Z. Wahrscheinlichkeitstheorie verw. Gebiete
48, 233–252 (1979)

Zeitschrift für
Wahrscheinlichkeitstheorie
und verwandte Gebiete

Asymptotic Minimax Theorems for the Sample Distribution Function

P.W. Millar*

Statistics Department, University of California, Berkeley, CA 94720, USA

Summary. Fix a family $\mathscr{C}$ of continuous distributions on the line. Sufficient and (different) necessary conditions on $\mathscr{C}$ are given in order that the sample distribution function be an optimal estimator in the asymptotic minimax sense. The abstract results are illustrated by a variety of concrete families $\mathscr{C}$ that have arisen in the literature; some of these illustrations settle known, but previously unsolved, problems. Methods involve systematic consideration of statistical experiments whose parameter lies in a Hilbert space, and the theory of abstract Wiener spaces.

1. Introduction

In a pioneering paper, Dvoretsky-Kiefer-Wolfowitz ([6], 1956) proved that the sample distribution function is asymptotically minimax among the collection of all continuous distributions (see Sects. 2–5 below for the formal definition of this concept). This paper has stood for over 20 years as one of the pivotal achievements of nonparametric decision theory. Recently, Kiefer and Wolfowitz, motivated in part by questions arising in reliability theory, reopened the question and proved, among many other things, that the sample distribution is asymptotically minimax (a.m.) among the class of all concave distributions ([14], 1976). This is, of course, harder to prove than the original problem, since the collection of concave df's is much smaller than the collection of all df's.

This suggests a general problem: given a collection $\mathscr{C}$ of distributions of the line, when is the sample distribution a.m. among the class $\mathscr{C}$ (am $\mathscr{C}$). This paper gives geometric sufficient conditions and some (different) necessary ones on a collection $\mathscr{C}$ in order that the sample df be am $\mathscr{C}$. In Sect. 6, illustrations of the use of these results show that the sample df is am $\mathscr{C}$ if $\mathscr{C}$ is the class of df's with increasing densities, the class of df's with increasing failure rate (IFR), the df's with decreasing failure rate (DFR), and the df's star ordered with respect to a

* Supported by NSF grant MCS75-10376

0044-3719/79/0048/0233/$04.00

given distribution; this list of examples could easily be augmented. Apparently Kiefer and Wolfowitz had settled the IFR case some time ago, but, according to Professor Kiefer, the DFR case had eluded everyone for some time. Indeed, to do the DFR case, a slight change of point of view must be adopted – one must center the argument around a special distribution located well within the class in question.

The geometric criterion on $\mathscr{C}$ forcing the sample df to be am$\mathscr{C}$ can be described roughly as follows (complete description is in Sect. 5). Select a distribution $F_0 \in \mathscr{C}$. Distributions "close" to F_0 can be parametrized by points of a certain Hilbert space. There are several ways of doing this – see Sect. 4. If $\mathscr{C}$ contains some df F_0 that, relative to this parametrization, is a "cluster point" of elements of $\mathscr{C}$, then the sample df will be am$\mathscr{C}$. Conversely, the sample df can fail to be am$\mathscr{C}$ for a variety of reasons. Two of these are discussed in Sect. 7. The gist of the discussion is that if $\mathscr{C}$ is too small then the sample df cannot be am$\mathscr{C}$. Here "small" has to do with dimensionality; it explains, for many of the standard families indexed by R^k, and for families of symmetric densities (for example), why the sample df cannot be asymptotically minimax.

Acknowledgements. This paper owes a very great deal to the work and thought of L. LeCam. This is perhaps most obvious in the ubiquitous use here of his abstract asymptotic minimax theorem. It has long been part of the statistical folklore that much can be gained – particularly in nonparametrics – by considering, as is done in this paper, experiments indexed by Hilbert spaces. In problems of signal detection, Hilbert structure was explicit from the start. In [12], there are examples (implicit) of it in a nonparametric setting; the Hilbert structure is explicit, for example, in Beran's papers ([2, 3]) where it is applied to both nonparametric and robustness problems. But it has been LeCam more than any other, it seems, who by casting it in the framework of Gaussian shift models like those of R^k, has illuminated the possibilities. Finally, I thank J. Kiefer for extensive and enlightening correspondence.

2. Decision Theoretic Preliminaries

The purpose of this section is to introduce some decision theoretic notions and to state the basic asymptotic minimax theorem used throughout the paper.

Let Θ be an index set. For each $\theta \in \Theta$, let P_θ be a probability on a measure space $(S, \mathscr{S})$. The collection $E = (P_\theta, \theta \in \Theta)$ is called an experiment. A decision space will be a measure space $(D, \mathscr{D})$; in all cases treated in this paper, D is a separable metric space and $\mathscr{D}$ its Borel sets. A procedure b is a Markov kernel of $(S, \mathscr{S})/(D, \mathscr{D})$:

for each $x \in S$, $b(x, \cdot)$ is a probability on $(D, \mathscr{D})$
for each $A \in \mathscr{D}$, $b(\cdot, A)$ is $\mathscr{S}$-measurable.

For each $\theta \in \Theta$, let $l(\theta; \cdot)$ be a loss function defined on D – here $l(\theta; \cdot)$ will always be non-negative and lower semicontinuous. In the decision problem (E, D, l) the risk function $\rho \geqq 0$, defined for $\theta \in \Theta$ and procedures b, is given by

$$(2.1) \quad \rho(\theta, b) = \int_S \int_D l(\theta; y)\, b(x, dy)\, P_\theta(dx).$$

When there are several experiments under discussion, write $\rho(\theta, b; E)$ for $\rho(\theta, b)$; D and l will always be clear from context.

Unfortunately the proper general statement of the asymptotic minimax theorem requires the notion of 'generalized procedures' – indeed, the result is false otherwise. The difficulty can be traced to a lack of compactness in the collection of ordinary (Markov kernel) procedures, so this class will now be enlarged in the usual manner. Define V_0 to be the collection of all finite linear combinations of the form $\sum a_i \mu_i$, a_i real, where μ_i is, for each i, a finite signed measure absolutely continuous with respect to some $P_{\theta_i}(\theta_i \in \Theta)$; define V to be the Banach space obtained by closing V_0 under the variation norm. Let $C(D)$ be the Banach space of all continuous bounded real functions on D (supremum norm). Define a generalized procedure b to be a positive bilinear functional on $V \times C(D)$ such that $\|b(\mu, c)\| \leqq \|\mu\| \, \|c\|$ for $\mu \in V$, $c \in C(D)$ and such that $b(\mu, 1) = \|\mu\|$ if $\mu \geqq 0$. An ordinary procedure $b(x, dy)$ is a generalized procedure via $b(\mu, c) \equiv \iint c(y) \, b(x, dy) \, \mu(dx)$, $c \in C(D)$, $\mu \in V$. The collection of generalized procedures is then compact for the topology of pointwise convergence on $V \times C(D)$ (it is Alaoglu's theorem!) and it is a simple exercise to show that the procedures of Markov kernel type are dense in the collection of generalized procedures. If $l(\theta, \cdot)$ is a loss function, continuous and bounded on D for each θ, then the risk of a generalized procedure b can be defined by $b(P_\theta, l(\theta, \cdot))$ $(P_\theta \in V, l(\theta, \cdot) \in C(D))$; if $l(\theta, \cdot)$ is only lower semicontinuous, then the risk function of b can be defined by $\sup_c b(P_\theta, c)$ where the supremum is computed over all continuous bounded functions c with $c(\cdot) \leqq l(\theta, \cdot)$. This is all there is to the required extension; however, to preserve intuitive content, we shall use throughout the Markov kernel notation for procedures even though generalized procedures might really be the ones used.

Fix now a sequence of experiments $E^n = \{P_\theta^n, \theta \in \Theta\}$, P_θ^n defined on $(S_n, \mathscr{S}_n)$, say. Single out $\theta_0 \in \Theta$ for special attention and compute the Radon-Nikodym derivatives $dP_\theta^n / dP_{\theta_0}^n \equiv \xi_\theta^n$ of the part of P_θ^n that is absolutely continuous with respect to $P_{\theta_0}^n$. Let $E = \{P_\theta, \theta \in \Theta\}$ be another experiment, with P_θ absolutely continuous with respect to P_{θ_0}, and set $\xi_\theta = dP_\theta / dP_{\theta_0}$. The experiments E^n are said to converge weakly to E if, for each θ, the $P_{\theta_0}^n$ distribution of ξ_θ^n converges to the P_{θ_0} distribution of ξ_θ. This notion of convergence is quite adequate for the purposes of this paper, despite obvious defects; for a better definition, see LeCam [16].

The following basic theorem is due, in its present form, to Hajek [11] and LeCam [16]. Fix l, D as above.

Proposition 2.1 (Hajek-LeCam asymptotic minimax theorem). *Let $E^n = \{P_\theta^n, \theta \in \Theta\}$ be a sequence of experiments converging weakly to $E = \{P_\theta, \theta \in \Theta\}$. Then*

$$(2.2) \quad \lim_n \inf_b \sup_\theta \rho(\theta, b; E^n) \geqq \inf_b \sup_\theta \rho(\theta, b; E).$$

Early versions of this go back to the mid 1950's – see the survey [4]. Hajek's version assumed P_θ normal with mean vector $\theta \in R^k$; LeCam's is completely general. Our use essentially will be to replace R^k by R^∞ in Hajek's formulation. We will make great use, however, of the fact that Θ is quite arbitrary. The appendix (Sect. 8) contains a very simple proof of this proposition that avoids the delicate calculations of Hajek's and which does not draw on LeCam's theory of experiments either; that it is a simple consequence of LeCam's version [16] is easily seen.

Obviously, if the infimum on the left side of (2.2) is replaced by an infimum over kernel procedures, the result still holds; in many applications, such as those of this paper, the limit experiment is nice enough to allow one to dispense with the generalized procedures there too.

3. Gaussian Experiments

In this paper, all sequences of experiments studied converge weakly to a Gaussian experiment whose parameter set Θ is a subset of some Hilbert space. This section sets forth the basic facts about such experiments.

Let H be a separable Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $|\cdot|^2 = \langle\cdot,\cdot\rangle$. The standard normal distribution on H is a cylinder probability measure on H whose characteristic functional is

$$(3.1)\quad \varphi(z)=\exp\{-\tfrac{1}{2}|z|^2\}$$

for $z\in H^*$, the dual of H (see [8, 15, 24]). If H is infinite dimensional, it is well known that this cylinder measure is finitely additive but not countably additive. The theory of abstract Wiener spaces [9] provides a procedure for dealing with this inconvenience.

Briefly, one searches for a separable Banach space B and a map

$$\tau: H\to B$$

such that (i) $\tau(H)$ is dense in B (ii) τ is linear, bounded, one-to-one (iii) the image in B of the standard normal on H is countably additive. Basic theory asserts that such a τ must be obtainable [5] via a "measurable norm" on H, but the statistical applications are so simple that we can avoid this theory: the particular (τ, H, B) used throughout is given below. Nevertheless, derivations of various statements are facilitated by the general formulation.

Given (τ, H, B) as above, let P_0 be the image on B of the standard normal of H. Then for fixed $x\in B$, the translate

$$(3.2)\quad A\to P_0(A+x), \qquad A\in\text{Borel sets of } B$$

is a probability absolutely continuous with respect to P_0 iff $x=\tau h$ for some $h\in H$. Denote by P_h the measures on B given by

$$(3.3)\quad P_h(A)=P_0(A+\tau h).$$

In this case

$$(3.4)\quad (dP_h/dP_0)(x)=\exp\{L_h(x)-\tfrac{1}{2}|h|^2\}$$

where L_h, $h\in H$ is a linear process on B, indexed by H, and defined by

$$L_h=\langle z,\cdot\rangle_B \quad \text{if } h=\tau^* z$$

and extended to all of H by a limiting argument. Here $\langle\cdot,\cdot\rangle_B$ is the duality relation between B^* and B, $z\in B^*$, and τ^* is the adjoint of τ. See [7, 10, 15, 24] for further details. The experiments $\{P_h, h\in H\}$ are then the basic experiments for this paper.

The concrete examples of this set up used in this paper are variants of the following classical case. Take H to be the Hilbert space of real functions on the unit interval such that $\int_0^1 h(s)\,ds=0$ and $\int_0^1 h^2(s)\,ds<\infty$. Let B_0 be the Banach space of all continuous real functions x on $[0,1]$ such that $x(0)=x(1)=0$ (supremum norm). Define $\tau\colon H\to B_0$ by

$$(3.5)\quad \tau h(t)=\int_0^t h(s)\,ds.$$

Then, as is well known, P_0 is the distribution of $\{W^0(t), 0\leqq t\leqq 1\}$, the standard Brownian bridge (readers unfamiliar with this fact can easily check it by computing characteristic functionals; the ch.f. of P_0 is evidently $\varphi(\tau^* y)$, $y\in B^*$, φ as in (3.1)). Moreover, P_h is the distribution of $\left\{W^0(t)+\int_0^t h(s)\,ds\right\}$, and L_h has a representation as the stochastic integral $L_h=\int h\,dW$. Proceeding to the slightly greater generality required below, let F be a continuous df on the line. Define the Hilbert space $H(F)$ of real functions h by

$$(3.6)\quad H(F)=\{h\colon \int h^2(s)\,dF(s)<\infty,\ \int h(s)\,dF(s)=0\}.$$

Here one can take B to be the Banach space (supremum norm) of continuous functions x on the line such that $x(t)=x_0(F(t))$, t in the support of F, for some $x_0\in B_0$, B_0 as defined above for the classical case. If τ is taken to be the evident analogue of (3.5), it is then easy to see that P_0 is the distribution of the process $\{W^0(F(t))\}$, and that P_h, dP_h/dP_0 have descriptions analogous to those of the classical case.

Fix now a (τ, H, B) and the corresponding Gaussian experiment $\{P_h, h\in H\}$. All decision theoretic problems in this paper will be of the following sort. The decision space will be B itself. A non-negative function l on B is subconvex if $\{x\colon l(x)\leqq a\}$ is closed, convex and symmetric for every real number a. If $B=B_0$, then, for example, $l(x)=\sup_t |x(t)|=\|x\|$, $l(x)=\int |x(t)|^2\,dt$, $l(x)=I\{x\colon \|x\|>c\}$ and so forth are all subconvex. Evidently such a function is lower semi continuous; in fact, there is even a sequence of uniformly continuous subconvex functions l_n with $l_n\uparrow l$, a technical convenience in some proofs. In addition, for every c, $\min\{l,c\}$ is again subconvex. Loss functions on the parameter set H and the decision space will have the form

$$l(h,x)=l(x-\tau h).$$

In these circumstances, the following evaluation of the minimax value was pointed out to me by L. LeCam (private communication). In LeCam's version, integration theory is developed for finitely additive Gaussian measures on the Hilbert space H; l is defined on H. This approach is no doubt preferable for a general theory, but it involves a bit of work, and the following simpler variant suffices for the purposes of this paper. The appendix contains a proof for the convenience of the reader.

Proposition 3.1. *Let $\{P_h, h \in H\}$ be a Gaussian experiment. Then*

$$(3.7) \quad \inf_b \sup_{h \in H} \iint l(x - \tau h)\, b(y, dx)\, P_h(dy) = \int l(x)\, P_0(dx).$$

4. Experiments Indexed by a Hilbert Space

In this section are exhibited sequences of experiments converging weakly to the Gaussian experiments of Sect. 3.

Fix a distribution function F_0 on the line. There are several ways of parametrizing 'neighborhoods' of F_0, some more convenient than others according to the problem at hand. Here are two parametrizations to be used later; such parametrizations have been used in the literature before, e.g. by Beran [2] and LeCam [18].

For the first, assume F_0 has a density f_0 with respect to a probability μ on the line. If f is another probability density with respect to μ, then $f^{\frac{1}{2}}$ and $f_0^{\frac{1}{2}}$ are elements of $L^2(d\mu)$. The orthogonal decomposition of $f^{\frac{1}{2}}$ into a multiple of $f_0^{\frac{1}{2}}$ and an orthogonal piece suggests the following parametrization of such f: define $f(h; x)$ by

$$(4.1) \quad f^{\frac{1}{2}}(h; x) = (1 - \tfrac{1}{4}|h|^2)^{\frac{1}{2}} f_0^{\frac{1}{2}} + h/2$$

where $h \in H^*(F_0)$, the Hilbert space given by

$$(4.2) \quad H^*(F_0) = \{h \in L^2(d\mu) \colon h \text{ orthogonal to } f_0^{\frac{1}{2}}\}.$$

Notice that not all elements of $H^*(F_0)$ can be used in this parametrization ($|h|$ must be less than 2, for example). Let H' denote the subset of such h that work; one can then parametrize by all of H^* by the trivial device of (say) defining $f(h; x)$ for $h \in H^* - H'$ by $f(h; x) = f(\pi h; x)$ where π is a mapping of $H^* - H'$ to H' (e.g., a projection). We assume that such trivial completions of the parametrization have been carried out when they are clearly needed. Let H_n denote the collection of $h \in H^*(F_0)$ such that $f(hn^{-\frac{1}{2}}; x)$, defined via (4.1), is a probability density. Evidently, $H_{n+1} \supset H_n$, and $\bigcup H_n = H^*(F_0)$.

The parametrization (4.1) gives reasonably full neighborhoods of F_0. The second method, to which we now turn, is rather narrower (e.g., no measure with a component singular to F_0 can arise); its simplicity, however, makes it useful. For F_0, f_0 as in the previous parametrization, define

$$(4.3) \quad f(h; x) = (1 + h(x))\, f_0(x)$$

where $h \in H(F_0)$ (defined in (3.6)) is chosen so that (4.3) gives a true density. Again only part of $H(F_0)$ is used in the parametrization, but the definition can be extended to all of H as before if necessary. Analogues of H_n here satisfy $H_{n+1} \supset H_n$, $\bigcup H_n$ dense in $H(F_0)$.

Each of these parametrizations can be used to construct a convergent sequence of experiments. For example, let us consider $f(h; x)$ defined by (4.3). Define P_h^n to be the probability on R^n having density

$$(4.4)\quad \prod_{i=1}^{n} f(hn^{-\frac{1}{2}}; x_i).$$

Standard asymptotic methods (e.g., short Taylor expansion of $\log(1+u)$ and estimation of remainder) yield an asymptotic expansion for $\log dP_h^n/dP_0^n$:

$$(4.5)\quad \log \prod_{1}^{n} f(hn^{-\frac{1}{2}}; x_i)/f_0(x_i) - n^{-\frac{1}{2}} \sum_{i=1}^{n} h(x_i) + \tfrac{1}{2}|h|^2 \to 0(P_0^n).$$

It is then immediate that the experiments P_h^n, $h \in H(F_0)$ converge weakly to the Gaussian experiment indexed by $H(F_0)$.

Similarly, if $f(h; x)$ is given by (4.1), one can again define probabilities P_h^n by (4.4). Restricting attention only to h with support contained in that of f_0, one can again obtain an expansion for $\log dP_h^n/dP_0^n$:

$$(4.6)\quad \log \prod_{1}^{n} f(hn^{-\frac{1}{2}}; x_i)/f_0(x_i) - n^{-\frac{1}{2}} \sum_{i=1}^{n} h(x_i)/f_0^{\frac{1}{2}}(x_i) + \tfrac{1}{2}|h|^2 \to 0(P_0^n).$$

To see this one may, for example, approximate $\log f(hn^{-\frac{1}{2}}; x_i)/f_0(x_i)$ by $f^{\frac{1}{2}}(hn^{-\frac{1}{2}}; x_i)/f_0^{\frac{1}{2}}(x_i) - 1$ as in [20]. For the special h in $H^*(F_0)$ just introduced, the trivial reparametrization of replacing h by $h/f_0^{\frac{1}{2}}$ and μ by $f_0\, d\mu$ reveals this sequence of experiments to be essentially asymptotically equivalent to that of the preceding paragraph. For future convenience, denote by $H_0^*(F_0)$ the Hilbert subspace of $H^*(F_0)$ consisting of h with support in that of f_0.

5. Sufficient Conditions that the Sample Distribution be Asymptotically Minimax

Let $\mathscr{C}$ be some collection of continuous distribution functions on the line. Take n independent, identically distributed observations, and let $\hat{F}_n$ be the sample distribution function, constructed piecewise linear and continuous. If F is a df, let F^n be the probability on R^n defined by

$$(5.1)\quad F^n(dx) = \prod_{i=1}^{n} F(dx_i).$$

Fix a subconvex loss function l on the space of continuous functions. $\hat{F}_n$ is asymptotically minimax in $\mathscr{C}$ (am $\mathscr{C}$) if

$$(5.2)\quad \lim_{n\to\infty} \frac{\sup_{F\in\mathscr{C}} \int l(n^{\frac{1}{2}}(\hat{F}_n - F))\, F^n(dx)}{\inf_b \sup_{F\in\mathscr{C}} \iint l(n^{\frac{1}{2}}(y-F))\, b(x, dy)\, F^n(dx)} = 1.$$

Here the basic decision space is the entire collection of continuous df's. The main result of this section is a sufficient condition on a collection $\mathscr{C}$ that forces $\hat{F}_n$ to be am$\mathscr{C}$.

To state this sufficient condition, first recall the definition of the Hilbert space $H(F)$ from (3.5). Let F_0 be a df with density f_0 with respect to a probability μ. Define F_0 to be a *radial cluster point* of $\mathscr{C}$ if there are subsets H_n of $H(F_0)$ such that $H_n \subset H_{n+1}$ and

(5.3a) if $h \in H_n$, then $f(hk^{-\frac{1}{2}}; x) = (1 + h(x)\, k^{-\frac{1}{2}}) f_0(x)$ is a probability density in $\mathscr{C}$ for all $k \geqq n$,

(5.3b) $\bigcup\limits_n H_n$ is dense in $H(F_0)$.

Fix now $F_0 \in \mathscr{C}$, radial cluster point; let B be the Banach space of continuous functions associated with $H(F_0)$ as in Sect. 3. Let l be a real function, defined on all continuous functions, subconvex on B, and satisfying

(5.4) $\int l(n^{\frac{1}{2}}(\hat{F}_n - F))\, dF^n$ is distribution free over $\mathscr{C}$ and converges to $El(W^0) = El(W^0(F_0))$ as $n \to \infty$.

This condition is much more restrictive than necessary – it was given so that the coming proof can be quick and unencumbered. Loss functions l that satisfy the condition (5.4) are those which are nice functions of $\|x\|$ (i.e., ones that measure loss between F and its estimator in terms of the Kolmogorov distance); see Remark 5 for a formulation covering a wider variety of cases, and for which the following proposition still holds.

Proposition 5.1. *Let $\mathscr{C}$ be a collection of continuous distribution functions admitting F_0 as a radial cluster point. Then the sample distribution function is asymptotically minimax in $\mathscr{C}$.*

Examples of the use of this result are given in Sect. 6.

Remark 1. One can prove a variant of this using the parametrization (4.3) as well – this offers a slightly greater variety of h's. To carry out the proof one considers first $l_1 \leqq l$, l_1 uniformly continuous, and lets l_1 increase to l (the Eq. (5.6) here will be only an approximation but the uniform continuity of l_1 permits passage to the local experiments much as before).

Remark 2. One can write a variant of this result in which densities $f_0(1 + hn^{-\frac{1}{2}})$ are not in $\mathscr{C}$ but are 'sufficiently close' to ones of $\mathscr{C}$. Since none of the applications discussed in Sect. 6 require this, we omit it; see, however, Sect. 7 for a possible formulation in terms of Hellinger distances.

Remark 3. Simple variants of the argument here can often be used to prove other statistics are asymptotically minimax in a given class, even when the sample df is not. See Example 6e, Sect. 6, for an illustration, and Proposition 5.2 below.

Remark 4. The classical minimax theorem of Dvoretsky-Kiefer-Wolfowitz [6] is of course immediate from Proposition 5.1. Here $\mathscr{C}$ is all continuous df's, and F_0 is, e.g., the uniform distribution on [0, 1]; it is obvious that F_0 is a radial cluster point of $\mathscr{C}$: H_n can be taken to be $\{h \in H(F_0): \sup\limits_{0 \leqq x \leqq 1} |h(x)| \leqq n^{\frac{1}{2}}\}$.

Remark 5. To extend the result to a greater number of common situations, one can weaken the hypotheses (5.4) on the loss function by making them true only asymptotically. A particular case, for example, that does not satisfy (5.4) exactly is the situation where after n observations one uses a loss function $l_n(F,\varphi)$ which is a fixed function of the normalized Cramer-von Mises 'distance' $\int n[F(t)-\varphi(t)]^2\,dF(t)$. Here is a formulation to cover such situations. Let l_n be a loss function, used at stage n, and let F_0 be a radial clusterpoint. Assume

(5.4a) $l_n(F;\hat{F}_n)$, under F^n, is distribution free for $F\in\mathscr{C}$,

(5.4b) there is a subconvex function l, defined on the usual Banach space of continuous functions associated with $H(F_0)$ such that

$$\int l_n(F;\hat{F}_n)\,dF^n \to El(W^0(F_0)),$$

(5.4c) for the experiments (4.4) and h in finite subsets H_0 of the dense subset of $H(F_0)$ specified in Definition (5.3)

$$\lim_{n\to\infty}\inf_{b}\sup_{h\in H_0}\int l_n(F(hn^{-\frac{1}{2}};\cdot),\varphi)\,b(x,d\varphi)\,dF^n(hn^{-\frac{1}{2}};x)$$
$$\geqq \inf_{b}\sup_{h\in H_0}\int l(\varphi-\tau h)\,b(x,d\varphi)\,dP_h.$$

If $l_n(F,\varphi)$ is a nice function G of $n^{\frac{1}{2}}\sup_t|\varphi(t)-F(t)|=n^{\frac{1}{2}}\|\varphi-F\|$ then one takes $l(x)=G(\|x\|)$; if $l_n(F;\varphi)$ is a nice function G of $\int n[F(t)-\varphi(t)]^2\,dF(t)$, then in the situation above l would be $G(\int|x(t)|^2\,dF_0(t))$. With these assumptions the conclusion of Proposition 5.1 continues to hold (the definition of asymptotic minimax being adjusted to the present situation) and extends to loss functions of the Cramer-von Mises variety (for example); only minor modifications in the argument are needed to carry the proof through again.

Proof of Proposition 5.1. The numerator in (5.2) converges, by hypothesis, to $El(W^0)$. To prove the result, it suffices to show that, in the limit, the denominator is at least as big as $El(W^0)$. Let F_0 be a radial cluster point of $\mathscr{C}$. Fix n, and let $k\geqq n$; if F is any df, F^n is defined by (5.1). Then, the denominator of (5.2) is bigger than

$$\lim_{k\to\infty}\inf_{b}\sup_{h\in H_n}\iint l(k^{\frac{1}{2}}(y-F(hk^{-\frac{1}{2}};\cdot)))\,b(x,dy)\,F^k(hk^{-\frac{1}{2}};dx). \tag{5.5}$$

Since $F(hk^{-\frac{1}{2}};\cdot)$ is the integral of $f_0(1+hk^{-\frac{1}{2}})$, the argument of l in (5.5) is

$$k^{\frac{1}{2}}(y-F_0)+(F_0-F(hk^{-\frac{1}{2}};\cdot))\,k^{\frac{1}{2}}=k^{\frac{1}{2}}(y-F_0)-\tau h \tag{5.6}$$

so (5.5) is equal to

$$\lim_{k\to\infty}\inf_{b}\sup_{h\in H_n}\iint l(\varphi-\tau h)\,b(x,d\varphi)\,F^k(hk^{-\frac{1}{2}};dx) \tag{5.7}$$
$$\geqq \inf_{b}\sup_{h\in H_n}\iint l(\varphi-\tau h)\,b(x,d\varphi)\,P_h(dx)$$

by Proposition 2.1 and (4.4); here $\{P_h, h\in H(F_0)\}$ is the standard normal shift associated with the Hilbert space $H(F_0)$ as in Sect. 3. Let now $n\to\infty$; since H_n

increases to a dense subset of $H(F_0)$ and since P_h is continuous in h, it follows from (5.7) that the denominator of (5.2) exceeds

$$(5.8)\quad \begin{aligned} &\inf_{b}\sup_{h\in H(F_0)} \iint l(\varphi-\tau h)\, b(x,d\varphi)\, P_h(dx) \\ &\quad = \int l(x)\, P_0(dx) \qquad \text{(Prop. 3.1)} \\ &\quad = E\, l(W^0(F_0)). \end{aligned}$$

Since $E\,l(W^0(F))=E\,l(W^0)$ by assumption, the denominator of (5.2) is indeed at least $E\,l(W^0)$. QED

There is a variant of Proposition 5.1, occasionally useful, which provides asymptotically minimax procedures when the sample df fails.

Let $\mathscr{C}$ be a collection of distributions satisfying the following conditions:

(5.9a) there is $F_0\in\mathscr{C}$ with density f_0 and a subspace $H_0\subset H_0^*(F_0)$ such that every $F\in\mathscr{C}$ has density f that is given by

$$f^{\frac{1}{2}}=(1-|h|^2)^{\frac{1}{2}} f_0^{\frac{1}{2}}+h$$

for some $h\in H_0$ ($H_0^*(F)$ was defined below (4.6)),

(5.9b) if h belongs to a dense subset of H_0 then

$$(1-|h|^2 n^{-1})^{\frac{1}{2}} f_0^{\frac{1}{2}}+h n^{-\frac{1}{2}}$$

is a density of $\mathscr{C}$ for all sufficiently large n.

If $(\tau, H^*(F), B)$ is as in Sect. 3, let π be a continuous projection of B to the closure in B of τH_0, say $\overline{\tau H_0}$; for π to exist we must assume that $\overline{\tau H_0}$ is a complemented subspace. As usual, $\hat{F}_n$ is the sample df.

Proposition 5.2. *If $\mathscr{C}$ satisfies* (5.9) *then $F_0+\pi(\hat{F}_n-F_0)$ is an asymptotic minimax estimator of distributions in $\mathscr{C}$ and the asymptotic minimax value is $\int l(\pi x)\, P_0(dx)$.*

The proof is sufficiently similar to that of Proposition 5.1 that it can be omitted.

Remarks. (a) The Remarks 1–5 to Proposition 5.1 apply here too.

(b) The df estimator $F_0+\pi(F_n-F_0)$ need not be a df itself. The worried reader has several time-worn recourses available; the simplest is to let $\hat{G}_n$ be the smallest positive increasing function bigger than $F_0+\pi(\hat{F}_n-F_0)$ and take as estimator the df $\hat{G}_n\wedge 1$ – this will not increase the loss if the loss is a function of the Kolmogorov distance, and so the revised estimator will be asymptotically minimax too. Other recourses are available, depending on l; sometimes no adjustments are necessary, as in Example 6e.

6. Examples

In using Proposition 5.1, the strategy is to choose an F_0 which 'strictly' satisfies the conditions defining the class $\mathscr{C}$; it is then a matter of fussing (often with some tedium) to show that one can fit 'around' F_0 lots of other elements of $\mathscr{C}$.

6a. Decreasing Densities on $[0, \infty)$

Proposition 6.1. *Let $\mathscr{C}$ be the collection of all distributions F supported on $[0, \infty)$ and having a decreasing density with respect to Lebesgue measure. Then the sample df is am $\mathscr{C}$.*

Remark. This obviously implies that F_n is asymptotically minimax among the concave distributions on $[0, \infty)$, since the latter is a larger class than the one of Proposition 6.1. This case was treated by another method in [14], where much more is proved.

Proof. Take F_0 to be the probability with density $f_0(x) = e^{-x}$, $x \geqq 0$. It suffices to show $\mathscr{C}$ radially dense at F_0. Densities of the form $f_0(1 + hn^{-\frac{1}{2}})$ will belong to $\mathscr{C}$ if they are decreasing. This will be ensured, for example, if h has a derivative h' and the derivative of $(1 + hn^{-\frac{1}{2}}) f_0$ is negative:

$$f_0'(1 + hn^{-\frac{1}{2}}) + f_0 h' n^{-\frac{1}{2}} \leqq 0.$$

Of course, $(1 + hn^{-\frac{1}{2}}) f_0$ will be a density if $\int h(x) e^{-x} dx = 0$ and $\sup\limits_x |h(x) n^{-\frac{1}{2}}| \leqq \frac{1}{2}$. Accordingly if we define

$$H_n = \{h\colon \int h(x) e^{-x} dx = 0, \sup_x |h(x) n^{-\frac{1}{2}}| \leqq \tfrac{1}{2}, h'(x) n^{-\frac{1}{2}} \leqq \tfrac{1}{2}\}$$

then $(1 + hk^{-\frac{1}{2}}) f_0 \in \mathscr{C}$ for $h \in H_n$, $k \geqq n$ and $\bigcup H_n$ is dense in $H(F_0)$; that is, $\mathscr{C}$ admits F_0 as a radial cluster point.

6b. Increasing Failure Rate

A distribution F on the line with density f is said to have increasing failure rate (IFR) if $f(x)(1 - F(x))^{-1}$ is increasing in x for x in the support of f. Properties of such distributions and their applications to reliability theory are discussed, for example, in [1].

Proposition 6.2. *The sample distribution is asymptotically minimax among the class of IFR distributions.*

Proof. If f is any density with a derivative everywhere, then $f(x)/(1 - F(x))$ will be increasing if

$$(6.1) \quad [f'(1 - F) + f^2] \geqq 0 \qquad \text{(on support } F).$$

Take f_0 to be the uniform density on $[0, 1]$; then densities of the form $f_0(1 + hn^{-\frac{1}{2}})$ will be IFR if they satisfy (6.1) – i.e., if

$$(6.2) \quad [f_0'(1 + hn^{-\frac{1}{2}}) + f_0 h'](1 - F_0) + f_0^2 (1 + hn^{-\frac{1}{2}})^2 \geqq 0.$$

If

$$H_n = \left\{h\colon \int_0^1 h(x) dx = 0, \sup_{0 < x < 1} |h(x)| n^{-\frac{1}{2}} \leqq \tfrac{1}{2}, \sup_{0 < x < 1} |h'(x)| n^{-\frac{1}{2}} \leqq \tfrac{1}{4}\right\}$$

then (6.2) will hold for any $h \in H_n$. Since $\bigcup H_n$ is dense in $H(F_0)$, $F_0 = \text{uniform}$ distribution, it is clear that the IFR class admits F_0 as a radial cluster point.

6c. Decreasing Failure Rate

A distribution function F with density f has decreasing failure rate (DFR) if $f(x)(1-F(x))^{-1}$ is decreasing in x for x in the support of f; see [1] for properties and applications.

Proposition 6.3. *The sample df is asymptotically minimax among the class of distributions which have DFR.*

Proof. If F_0 is a distribution with density f_0, then any distribution with density $f_0(1+hn^{-\frac{1}{2}})$ will be DFR if it satisfies (6.2) with the inequality reversed. Choose $f_0(t) = \frac{1}{2}\exp(-t^{\frac{1}{2}})$, $t \geqq 0$ so $1-F_0(t) = (t^{\frac{1}{2}}+1)\exp(-t^{\frac{1}{2}})$. One may check that F_0 is a radial cluster point for the DFR distributions by choosing H_n to be (for example):

$$H_n = \Big\{ h\colon \int_0^\infty h(t)\exp(-t^{\frac{1}{2}})\,dt = 0,\ \text{support}\, h \subset [0, \log\log n],$$
$$\sup_x |h(x)|(\log n)^2 n^{-\frac{1}{2}} \leqq 1/8,\ \sup_x |h'(x)| \leqq n^{\frac{1}{2}}(\log\log n)^{-1}/16 \Big\}.$$

The verification is slightly tedious; again $\bigcup H_n$ is dense and so $\hat{F}_n$ is asymptotically minimax.

6d. Transformations of a Given Distribution

Fix a continuous distribution G on the line, and let $\mathscr{D}$ be a collection of monotone transformations g. A basic class $\mathscr{C}$ of distributions is of the form

$$\mathscr{C} = \{F : F = G \circ g \text{ for some } g \in \mathscr{D}\}.$$

Classes $\mathscr{D}$ that have arisen in the statistical literature are

$$\mathscr{D} = \{\text{convex functions}\} \quad \text{and} \quad \mathscr{D} = \{\text{star-shaped functions}\}$$

(see [1] for example).

In order to show that the sample distribution is asymptotically minimax for such $\mathscr{C}$, one seeks, for a fixed $g \in \mathscr{D}$, a dense collection of $h \in H(G \circ g)$ and functions $g_{n,h} \in \mathscr{D}$ such that

$$(6.3) \quad G \circ g_{n,h}(t) = \int^t (1 + h(s)n^{-\frac{1}{2}})\, dG \circ g(s).$$

Such a $g_{n,h}$ evidently must be given by

$$(6.4) \quad g_{n,h}(t) = G^{-1}\Big\{\int^t (1 + h(s)n^{-\frac{1}{2}})\, dG \circ g(s)\Big\}.$$

From (6.4) it is clear that if G is reasonably smooth and if $\mathscr{D}$ contains a fairly full neighborhood of g, then such $g_{n,h}$ will exist and so the sample df will be am $\mathscr{C}$. Evidently this problem is better attacked via Remark 2 of Sect. 5; until general statistical need arises, we shall not bother (it is pretty messy).

To illustrate a simple case, however, fix a G and let $\mathscr{D}$ be the collection of df's of the form $G \circ g$, where g is convex and increasing on the support of G. Assume G is twice differentiable, and G^{-1} continuous and differentiable. The function $g_{n,h}$ of (6.4) will belong to $\mathscr{D}$ (i.e., will be convex) if, for example, its second derivative is positive. If h is bounded, with bounded first derivative, and if g is chosen (for example) to have a second derivative that is bounded away from 0 (so take $g(t)=e^t-1$ if support G is $[0, \infty)$), then routine calculation shows that $g_{n,h}$ will have positive second derivative. Imposing the further condition that $(1+hn^{-\frac{1}{2}})\,dG \circ g$ be a probability, we obtain the required dense subset of $H(G \circ g)$. This shows that, for 'smooth' G, the sample df is asymptotically minimax among the collection of df's $G \circ g$, g convex; a fortiori, it is am among the df's obtained with g assumed "starshaped."

6e. Symmetric Density

Let $\mathscr{C}$ be the class of df's on $[0, 1]$ having a symmetric density f with respect to lebesgue measure: $f(x)=f(1-x)$, $0<x<1$. It turns out that the sample df is not asymptotically minimax here – see Sect. 7, Example 1. If f_0 is the uniform density on $[0, 1]$, then every density f in $\mathscr{C}$ has a representation

$$f^{\frac{1}{2}}=(1-|h|^2)^{\frac{1}{2}} f_0^{\frac{1}{2}}+h$$

with $h \in H^*(f_0)$, *h symmetric*. The natural parameter space here is stable under multiplication by small positive constants. Accordingly, the argument of Proposition 5.1 may be repeated (with the relevant Hilbert space being the symmetric functions of $H^*(F_0)$) to show that the symmetrized sample df, $\frac{1}{2}(\hat{F}_n(x)+1-\hat{F}_n(1-x))$, is am $\mathscr{C}$. Here the process that arises is not the Brownian bridge $W^0(t)$, but rather the process $\frac{1}{2}(W^0(t)-W^0(1-t))$. One could use Proposition 5.2 as well, π here being the projection $(\pi x)(t)=\frac{1}{2}(x(t)-x(1-t))$.

6f. Density Estimation

There is a vast literature on the problem of producing good estimators f_n of a density f. See [23] for a partial survey. If one has a good density estimator f_n, it might be tempting to use the integral of f_n as an estimator for the distribution function. Usually such 'good' density estimators are shown to exist under conditions specifying several derivatives of f, compact support, boundedness of $f'(x)/f^{\frac{1}{2}}(x)$ and so forth. However such classes of densities are so broad that they contain radial cluster points. Accordingly, within such a class, $\int^t f_n$ as a df estimator can be no better (in the asymptotic minimax sense) than the sample df.

Therefore, whether such estimators are actually as good (*even within the class in question*) remains to be proved, and cannot be taken for granted. The same caveat obviously applies to df estimators constructed from estimators of failure rates.

7. Necessary Conditions

According to the results of Sect. 5 and the examples of Sect. 6, very little appears to be required of a family of df's in order to force the sample df to be am within the given class: the class in question need 'only' contain a distinguished distribution F_0 that is a radial cluster point, and very few members of the class are needed to force F_0 into this role. This section is devoted to an assertion to the effect that, on the other hand, if $\mathscr{C}$ is 'too small', the sample df cannot be am $\mathscr{C}$; since we insist here on forcing the problem into a geometric mold, smallness will have something to do with dimensionality.

Consider a class of continuous df's $\mathscr{C}=\{F_\theta, \theta\in\Theta\}$; assume each F_θ is absolutely continuous with respect to some sigma finite measure μ. Then each F_θ has a density f_θ which can be expressed, for any fixed θ_0, by

(7.1) $f_\theta^{\frac{1}{2}}=(1-|h_\theta|^2)^{\frac{1}{2}} f_{\theta_0}^{\frac{1}{2}}+h_\theta$,

where $h_\theta\in H_0^*(F_{\theta_0})$ (defined after (4.6)). Essentially the result of this section says that if the closed linear span of the h_θ's that arise in this manner is a proper subspace of H_0^*, then the sample df cannot be asymptotically minimax in $\mathscr{C}$. Of course, if f_{θ_0} were a radial cluster point for $\mathscr{C}$, then the span just mentioned would be all of H^*. Replacing the exact class of h_θ's that arise in (7.1) by their span is very crude but the result nevertheless covers a number of standard situations – see the examples later on.

The formulation just given is sometimes inconvenient for asymptotic purposes, and so will be replaced by an asymptotic variant. Here are the necessary assumptions and definitions for the rest of this section.

(7.2a) Assume $\mathscr{C}=\{F(\theta;\cdot), \theta\in\Theta\}$ is a family of distributions indexed by a cone Θ: if $a>0$, then a $\Theta=\Theta$ (if Θ does not have this property, it is often very easy to reparametrize or to extend the parameter set so that it does).

(7.2b) Let F_0 be a continuous df with density f_0 with respect to some sigma finite measure μ; write $F(h; dx)$ for the df with density $f_0(h; x)$ given by (4.3). Assume that for each $\theta\in\Theta$ there is an $h_\theta\in H_0^*(F_0)$ such that

$$n^{\frac{1}{2}}\zeta(F(\theta n^{-\frac{1}{2}};\ ), F(h_\theta n^{-\frac{1}{2}};\ ))\to 0.$$

Here $\zeta(P,Q)$ is the Hellinger distance between probabilities P, Q:

$$\zeta^2(P,Q)=\int(\sqrt{dP}-\sqrt{dQ})^2.$$

The vectors h arising in (b) have, under the mapping τ of Sect. 3, images in a Banach space B of continuous functions. Let $sp\{h_\theta\}$ denote the linear span in $H_0^*(F_0)$ of the h_θ, and $sp\{\tau h_\theta\}$ the closure in B of the image of $sp\{h_\theta\}$ under τ.

Proposition 7.1. *If $sp\{\tau h_\theta\}$ is a complemented subspace of B, then the sample distribution is not asymptotically minimax in $\mathscr{C} = \{F_\theta, \theta \in \Theta\}$.*

Proof. Let P_h, $h \in H^*(F_0)$ be the usual Gaussian shift family on B. For any df F, define F^n by (5.1); let l be a uniformly continuous subconvex loss function. By arguments given in the proof of Proposition 5.1

$$(7.3)\quad \liminf_{n\to\infty} \inf_{b} \sup_{F\,\mathrm{cont}} \iint ((n^{\frac{1}{2}}(y-x))\, b(x,dy)\, F^n(dx)$$
$$= \liminf_{n\to\infty} \inf_{b} \sup_{h\in H_0^*(F_0)} \iint l(n^{\frac{1}{2}}(y-x))\, b(x,dy)\, F^n(hn^{-\frac{1}{2}}; dx)$$
$$= \int l(x)\, P_0(dx)$$

and this last number is achieved by the sample df. On the other hand, since Θ is a cone, similar considerations show

$$(7.4)\quad \liminf_{n\to\infty} \inf_{b} \sup_{\theta\in\Theta} \iint l(n^{\frac{1}{2}}(y-F_\theta)\, b(x,dy)\, F^n(\theta; dx)$$
$$= \liminf_{n\to\infty} \inf_{b} \sup_{\theta\in\Theta} \iint l(n^{\frac{1}{2}}(y-F(\theta n^{-\frac{1}{2}})))\, b(x,dy)\, F^n(\theta n^{-\frac{1}{2}}; dx)$$
$$\leq \liminf_{n\to\infty} \inf_{b} \sup_{h\in sp\{h_\theta\}} \iint l(n^{\frac{1}{2}}(y-F(hn^{-\frac{1}{2}};\cdot)))\, b(x,dy)\, F^n(hn^{-\frac{1}{2}}; dx)$$
$$= \inf_{b} \sup_{h\in sp\{h_\theta\}} \iint l(y-\tau h)\, b(x,dy)\, b(x,dy)\, P_h(dx).$$

Here condition (7.2b) forced the experiments $F^n(\theta n^{-\frac{1}{2}})$ and $F^n(h_\theta n^{-\frac{1}{2}})$ to be asymptotically equivalent; the uniform continuity of l, together with (7.2b), permitted the replacement of $F(\theta n^{-\frac{1}{2}})$ by $F(h_\theta n^{-\frac{1}{2}})$ in the argument of l. Let π be a projection of B to $sp\{\tau h_\theta\}$. Then

$$(7.5)\quad \inf_{b} \sup_{h\in sp\{h_\theta\}} \iint l(y-\tau h)\, b(x,dy)\, P_h(dx)$$
$$= \sup_{h\in sp\{h_\theta\}} \int (\pi x - \tau h)\, P_h(dx)$$
$$= \int l(\pi x)\, P_0(dx) \qquad \text{(see Proposition 5.2)}.$$

Accordingly, the sample df will fail to be asymptotically minimax for any l such that

$$(7.6)\quad \int l(\pi x)\, P_0(dx) < \int l(x)\, P_0(dx).$$

However, since $sp\{\tau h_\theta\}$ is a proper closed subspace, it has measure 0 under P_0. This can be deduced, for example, from the fact (e.g., [13]) that the support of a Gaussian measure on C is the closure of its reproducing kernel Hilbert space, plus standard results on perpendicularity of Gaussian measures; more direct proofs are also possible. Strict inequality (7.6) will then hold whener l satisfies

$$(7.7)\quad l(\pi x) < l(x), \qquad x \in B - sp\{\tau h_\theta\}.$$

Of course (7.6) is much weaker than (7.7) so other choices of l are possible too. The result extends beyond the uniformly continuous l as well. Indeed, assume that l is subconvex loss for which there exist uniformly continuous subconvex $l_n \geq l$, $l_n \downarrow l_0$, $\int l_n(x)\, dP_0 < \infty$ and $P_0(x\colon l(\pi x) = l_0(\pi x)) = 1$. Then it is easy to see that the result continues to hold for l; this permits one to bring in the usual discontinuous l such as $l(x)$=indicator of the set of $y \in B$ such that $\|y\|>c$. QED

Abstractly, the point here is nearly the same as realizing that the minimax risk available in a normal experiment on R^{k+1} is strictly bigger than that on R^k. Here are two examples of its use.

Example 1. *Symmetric densities.*

Let $\mathscr{C}$ be the distributions on $[0, 1]$ with symmetric densities f: $f(x)=f(1-x)$, $0<x<1$. Take f_0 to be the uniform density on $[0, 1]$. Every f in the class in question has a representation

$$(7.8) \quad f^{\frac{1}{2}}=(1-|h|^2)^{\frac{1}{2}} f_0^{\frac{1}{2}}+h.$$

The only h which arise, however, are symmetric h on $[0, 1]$. Proposition 7.1 then easily implies that the sample df is not am here. Of course this fact can also be proved be direct calculation: indeed, Sect. 6 showed that, for the class of symmetric densities, the symmetrized sample df is a.m. and it is easy to see that here it is better (strictly) than the sample df.

Example 2. *Families parametrized by R^k.*

Let $\{F_\theta, \theta \in \Theta\}$ be a family of continuous distributions where Θ is an open subset of R^k; assume $0 \in \Theta$ for convenience. Let F_θ have density $f(\theta;\ )$ with respect to a probability μ. Assume the quadratic mean derivative of $\theta \to f^{\frac{1}{2}}(\theta; x)$ exists at $\theta=0$ and is given by the vector $V(x) \in R^k$. (See [20], [11] regarding criteria for q.m. differentiation and discussion of its statistical relevance). It is then clear that

$$n^{\frac{1}{2}} \zeta(f(\theta n^{-\frac{1}{2}}; \cdot), g(h_\theta n^{-\frac{1}{2}}; \cdot)) \to 0$$

where $g(h_\theta n^{-\frac{1}{2}};\ )=(1-|h_\theta|^2 n^{-1})^{\frac{1}{2}} f_0^{\frac{1}{2}}+h_\theta n^{-\frac{1}{2}}$ with $h_\theta=\langle\theta, V\rangle$ (brackets denote inner product of R^k). Under further mild, well-known conditions

$$\log f(\theta n^{-\frac{1}{2}}; x_i)/f(0; x_i) - n^{-\frac{1}{2}} \sum_1^n \langle\theta, V(x_i)\rangle f_0^{-\frac{1}{2}}(x_i)+\tfrac{1}{2} E\langle\theta, V\rangle^2 \to 0.$$

The linear span $sp\{h_\theta\}$ is the span of the $\langle\theta, V(\cdot)\rangle$, $\theta \in \Theta$ and this is evidently finite dimensional. Proposition 7.1 then shows that the sample df cannot possibly be a.m. in the class $\{F_\theta, \theta \in \Theta\}$. The afficionado of parametric estimators no doubt can produce an alternative argument to show this; however, the point is that here the result follows simply, with no calculations, from the geometry of the parameter set.

Example 3. *Other finite dimensional problems.*

Let $\{F_\theta, \theta>0\}$, be the uniform distributions on $[0, \theta]$. Here the linear span of the h_θ's that arise in (7.1) is the entire Hilbert space, so Proposition 7.1 yields no information. Nevertheless, the sample df is still not a.m. One possible explanation runs as follows. Whenever $\{F_\theta, \theta \in \Theta\}$ is a family of df's indexed by a finite dimensional Θ, there are excellent estimators θ_n of θ; in fact, θ_n may be chosen so that $n^{\frac{1}{2}}\zeta(F(\theta_n;\ ), F(\theta;\ ))$ remains bounded in probability (see [21]). The loss functions l used in this paper may be chosen to measure the Kolmogorov distance between two measures. Accordingly, if the Hellinger distance between F_θ, F_0 is an order of magnitude greater than the Kolmogorov distance, then the estimate $F(\theta_n;\ )$ of $F(\theta;\ )$ should converge to F_θ (Kolmogorov sense) an order of magnitude faster than the sample distribution. In such circumstances, the sample df will clearly not be a.m. This is indeed what happens in the uniform $[0, \theta]$ case. In the standard asymptotic normal situation (cf., Example 2) the Hellinger and Kolmogorov distances are comparable, so it is not quite so trivial to dismiss the sample df.

8. Appendix

This section provides proofs of Propositions 2.1 and 3.1.
(a) *Proof of Proposition 2.1.*

Let l be a lsc loss function, bounded from below. Let (E, D, l) be a statistical decision problem as in Sect. 2. Let $\mathscr{M}(\Theta)$ be the collection of all probability measures on Θ supported by a finite number of points. Set

$$\rho(\mu; E) = \inf_b \int \rho(\theta, b; E)\,\mu(d\theta), \qquad \mu\in\mathscr{M}(\Theta).$$

Then by a standard minimax theorem (the fact that the collection of generalized procedures is compact, convex as well as the lower semicontinuity of the map $b\to\rho(\theta, b)$ is used at this point):

$$(8.1)\quad m \equiv \inf_b \sup_\theta \rho(\theta, b; E) = \sup_{\mu\in\mathscr{M}(\Theta)} \rho(\mu; E).$$

If E^n is a sequence of experiments converging weakly to E as in Sect. 2, then for each $\mu\in\mathscr{M}(\Theta)$

$$(8.2)\quad \varliminf_{n\to\infty} \rho(\mu, E^n) \geqq \rho(\mu, E).$$

Assuming (8.2), Proposition 2.1 is immediate since if $\varepsilon > 0$ and if $\mu_0\in\mathscr{M}(\Theta)$ comes within ε of achieving (8.1), then

$$\sup_{\mu\in\mathscr{M}(\Theta)} \rho(\mu, E^n) \geqq \rho(\mu_0, E^n) \geqq m - 2\varepsilon$$

for all sufficiently large n.

To establish (8.2), fix $\mu\in\mathscr{M}(\Theta)$, let Θ_0 be the support of μ and let $\mu_\theta = \mu\{\theta\}$. If $x\in R^{\Theta_0}$, let x_θ denote the θth co-ordinate of x. Define measures $Q = \sum_{\theta\in\Theta_0} P_\theta$ and similarly Q^n. The vector process $\{dP_\theta/dQ, \theta\in\Theta_0\}$ takes its values in the subset of R^{Θ_0} consisting of points x with $x_\theta \geqq 0$, $\sum x_\theta = 1$. Denote by Q_0 (resp. Q_0^n) the distribution of this process on R^{Θ_0}. Let $q = \liminf \rho(\mu_0, E^n)$; extract a subsequence (which we continue to denote by $\{n\}$) such that $\rho(\mu, E^n)\to q$. The hypothesis implies that dP_θ^n/dP_0^n converges weakly to dP_θ/dP_0 for each $\theta\in\Theta_0$;

extract a further subsequence (still to be denoted by $\{n\}$) so that the vector process $\{dP_\theta^n/dP_0^n,\ \theta\in\Theta_0\}$ converges in distribution. Then Q_0^n converges weakly to Q_0.

Moreover,

$$(8.3)\quad \begin{aligned}\rho(\mu, E) &= \inf_b \sum_\theta \mu_\theta \iint l(\theta, y)\, b(x, dy)\, dP\\ &= \inf_b \sum_\theta \mu_\theta \iint l(\theta, y)\, b(x, dy)\, x_\theta\, Q_0(dx)\\ &= \inf_b \iint \sum_\theta \mu_\theta\, l(\theta, y)\, x_\theta\, b(x, dy)\, Q_0(dx)\\ &= \int \inf_{y\in D} (\sum_\theta \mu_\theta\, l(\theta, y)\, x_\theta)\, Q_0(dx)\end{aligned}$$

and similarly for $\rho(\mu, E^n)$. The functions $x\to\sum_\theta \mu_\theta\, l(\theta, y)\, x_\theta$ are (for each y) linear on R^{Θ_0}, so their infimum over y is concave and continuous. If this concave function is denoted by h, then

$$(8.4)\quad \rho(\mu, E) = \int h(x)\, Q_0(dx) = \lim \int h(x)\, Q_0^n(dx) = \lim \rho(\mu, E^n),$$

proving the result. QED

(b) *Proof of Proposition 3.1.*

Let m denote the left side of (3.6). It is clear that $m \leqq \int l(x)\, P_0(dx)$, since the procedure $b(x, dy)$ = unit mass at $\{x\}$ produces the latter. To show equality, notice first that if l is subconvex on B, then there is a sequence of real functions l_k on R^k and a sequence $\{z_i\}$, $z_i\in B^*$, such that $l_k(\langle z_1, x\rangle, \ldots, \langle z_k, x\rangle)$, $x\in B$, is subconvex on B and increases to $l(x)$. This is easily proved by standard approximations, together with the fact that a closed convex set in B is a countable intersection of hyperplanes containing it. If τ^* is the adjoint of τ, then the z_i may be chosen, moreover, so that $e_i = \tau^* z_i$, $i \geqq 1$ forms an orthonormal basis of H. If k is fixed, evidently

$$(8.5)\quad m \geqq \inf_b \sup_h \iint l_k(x - \tau h)\, b(y, dx)\, P_h(dy).$$

If H_k is the subspace spanned by $e_1, \ldots, e_k$ and if A_k is the collection of procedures for the decision space τH_k, then by looking at (8.5) and l_k

$$(8.6)\quad m \geqq \inf_{b\in A_k} \sup_{h\in H_k} \iint l_k(x - \tau h)\, b(y, dx)\, P_h(dy).$$

However, each $h\in H_k$ has an expansion $h = \sum_1^k a_i e_i$, $(a_1, \ldots, a_k)\in R^k$, and for such h

$$(8.7)\quad dP_h/dP_0 = \exp\{\sum a_i L_{e_i}(x) - \sum a_i^2\}$$

where the L_{e_i}, described in Sect. 3, are i.i.d. $N(0, 1)$ random variables on B. From this it is immediate that the experiment $\{P_h, h\in H_k\}$, reparametrized in the obvious way by points of R^k, is equivalent to the standard normal shift experiment on R^k, $\{P_a^k, a \in R^k\}$. Here P_a^k is the normal distribution on R^k, with mean vector a, covariance the identity. Moreover, if $b \in A_k$, then such a b effectively chooses a point in R^k. Hence the quantity on the right in (8.6) is exactly equal to the minimax risk of the standard normal experiment on R^k, when the decision space is R^k itself and when the loss function is the subconvex function $l_k(a_1, \ldots, a_k)$ on R^k. But the minimax risk for the R^k problem just mentioned is

$$(8.8) \quad \int l_k(x)\, P_0^k(dx);$$

this can be established in the standard way by taking normal priors on R^k with covariance nI, using Anderson's famous lemma (Proc. Amer. Math. Soc. **6**, 170–76, 1956), and letting $n \to \infty$. However the quantity in (8.8) is equal to

$$(8.9) \quad \int l_k \left(\sum_1^k (\tau e_i)(x) \right) P_0(dx),$$

and so this, of course, is equal to the quantity on the right in (8.6). Since B is obtained via a measurable norm, if $k \uparrow \infty$, $\sum_1^k (\tau e_i)(x) \to x$ in P_0 probability. Using this, the lower semi continuity of l_k, and the fact that $l_k(\langle z_1, x \rangle, \ldots, \langle z_k, x \rangle) \uparrow l(x)$, one shows $m \geqq \int l(x)\, P_0(dx)$ by letting $k \to \infty$. QED

References

1. Barlow, R.E., Bartholomew, D.J., Bremmer, J.M., Brunk, H.D.: Statistical inference under order restrictions. New York: Wiley 1972
2. Beran, R.J.: Estimating a distribution function. Ann. Statist. **5**, 400–404 (1977)
3. Beran, R.J.: Rank spectral processes and tests for serial dependence. Ann. Math. Statist. **43**, 1749–1766 (1972)
4. Chernoff, H.: Large sample theory: parametric case. Ann. Math. Statist. **27**, 1–22 (1956)
5. Dudley, R.M., Feldman, J., LeCam, L.: On semi-norms and probabilities and abstract Wiener spaces. Ann. Math. **93**, 390–408 (1971)
6. Dvoretsky, A., Kiefer, J., Wolfowitz, J.: Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. Ann. Math. Statist. **27**, 642–669 (1956)
7. Feldman, J.: Equivalence and perpendicularity of Gaussian processes. Pacific J. Math. **8**, 699–708 (1958)
8. Gelfand, I.M., Vilenkin, N.Ya.: Generalized Functions, Vol. 4. New York: Academic Press 1964
9. Gross, L.: Abstract Wiener spaces, Proc. Fifth Berkeley Sympos. Math. Statist. Probab. Univ. Calif. **2**, 31–42 (1965)
10. Hajek, J.: A property of J-divergences of marginal probability distributions. Czechoslovak Math. J. **8**, 460–63 (1958)
11. Hajek, J.: Local asymptotic minimax and admissibility in estimation. Proc. sixth Berkeley Sympos. Math. Statist. Probab. Univ. Calif. **1**, 175–194 (1972)
12. Hajek, J., Sidak, Z.: Theory of Rank Tests. New York: Academic Press 1967
13. Kallianpur, G.: Abstract Wiener processes and their reproducing kernel Hilbert spaces. Z. Wahrscheinlichkeitstheorie verw. Gebiete **17**, 113–123 (1971)
14. Kiefer, J., Wolfowitz, J.: Asymptotically minimax estimation of concave and convex distribution functions. Z. Wahrscheinlichkeitstheorie verw. Gebiete **34**, 73–85 (1976)
15. Kuo, Hui-Hsiung: Gaussian Measures in Banach Spaces. New York: Springer 1975
16. LeCam, L.: Limits of experiments. Proc. Sixth Berkeley Sympos. Math. Statist. Probab. Univ. Calif. **1**, 245–61 (1972)

17. LeCam, L.: Asymptotic Decision Theory. Preprint
18. LeCam, L.: On some results of J. Hajek concerning asymptotic normality. Preprint
19. LeCam, L.: Sufficiency and asymptotic sufficiency. Ann. Math. Statist. **35**, 1419–1455 (1964)
20. LeCam, L.: Théorie Asymptotique de la Decision Statistique Les Presses de l'Universite de Montreal. Montreal, 1968
21. LeCam, L.: Convergence of estimates under dimensionality restrictions. Ann. Statist. **1**, 38–53 (1973)
22. Moussatat, M.: On the asymptotic theory of statistical experiments and some of its applications. Thesis, Univ. of Calif., Berkeley, 1976
23. Rosenblatt, M.: Curve estimates. Ann. Math. Statist. **42**, 1815–42 (1971)
24. Skorokhod, A.V.: Integration in Hilbert Space. K. Wickwire transl. New York: Springer 1974

Received September 15, 1978