function satisfying (9.31). The minimum variance is given by

$$(9.33) \qquad \mathrm{Var}[\hat{f}] = \|E^*[g|\bar{\mathscr{M}}]\|^2_{H(R)} \ .$$

It should be clear that the best estimate $\hat{f}$ should depend on $\bar{\mathscr{M}}$, rather than on $\mathscr{M}$, since if (9.31) holds for all m in $\mathscr{M}$, it then holds for all m in $\bar{\mathscr{M}}$.

To prove (9.32), let $\varphi_1,\varphi_2,\ldots$ be an orthonormal basis for $\bar{\mathscr{M}}$. Letting $m_2 = m$ and $\beta_n = (m_1-m_o,\varphi_n)$ we obtain

$$K(m_1,m) = \exp \sum_{n=1}^{\infty} \beta_n(m-m_o,\varphi_n) \ .$$

Differentiating we obtain

$$\frac{\partial}{\partial \beta_n} K(m_1, m) \Bigg|_{m_1 = m_o} = (m - m_o, \varphi_n) \quad .$$

Consequently,

$$\begin{aligned} f(m) &= (g, m_o) + (g, m - m_o) \\ &= (g, m_o) + \sum_{n=1}^{\infty} (g, \varphi_n)(m - m_o, \varphi_n) \\ &= (g, m_o) + \sum_{n=1}^{\infty} (g, \varphi_n) \frac{\partial}{\partial \beta_n} K(m_o, m) \quad . \end{aligned}$$

Consequently the best unbiased estimate of f is given by

$$\psi(f) = (g, m_o) + \sum_{n=1}^{\infty} (g, \varphi_n) \frac{\partial}{\partial \beta_n} k(m_o) \quad .$$

Now, letting $\beta_n = (m - m_o, \varphi_n)$

$$k(m) = \exp \left\{ \sum_{n=1}^{\infty} \beta_n \; [< X, \varphi_n > - (m, \varphi_n)] - \frac{1}{2} \sum_{n=1}^{\infty} \beta_n^2 \right\}$$

$$\frac{\partial}{\partial \beta_n} k(m_o) = < X, \varphi_n > - (m_o, \varphi_n) \quad .$$

Finally,

$$\psi(f) = \sum_{n=1}^{\infty} (g,\varphi_n) < X,\varphi_n > \; = \; < X,E^*[h|\bar{\mathcal{M}}] > \quad .$$

The proof of (9.32) is complete.

By (9.32) we may give explicit expressions for the best estimates of the mean $m(t_o)$ or its derivative $m'(t_o)$:

$$(9.34) \qquad \hat{m}(t_o) = < X,E^*[R(\cdot,t_o)|\bar{\mathcal{M}}] >$$

$$\hat{m}'(t_o) = < X,E^*[\frac{\partial}{\partial s} R(\cdot,t_o)|\bar{\mathcal{M}}] > \quad .$$

In the foregoing, we have assumed that the class of possible means was a set $\mathcal{M}$ in H(R). However, the set $\mathcal{M}$ is in turn often paramatrized by a vector parameter $\beta = (\beta_1,\ldots,\beta_q)$ varying in a set B of q-dimensional Euclidean space. In particular let us suppose that $\mathcal{M}$ consists of the set of functions of the form

$$m(t) = \beta_1\varphi_1(t) + \ldots + \beta_q\varphi_q(t)$$

for some β in B and functions $\varphi_1,\ldots,\varphi_2$ in H(R) . Under suitable conditions on B, one may obtain a formula analogous to (8.30); for any estimable function

$$(9.35) \qquad f(\beta) = c_1\beta_1 + \ldots + c_q\beta_q$$

define

$$f^* = \sum_{i,j=1}^{q} c_i \, M_{ij}^{-1} \, \varphi_j \tag{9.36}$$

where $M_{ij} = (\varphi_i, \varphi_j)$. Then

$$\hat{f} = < X, f^* > \tag{9.37}$$

$$\mathrm{Var}[\hat{f}] = \|f^*\|^2_{H(R)} = \sum_{i,j=1}^{q} c_i M_{ij}^{-1} c_j \quad . \tag{9.38}$$

To conclude this section we point out how one may formulate the expressions for $\hat{f}$ and $\mathrm{Var}[\hat{f}]$ given by (9.31) and (9.32) in terms of the representation of the random function $\{X(t), t \in T\}$ with respect to an orthogonal random set function. Suppose that the proper covariance kernel R admits a representation

$$R(s,t) = \int_Q f(s)f(t)du \tag{9.39}$$

as in (4.1). Then any function g in $H(R)$ may be represented

$$g(t) = \int_Q g^* f(t)du \tag{9.40}$$

in terms of a unique function g^* in $L_2(f(\epsilon), t \in T)$. Further

$\{X(t), t \in T\}$ has an orthogonal decomposition

$$(9.41) \qquad X(t) = \int_Q f(t)\,dZ$$

in terms of which we may express

$$(9.42) \qquad < X, g > = \int_Q g^* \, dZ \ .$$

For any set $\mathcal{M}$ of functions in H(R), let $\mathcal{M}^*$ be a set of functions in $L_2(f(t), t \in T)$ which correspond to the functions in $\mathcal{M}$ under the correspondence (9.40) . Then for any g in H(R)

$$(9.43) \qquad E^*[g| \mathcal{M}] = E^*[g^*|\mathcal{M}^*] \ .$$

These formulas will play an important role in our later study of the large sample properties of minimum variance estimates.

10. Minimum variance linear unbiased estimates.

Let $\{X(t), t \in T\}$ be a random function whose proper covariance kernel R is known but whose mean value function $m(t) = E[X(t)]$ is not assumed to be known, but is only assumed to belong to a known class $\mathcal{M}$ which is a subset of H(R). We do not assume a knowledge of the probability law of the random function. Consequently, in order to estimate the value $m(t_0)$ of the mean value function m at any point

t_o in T, we consider only random variables which are linear functionals over the observed random variables $X(t)$. We then adopt as our estimate of $m(t_o)$ that unbiased linear estimate which has (uniformly) minimum variance.

In this section we show how the problem of finding the minimum variance unbiased linear estimate may be reduced to a minimization problem in a reproducing kernel Hilbert space. To illustrate the ideas involved, we first consider the case where the index set $T = \{1,2,\dots,n\}$, and where the inverse matrix $\{R^{-1}(s,t)\}$ exists.

Given the finite family of random variables $\{X(t), t \in T\}$ any random variable $U = \sum_{t \in T} c(t)X(t)$ in $L(X(t), t \in T)$ is called a linear estimate. It is an unbiased linear estimate of $m(t_o)$ if, for every m in $\mathcal{M}$,

$$(10.1) \qquad E_m\left[\sum_{t \in T} c(t)X(t)\right] = \sum_{t \in T} c(t)m(t) = m(t_o) \quad .$$

The variance of an unbiased estimate

$$(10.2) \qquad \mathrm{Var}\left[\sum_{t \in T} c(t)X(t)\right] = \sum_{s,t \in T} c(s)R(s,t)c(t)$$

is independent of m. Consequently the uniformly minimum variance linear unbiased estimates $\widehat{m(t_o)}$ of $m(t_o)$ always exist.

Now any linear estimate U can be written in the form

$U = (g,X)_n$ for some vector g in $H(R)$, where the inner product $(a,b)_n$ is defined by (8.19). Further one may verify that

$$(10.3) \qquad E_m[(g,X)_n] = (g,m), \mathrm{Var}[(g,X)_n] = (g,g)_n \quad .$$

One therefore sees that the minimum variance linear unbiased estimate may be represented in the form

$$(10.4) \qquad \hat{m}(t_o) = (m^*(t_o),X)_n$$

where $m^*(t_o)$ is the vector in $H(R)$ which has minimum norm $\|g\|_n$ among all vectors in R subject to the condition

$$(10.5) \qquad (g,m)_n = m(t_o) \qquad \text{for all } m \text{ in } \mathcal{M} .$$

<u>We have thus shown that the problem of uniformly minimum variance linear unbiased estimation may be posed as a minimization problem in a suitable reproducing kernel Hilbert space</u>. From this fact we can immediately write an explicit formula for $\hat{m}(t_o)$. Let $\bar{\mathcal{M}} = L(m, m\in\mathcal{M})$ be the smallest linear space containing all posssible mean value functions in $\mathcal{M}$. The vector which has minimum norm among all vectors satisfying (10.5) may also be characterized as the unique vector in $\bar{\mathcal{M}}$ satisfying (10.5). <u>We thus see that the space $\mathcal{M}$ may just as well be taken to be the linear space $\bar{\mathcal{M}}$, since the expression for the</u>

<u>uniformly minimum variance estimate is the same in either case.</u>

To give an explicit formula for $\hat{m}(t_o)$, we write

$$(10.6) \qquad \hat{m}(t_o) = (X, E^*[R(\cdot,t_o|\bar{\mathcal{M}}])_n$$

In particular, if $\varphi_1,\ldots,\varphi_q$ are q linearly independent functions spanning $\bar{\mathcal{M}}$, with non-singular covariance matrix $M_{ij} = (\varphi_i,\varphi_j)_n$, then

$$(10.7) \qquad \hat{m}(t_o) = \sum_{i,j=1}^{q} \varphi_i(t_o) M_{ij}^{-1} (\varphi_j, X)_n$$

$$(10.8) \qquad \operatorname{Var}[\hat{m}(t_o)] = \sum_{i,j=1}^{q} \varphi_i(t_o) M_{ij}^{-1} \varphi_j(t_o) \quad .$$

Comparing (10.6) with (9.31), we see that for a Normal random function $\{X(t), t \in T\}$, the minimum variance unbiased linear estimate coincides with the minimum variance unbiased estimate. This result, proved here in the case that T is finite, will be proved below for arbitrary T.

We next consider the problem of uniformly minimum variance linear unbiased estimation in the case of a random function $\{X(t), t \in T\}$ with non-finite index set T, with proper covariance kernel R, and wsith mean value m belonging to a subset $\mathcal{M}$ of $H(R)$.

The first problem is to define what is meant by a <u>linear estimate</u>. Let $L(X(t), t \in T)$ and $L(R(\cdot,t), t \in T)$ be the linear spaces spanned respectively by $\{X(t), t \in T\}$ and $\{R(\cdot,t), t \in T\}$. Any function g in $L(R(\cdot,t), t \in T)$ is of the form

$$g = \sum_{i=1}^{n} c_i R(\cdot, t_i) \tag{10.9}$$

for some integer n, real constants $c_1, \ldots, c_n$ and points $t_1, \ldots, t_n$ in T. Define

$$< X,g > = \sum_{i=1}^{n} c_i X(t_i) \quad . \tag{10.10}$$

That $< X,g >$ is well defined, no matter what the true mean m, follows by the fact that

$$\begin{aligned} E_m \Big| \sum_{i=1}^{n} c_i X(t_i) \Big|^2 &= \sum_{i,j=1}^{n} c_i c_j \, \{R(t_i,t_j) + m(t_i)m(t_j)\} \\ &= \Big\| \sum_{i=1}^{n} c_i R(\cdot,t_i) \Big\|^2 + \Big|(m, \sum_{i=1}^{n} c_i R(\cdot,t_i))\Big|^2 \quad . \end{aligned} \tag{10.11}$$

It is clear that $< X,g >$ defines a one-one linear mapping from $L(R(\cdot,t), t \in T)$ onto $L(X(t), t \in T)$ with the following properties: for any t in T, and functions g,h in $L(R(\cdot,t), t \in T)$,

$$< X, R(\cdot,t) > = X(t) \tag{10.12}$$

$$E_m[< X,g >] = (m,g) \tag{10.13}$$

$$\mathrm{Cov}[< X,g >, < X,h >] = (g,h) \quad . \tag{10.14}$$

Now any function g in $H(R)$ may be represented as the limit of a sequence $g_1, g_2, \ldots,$ in $L(R(\cdot,t), t \in T)$. The coresponding random variables $< X,g_1 >, < X,g_2 >, \ldots$ are a Cauchy sequence, and have a limit which we denote by $< X,g >$. We have thus defined a transformation $< X,g >$ from $H(R)$ onto $L_2(X(t), t \in T)$ which is one-one, linear, and satisfies (10.12) , (10.13) , and (10.14) .

By a linear estimate we mean a random variable which is the image $< X,g >$ of some function g in $H(R)$.

We may now define the notion of an estimable function. A real valued function f defined on $\mathcal{M}$ is said to be estimable if, and only if, there exists an unbiased linear estimate of f . More precisely, f is estimable if, and only if, there exists a function g in $H(R)$ such that

$$E_m[< X,g >] = (m,g) = f(m) \tag{10.15}$$

for every m in $\mathcal{M}$.

A linear estimate $< X,g >$ satisfying (10.15) is called an unbiased linear estimate of f . If we introduce the reproducing kernel Hilbert space $H(K_{\mathcal{M}})$, coresponding to the kernel $K_{\mathcal{M}}$ defined on $\mathcal{M} \times \mathcal{M}$ by

$$K_{\mathcal{M}}(m_1, m_2) = (m_1, m_2) \tag{10.16}$$

then f is estimable if, and only if, $f \in H(K_{\mathcal{M}})$.

In particular, for any t_o in T , the mean $m(t_o) = E_m[X(t_o)] = (m, R(\cdot, t_o))$ is estimable. For any g in H(R), $< X, g >$ is an unbiased linear estimate of $m(t_o)$ if, and only if,

$$E_m[< X, g >] = (m, g) = m(t_o) \tag{10.17}$$

for every m in $\mathcal{M}$.

Using the ideas assembled to this point we see immediately the validity of the following theorem.

Theorem 10A: Let $\{X(t), t \in T\}$ have known proper covariance kernel R , and unknown mean value function m, belonging to a subset $\mathcal{M}$ of H(R) . To every estimable function f defined on $\mathcal{M}$ there exists a unique linear estimate, denoted by $\hat{f} = < X, f^* >$, which is the uniformly minimum variance linear unbiased estimate of f ; further

$$\operatorname{Var}[\hat{f}] = \|f^*\|^2_{H(R)} \tag{10.18}$$

A linear estimate $< X, f^* >$ is the uniformly minimum variance unbiased linear estimate of f if, and only if, f^* satisfies any one of the following equivalent conditions (where $\bar{\mathcal{M}}$ is the Hilbert space spanned by the functions m in $\mathcal{M}$):

(i) f^* is the function in H(R) which has minimum norm among all functions satisfying (10.15);

(ii) f^* is the unique function in $\bar{\mathcal{M}}$ satisfying (10.15);

(iii) $f^* = E^*[g | \bar{\mathcal{M}}]$, where g is any function in H(R) satisfying (10.15) .

From Theorem 10A, one sees that the basic formula for the best estimate of $f = (m,g)$ is

(10.19) $$\hat{f} = < X, E^*[g | \bar{\mathcal{M}}] > .$$

Equation (10.19) provides a general solution to the problem of minimum variance unbiased linear estimation. That (10.19) constitutes a practically useful formula can only be shown by exhibiting examples of its use. We shall do this in later papers.

We note here two important special cases of (10.19) . If each function m in $\mathcal{M}$ is of the form

(10.20) $$m(t) = \beta_1 \varphi_1(t) + \ldots + \beta_q \varphi_q(t)$$

then for any estimable function $f(\beta) = c_1\beta_1 + \ldots + c_q\beta_q$

$$(10.21) \quad \hat{f} = \langle X, f^* \rangle, \; f^* = \sum_{i,j=1}^{q} c_i M_{ij}^{-1} \varphi_j, \; \mathrm{Var}[\hat{f}] = \|f^*\|^2 ,$$

where $M_{ij} = (\varphi_i, \varphi_j)$. If each function m in $\mathcal{M}$ may be expanded in terms of orthonomal functions $\{\varphi_n\}$ by

$$(10.22) \qquad m(t) = \sum_{n=1}^{\infty} \beta_n \varphi_n, \quad \sum_{n=1}^{\infty} \beta_n^2$$

then for any estimable function $f(\beta) = \sum_{n=1}^{\infty} c_n \beta_n$

$$(10.23) \qquad \hat{f} = \langle X, f^* \rangle , \; f^* = \sum_{n=1}^{\infty} c_n \varphi_n .$$

11. Minimum variance unbiased prediction.

The problem of prediction of the value of a random variable Z on the basis of the observed values of a random function $\{X(t), t \in T\}$ was mentioned in section 6. If one knows the joint probability law of Z and $\{X(t), t \in T\}$, then the predictor which minimizes the mean square error of prediction is the conditional expectation $E[Z|X(t), t \in T]$. On the other hand if one only knows the second moments $E[Z^2], E[X(s)X(t)]$ and $E[ZX(t)]$ for any s, t in T, then one may form the linear predictor which minimizes the mean square error of prediction, given by the projection

$E^*[Z|X(t), t \in T]$. In this section we are concerned with the problem of prediction in the case that the true probability law or second moments are not known but must be estimated.

We first discuss the problem of <u>uniformly minimum variance linear unbiased prediction</u>, which arises as follows. Assume, as in section 10, that $\{X(t), t \in T\}$ is a random function whose proper covariance kernel R is known, but whose mean value function m is known only to belong to a class $\mathcal{M}$ of possible mean value functions. Let Z be a random variable such that

$$(11.1) \quad Var[Z], \rho_Z(t) = Cov[Z, X(t)] = E_m[(Z-E_m[Z])(X(t)-m(t))]$$

are known and do not depend on the true value of m ; however, the mean of Z,

$$(11.2) \qquad e_Z(m) = E_m[Z] ,$$

depends on the true value of m . The functional form of e_Z as a function of m is assumed to be known. One case of particular importance is when $Z = X(t_o)$, where $t_o \notin T$; then $e_Z(m) = m(t_o)$.

Conditions that the function e_Z be estimable are given by equation (10.15) . We shall see in the course of the discussion that e_Z is estimable if and only if Z is predictable.

We define a random variable Z to be <u>predictable</u> if and only if there exists an unbiased linear estimate of Z ; more precisely, Z is predictiable if and only if there exists a function g in $H(R)$ such that

$$(11.3) \qquad E_m[< X,g >] = E_m[Z] \qquad \text{for every } m \text{ in } \mathcal{M}$$

or equivalently

$$(11.4) \qquad (g,m) = e_Z(m) \qquad \text{for every } m \text{ in } \mathcal{M}.$$

Comparing (11.4) with (10.15), we see that Z is predictable if and only if its mean function e_Z is estimable.

A linear estimate $< X,g >$ satisfying (11.3) is called an unbiased linear estimate of Z . The mean square error is then given, independently of m by

$$(11.5) \quad E_m|Z-<X,g>|^2 = \mathrm{Var}[Z-<X,g>]=\mathrm{Var}[Z]+\mathrm{Var}[<X,g>]-2\,\mathrm{Cov}[Z,<X,g>] \;.$$

We leave it for the reader to verify that for any random variable Z

$$(11.6) \qquad \rho_Z(t) = \mathrm{Cov}\ [Z,X(t)] \in H(R)$$

(11.7) $$\text{Cov}\ [Z, < X,g >] = (\rho_Z, g) \qquad \text{for any}\ \ g \in H(R)\ .$$

In view of (11.7), the mean square prediction error of the linear estimate $< X,g >$ may be written

(11.8) $$\begin{aligned} E_m|Z - < X,g >|^2 &= \text{Var}\ [Z] + (g,g) - 2(\rho_Z, g) \\ &= \text{Var}\ [Z] - \|\rho_Z\| + \|g-\rho_Z\|\ . \end{aligned}$$

We have thus reduced the problem of finding the best unbiased linear predictor $\hat{Z}$ of Z to finding the vector g in $H(R)$ which among all vectors satisfying (11.4) minimizes the norm $\|g-\rho_Z\|$. Alternately, we may write

(11.9) $$\hat{Z} = < X, \rho_Z + h >$$

where h is the vector of minimum norm satisfying

(11.10) $$(h,m) = e_Z(m) - (\rho_Z, m) \qquad \text{for every}\ \ m\ \ \text{in}\ \ \mathcal{M}\ .$$

We are thus led to the following theorem.

<u>Theorem 11A</u>: Let $\{X(t), t \in T\}$ have proper covariance kernel R and unknown mean value function m, belonging to $\mathcal{M} \subset H(R)$. Let Z

be a random variable satisfying (11.1), and whose mean e_Z is an estimable function on $\mathscr{M}$. Then there exists a unique linear estimats, denoted by $\hat{Z} = \langle X, Z^* \rangle$, which is the uniformly minimum variance unbiased linear predictor of Z; further, for every m in $\mathscr{M}$

$$(11.11) \qquad E_m|\hat{Z}-Z|^2 = \mathrm{Var}[Z] - \|\rho_Z\|^2 + \|Z^*\|^2 .$$

A linear estimate Z^* is the uniformly minimum variance linear unbiased predictor of Z if, and only if, it satisfies any one of the following equivalent conditions (letting $\bar{\mathscr{M}}$ be the Hilbert space spanned by the functions m in $\mathscr{M}$):

(i) $Z^* = \rho_Z + h$, where h is the function in $H(R)$ which has minimum norm among all functions satisfying (11.8);

(ii) $Z^* = \rho_Z + h$, where h is the unique function in $\bar{\mathscr{M}}$ satisfying (11.8),

(iii) $Z^* = \rho_Z + E^*[g-\rho_Z|\bar{\mathscr{M}}]$, where g is any function in $H(R)$ satisfying (11.4).

To illustrate the use of Theorem 11A, let us note that in the case that $\bar{\mathscr{M}}$ is finite dimensional and (10.19) holds then

$$(11.12) \qquad \hat{Z} = \langle \rho_Z, X \rangle + \sum_{i,j=1}^{q} d_i M_{ij}^{-1} \langle \varphi_j, X \rangle$$

where

$$d_j = (g,\varphi_j)-(\rho_Z,\varphi_j) \quad . \tag{11.13}$$

Also

$$E|\hat{Z}-Z|^2 = \mathrm{Var}[Z]-\|\rho_Z\|^2 + \sum_{i,j=1}^{q} d_i M_{ij}^{-1} d_j \quad . \tag{11.14}$$

We next discuss briefly the problem of minimum variance unbiased prediction, mainly for the purpose of showing that for Normally distributed random variables the best unbiased predictor is the best unbiased linear predictor.

The general problem of minimum variance unbiased prediction may be posed as follows. Let $\{X(t), t \in T\}$ be a random function defined on a probabilizable space $(\Omega, \mathcal{A})$. Assume that the true probability measure over $(\Omega, \mathcal{A})$ is known only to belong to a family $\{P_\theta, \theta \in \Lambda\}$. Let $\mathcal{A}_X$ denote the smallest sigma-field, contained in $\mathcal{A}$, with respect to which each random variable $X(t)$ is measurable. Fix $\theta_o \in \Lambda$. Given a random variable Z in $L_2(\Omega, \mathcal{A}, P_{\theta_o})$, we define the θ_o -efficient predictor of Z to be that random variable U in $L_2(\Omega, \mathcal{A}_X, P_{\theta_o})$ which is closest to Z , in the sense of minimizing the mean square error $E_{\theta_o}|U-Z|^2$. It is well known that the θ_o -efficient predictor of Z is the conditional expectation $E_{\theta_o}[Z|X(t), t \in T]$. However, this conditional expectation will, in general, depend functionally on θ .

If one desires a uniformly efficient predictor of Z, one may desire to restrict consideration to unbiased predictors.

Assume that for each θ_o in Λ, the Radon-Nikodym derivative

$$k_{\theta_o}(\theta) = \frac{dP_\theta}{dP_{\theta_o}} \tag{11.15}$$

exists for each θ in Λ, and is square integrable with respect to P_{θ_o}. A random variable U in $L_2(\Omega, \mathcal{Q}_X, P_{\theta_o})$ is said to be an <u>unbaised predictor</u> of Z if

$$E_\theta[U] = E_{\theta_o}[U\, k_{\theta_o}(\theta)] = E_\theta[Z] \qquad \text{for every } \theta \text{ in } \Lambda. \tag{11.16}$$

Writing the mean square error of U as a predictor of Z in the form

$$E_{\theta_o}|U-Z|^2 = E_{\theta_o}|U-E_{\theta_o}[Z|X(t), t\epsilon T]|^2 + E_{\theta_o}|Z-E_{\theta_o}[Z|X(t), t\epsilon T]|^2 , \tag{11.17}$$

we see that the θ_o-efficient unbiased predictor of Z is equivalent to the θ_o-efficient unbiased predictor of $E_{\theta_o}[Z|X(t), t \in T]$, which in turn may be seen to be of the form

$$E_{\theta_o}[Z|X(t), t \in T] + V^* , \tag{11.18}$$

where V^* is the random variable in $L_2(\Omega, \mathcal{A}_X, P_{\theta_o})$ of minimum norm $E_{\theta_o}|V^*|^2$ among all random variables V satisfying, for all θ in Λ.

$$(11.19) \qquad E_{\theta_o}[V\, k_{\theta_o}(\theta)] = E_\theta[Z] - E_{\theta_o}[E_{\theta_o}[Z|X(t), t \in T] k_{\theta_o}(\theta)]$$

$$= E_\theta[Z] - E_\theta[E_{\theta_o}[Z|X(t), t \in T]] \quad .$$

To find the solution of (11.19) of minimum norm, one uses the techneques discussed in section 7 .

Let us apply these considerations to the case of jointly Normally distributed random variables Z and $\{X(t), t \in T\}$ for which

$$R(s,t) = \mathrm{Cov}[X(s),X(t)], \rho_Z(t) = \mathrm{Cov}[Z,X(t)], \mathrm{Var}[Z]$$

are known, but

$$m(t) = E_m[X(t)], e_Z(m) = E_m[Z]$$

are unknown depending on an unknown member m of a known subspace $\mathcal{M}$ of $H(R)$. It may be shown that

$$(11.20) \qquad E_m[Z|X(t), t \in T] = \langle \rho_Z, X \rangle + e_Z(m) - (\rho_Z, m)$$

$$(11.21) \qquad E_m|Z-E_m[Z|X(t),t \in T]|^2 = Var[Z]-\|\rho_Z\|^2 \quad .$$

Therefore

$$(11.22) \qquad E_m[Z-E_{m_o}[Z|X(t),t\in T]] = e_Z(m)-(\rho_Z,m-m_o) \quad .$$

The m_o- best unbiased estimater of this parametric function of m is, in view of (10.12), given by

$$(11.23) \qquad < X,E^*[g-\rho_Z | \mathscr{M}] > - e_Z(m_o) + (\rho_Z,m_o)$$

where g is any function in $H(R)$ such that $(g,m) = e_Z(m)$ for all m in $\mathscr{M}$. Now the best predictor of Z is given by the sum of (11.23), and (11.20) with $m = m_o$. Therefore, the best predictor of Z is

$$(11.24) \qquad < \rho_Z,X > + < X,E^*[g-\rho_Z | \tilde{\mathscr{M}}] >$$

which agrees with Theorem 11A.

The reader should compare (11.24) with (11.20), and note that the best unbiased predictor of Z is obtained from the best predictor for known m by replacing $e_Z(m) - (\rho_Z,m)$ by its best unbiased estimate.

[1] T.W. Anderson, Introduction to Multivariate Statistical Analysis, New York, John Wiley, 1958.

[2] N. Aronszajn, "La théorie générale des noyaux réproduisants et ses applications," Proc. Cambridge Philos. Soc., Vol.39 (1944), pp.133-153.

[3] N. Aronszajn, "Theory of reproducing kernels," Trans. Amer. Math. Soc., Vol.68 (1950), pp.337-404.

[4] E.W. Barankin, "Locally best unbiased estimates," Ann. Math. Stat., Vol.20 (1949), pp.477-501.

[5] R.R. Bahadur, "Measurable subspaces and subalgebras," Proc. Amer. Math. Soc., Vol.6 (1955), pp.565-570.

[6] J.L. Doob, Stochastic Processes, New York, John Wiley, 1953.

[7] K. Karhunen, "Über lineare Methoden in der Wahrscheinlichkeitsrechnung," Ann. Acad. Sci. Fennicae, I. Math.-Physica, No.37 (1947) 79 pp.

[8] A.N. Kolmogorov, Foundations of the Theory of Probability, (English translation) New York, Chelsea, 1950.

[9] M. Loève, "Fonctions aléatoires du second ordre," Supplement to P.Lévy, Processus stochastiques et mouvement Brownien, Paris, Gauthier-Villars, 1948.

[10] M. Loève, Probability Theory, New York, Van Nostrand, 1955.

[11] C.R. Rao, Advanced Statistical Methods in Biometric Research, New York, John Wiley, 1952.

[12] C. Stein, "Unbiased estimates with minimum variance," Ann. Math. Stat., Vol.21 (1950), pp.406-415.

[13] A. Taylor, Functional Analysis, New York, John Wiley, 1958.

[14] A.C. Zaanen, Linear Analysis, Amsterdam, North Holland Publishing Co., 1956.