ON SOME RESULTS OF J. HÁJEK
CONCERNING ASYMPTOTIC NORMALITY

By L. Le Cam

University of California, Berkeley

1. <u>Introduction</u>. This is a sequel to the papers [1] [2] [3] of J. Hájek and to the papers [4] [5] of the present author. Sometime before [4] appeared, but too late to make revisions, Hájek communicated to me two results which would have simplified and improved the statements of [4].
In the Summer of 1973, we debated the subject briefly. Hájek expressed the wish that someone should write it out properly. His premature disappearance may well delay realization of this wish for some length of time. In the hope that the present paper may help in the eventual fulfillment of Hájek's desire, I have attempted to summarize the situation, indicating in passing where additional research seems necessary.

For simplicity, the paper deals only with "local" asymptotic properties, as described in [1] [2] and [3]. Problems of a global nature, such as the existence of consistent estimates, will not be mentioned.

This research was partially supported by the U. S. Army Researc Office-Durham, Grant No. DA-ARO-D-31-124-73-G31 .

Within this restricted framework, the paper consists of two distinct parts. The first refers to the local admissibility results of [3], extending them to two dimensional parameter sets, and, to a certain extent, to higher dimensions. The second part refers to [4] and to the conditions under which, fpr independent identically distributed observations, the local asymptotic normality requirements of [3] are satisfied.

More specifically, the contents are as follows. Section 2 recalls some definitions and propositions from decision theory, following essentially [5]. The main results are two propositions which can be regarded as abstract versions of Hájek's Theorem 4.1 in [3]. The theorem in question consists of two parts. The first part, relating to minimax properties or analogous statements, is covered by our Proposition 1. The second part of Hájek's Theorem 4.1 says that all sequences which have the required optimal asymptotic behavior must be asymptotically equivalent in probability. A general version of this statement is the subject of our Theorem 1. It should be noted that the framework of Proposition 1 or Theorem 1 does not in any way refer to asymptotic normality. Thus, even though the arguments are abstracted from Hájek's proof, we bypass entirely the fine machinery represented by Hájek's Lemmas 3.1, 3.2 and 3.3 in [3].

Section 3 describes Gaussian shift experiments and asymptotically Gaussian experiments. The results of Section 2 are applied there to clarify the situation described by Hájek in [3].

Section 4 returns to the independent identically distributed case and improves the results of [4] according to remarks made by J. Hájek.

Section 5 deals with a definition of Fisher information and differentiability in quadratic mean. It refines some results of [3] and [4].

Section 2. Experiments and limits of experiments.

Let Θ be a set. One can consider that an experiment $\mathcal{E}$ indexed by Θ consists of a σ-field $\mathcal{A}$ and of a map $\theta \rightsquigarrow P_\theta$ from Θ to the space of probability measures on $\mathcal{A}$. The set $\mathcal{X}$ which carries the σ-field $\mathcal{A}$ will not play an essential role here. The **L**-space $L(\mathcal{E})$ of the experiment $\mathcal{E}$ is the smallest band which contains all the P_θ.

To specify a decision problem one needs an experiment $\mathcal{E}$, a space Z of possible decisions and a loss function. It will be assumed that a decision space consists of a set Z and of a uniform lattice Γ of bounded functions from Z to $(-\infty, +\infty)$. The set Γ is a uniform lattice if it is a vector lattice for the pointwise operations, contains the constant functions and is complete for the uniform norm.

The loss function W is a function from $\Theta \times Z$ to $(-\infty, \infty]$ subject to the restriction that, for each θ, one has $\inf_z \{W_\theta(z) \; ; \; z \in Z\} > -\infty$.

The space $\mathcal{D}(\mathcal{E},\Gamma)$ of decision procedures available on $\mathcal{E}$ for the decision space (Z,Γ) will be taken equal to the space of all transitions from $L(\mathcal{E})$ to the dual Γ^* of the uniform lattice Γ. Equivalently, a decision procedure ρ is a bilinear map from $\Gamma \times L(\mathcal{E})$ to $(-\infty, +\infty)$ such that, for all pairs $(\gamma,\mu) \in \Gamma \times L(\mathcal{E})$ one has $\gamma^+ \rho \, \mu^+ \geq 0$ and

$1\rho\,\mu^+ = ||\mu^+||$. This last symbol designates the ordinary (total variation) norm of $\mu \in L(\mathcal{E})$.

On $\mathcal{D}(\mathcal{E},\Gamma)$ we shall consider two topologies. The simple topology of convergence pointwise on $\Gamma \times L(\mathcal{E})$ and the topology of convergence in measure defined by the property that $\rho_\nu \to \rho$ in measure if $\langle|\gamma\rho_\nu - \gamma\rho|,\mu\rangle \to 0$ for each $\mu \geq 0$, $\mu \in L(\mathcal{E})$ and each $\gamma \in \Gamma$.

The nonrandomized procedures are the extreme points of $\mathcal{D}(\mathcal{E},\Gamma)$. It is easily verified that they are characterized by the property that they correspond to multiplicative maps of Γ into the dual $M(\mathcal{E})$ of $L(\mathcal{E})$. Equivalently, ρ is nonrandomized if

$$\gamma^2\rho - (\gamma\rho)^2 = 0 \qquad \text{for all} \quad \gamma \in \Gamma .$$

Suppose that W is a loss function on the decision space (Z,Γ). We shall <u>define</u> the risk $R(\theta,\rho)$ of a procedure $\rho \in \mathcal{D}(\mathcal{E},\Gamma)$ at the point θ by the relation

$$R(\theta,\rho) = \sup_{\gamma}\{\gamma\rho\,P_\theta\ ;\ \gamma \in \Gamma\ ,\ \gamma \leq W_\theta\} .$$

With this definition, the function $\rho \rightsquigarrow R(\theta,\rho)$ is always lower semicontinuous on the space $\mathcal{D}(\mathcal{E},\Gamma)$ topologized by pointwise convergence on $\Gamma \times L(\mathcal{E})$. For this same topology $\mathcal{D}(\mathcal{E},\Gamma)$ is a compact Hausdorff space.

Let $\mathcal{E}$ be an experiment and let (Z,Γ,W) be a decision space with a loss function W. We shall denote $\mathcal{R}(\mathcal{E},\Gamma,W)$

the space of all functions f from Θ to $(-\infty,+\infty]$ which are such that there is a $\rho \in \mathcal{D}(\mathcal{E},\Gamma)$ satisfying $f(\theta) \geq R(\theta,\rho)$ for all $\theta \in \Theta$. If m is a probability measure with finite support on Θ, let $\chi(\mathcal{E},\Gamma,W,m) = \inf_f \{ \int f(\theta)\, m(d\theta)\ ,\ f \in \mathcal{R}(\mathcal{E},\Gamma,W) \}$. The minimax theorem says that a function g from Θ to $(-\infty, +\infty]$ belongs $\mathcal{R}(\mathcal{E},\Gamma,W)$ if and only if

$$\int g\, dm \geq \chi(\mathcal{E},\Gamma,W,m)$$

for all probability measures in which have finite support on Θ.

Consider now two experiments $\mathcal{E} = \{P_\theta\ ;\ \theta \in \Theta\}$ and $\mathfrak{F} = \{Q_\theta\ ;\ \theta \in \Theta\}$ indexed by the same set Θ but corresponding to different σ-fields $\mathcal{A}$ and $\mathcal{B}$ respectively. The <u>deficiency</u> $\delta(\mathcal{E},\mathfrak{F})$ is the smallest $\varepsilon \in [0,1]$ for which there is a transition T from $L(\mathcal{E})$ to $L(\mathfrak{F})$ such that

$$\sup_\theta \tfrac{1}{2} \|T P_\theta - Q_\theta\| \leq \varepsilon \ .$$

Equivalently, $\delta(\mathcal{E},\mathfrak{F}) \leq \varepsilon$ if for all decision spaces (Z,Γ) and all loss functions W such that $0 \leq W \leq 1$, for every $g \in \mathcal{R}(\mathfrak{F},\Gamma,W)$ there is an $f \in \mathcal{R}(\mathcal{E},\Gamma,W)$ such that $f \leq g + \varepsilon$. For the purposes of this kind of definition one can restrict the triplets (Z,Γ,W) drastically without modifying the resulting number $\delta(\mathcal{E},\mathfrak{F})$. See [6] for instance.

The "distance" between $\mathcal{E}$ and $\mathfrak{F}$ will be taken equal to

$$\Delta(\mathcal{E},\mathfrak{F}) = \max\{\delta(\mathcal{E},\mathfrak{F}),\ \delta(\mathfrak{F},\mathcal{E})\}\ .$$

This is only a pseudometric. If one says that $\mathcal{E}$ and $\mathfrak{F}$ are <u>equivalent</u>, or are of the same <u>type</u>, when $\Delta(\mathcal{E},\mathfrak{F}) = 0$, the set $\mathbb{E}(\Theta)$ of experiment types becomes a complete metric space for Δ.

Suppose that $\mathcal{E} = \{P_\theta\ ;\ \theta \in \Theta\}$ is an experiment indexed by Θ and that A is a subset of Θ. Denote $\mathcal{E}(A)$ the restriction $\mathcal{E}(A) = \{P_\theta\ ;\ \theta \in A\}$ of $\mathcal{E}$ to A. The weak topology of $\mathbb{E}(\Theta)$ is the weakest which renders continuous the restriction map $\mathcal{E} \rightsquigarrow \mathcal{E}(A)$ from $\mathbb{E}(\Theta)$ to $\mathbb{E}(A)$ for each finite A. For this topology $\mathbb{E}(\Theta)$ is compact.

Assuming Θ, Z and Γ fixed, let W and V be two loss functions on (Z,Γ). The inequality $V \leq W$ will mean that $V_\theta(z) \leq W_\theta(z)$ for all pairs $(\theta,z) \in \Theta \times Z$.

A loss function V will be called <u>special</u> if for each θ the function $z \rightsquigarrow V_\theta(z)$ is an element of Γ.

The following proposition strengthens the statement of Proposition 5 in [5]. Since the argument in [5] is inadequate, we give a complete proof. As mentioned in the Introduction, the intent of this Proposition is essentially the same as that of the first half of Theorem 4.1 in [3].

<u>Proposition</u> 1. <u>Let</u> $\mathfrak{F}$ <u>be an experiment indexed by</u> Θ. <u>Let</u> (Z,Γ,W) <u>be a decision space with a loss function</u> W. <u>Let</u> f

be a function from Θ *to* $(-\infty, +\infty]$ *which does not belong to* $\mathfrak{R}(\mathfrak{F}, \Gamma, W)$.

Then, there is a special loss function V *such that* $V \leq W$, *a probability measure* m *with finite support on* Θ, *a number* $\alpha > 0$ *and a weak neighborhood* U *of* $\mathfrak{F}$ *in* $\mathbb{E}(\Theta)$ *such that*

$$\int (f+\alpha)\,dm < \chi(\mathcal{E}, \Gamma, V, m)$$

for all $\mathcal{E} \in U$.

Proof. Let $\mathfrak{F} = \{Q_\theta\ ;\ \theta \in \Theta\}$. Since f does not belong to $\mathfrak{R}(\mathfrak{F}, \Gamma, W)$, there is an m with a finite support S such that $\int f\,dm < \chi(\mathfrak{F}, \Gamma, W, m)$. Take two numbers a and b such that

$$\int f\,dm < a < b < \chi(\mathfrak{F}, \Gamma, W, m)\ .$$

Let $\{\gamma_\theta\ ;\ \theta \in S\}$ be any set of elements of Γ such that $\gamma_\theta \leq W_\theta$ for $\theta \in S$. The set of procedures $\rho \in \mathcal{D}(\mathfrak{F}, \Gamma)$ such that $\int (\gamma_\theta \rho Q_\theta)\,dm > b$ is open for the pointwise topology of $\mathcal{D}(\mathfrak{F}, \Gamma)$. Since $R(\theta, \sigma) = \sup_\gamma \{\gamma\,\sigma\,Q_\theta\ ;\ \gamma \in \Gamma\ ,\ \gamma \leq W_\theta\}$ the open sets in question cover the entire compact $\mathcal{D}(\mathfrak{F}, \Gamma)$. Thus one can extract a finite subcover, yielding a finite family $\{\gamma_{\theta,i}\ ;\ i \in I\}$ such that if $\sigma \in \mathcal{D}(\mathfrak{F}, \Gamma)$ then $\int (\gamma_{\theta,i}\ \sigma\ Q_\theta)\,dm > b$ for at least one $i \in I$. Let V_θ be $V_\theta = \sup_i \{\gamma_{\theta,i}\ ;\ i \in I\}$ if $\theta \in S$. For values $\theta \notin S$ take for V_θ any element of Γ such that $V_\theta \leq W_\theta$. Then $V \leq W$.

Furthermore V is special and such that $\int (V_\theta \, \sigma \, Q_\theta) dm > b$ for all σ.

Let $K = \sup\{|V_\theta(z)| \; ; \; \theta \in S , z \in Z\}$ and let U be the subset of $\mathbb{E}(\Theta)$ formed by those $\mathcal{E}$ whose restriction to S satisfy the inequality

$$\Delta[\mathcal{E}(S) \, , \, \mathfrak{F}(S)] \le \frac{b-a}{2K} .$$

This set is a weak neighborhood of $\mathfrak{F}$ in $\mathbb{E}(\Theta)$. For each $\mathcal{E} = \{P_\theta \; ; \; \theta \in \Theta\}$, $\mathcal{E} \in U$ one has

$$\int (V_\theta \, \rho \, P_\theta) dm > \chi[\mathfrak{F}, \Gamma, V, m] - \frac{(b-a)}{2} \ge \frac{a+b}{2}$$

for all $\rho \in \mathcal{D}(\mathcal{E},\Gamma)$. Therefore, the function $f + (\frac{b-a}{2})$ does not belong to $\mathcal{R}(\mathcal{E},\Gamma,V)$. This completes the proof of the Proposition.

It was shown in [5] that certain results of [3] can be obtained by application of this Proposition. However the second statement of Theorem 4.1 in [3] does not follow from this. To obtain it we shall first prove the following abstract version of Hájek's argument.

Consider a given system (Θ, Z, Γ, W) and particular experiment $\mathfrak{F} = \{Q_\theta \; ; \; \theta \in \Theta\}$.

Theorem 1. *Let* f *be an admissible element of* $\mathcal{R}(\mathfrak{F},\Gamma,W)$. *Assume that there is only one procedure* $\sigma \in \mathcal{D}(\mathfrak{F},\Gamma)$ *such that* $R(\theta,\sigma) \equiv f(\theta)$ *and that this* σ *is nonrandomized*.

Assume that $\mathcal{E}_\nu \to \mathfrak{F}$ weakly and let $\rho_{\nu,i'}$, $i = 1 , 2$ be two elements of $\mathcal{D}(\mathcal{E}_\nu,\Gamma)$. Suppose that for all special loss functions $V \leq W$ and all θ one has

$$\lim_\nu \sup V_\theta \, \rho_{\nu,i} \, P_{\theta,\nu} \leq f(\theta)$$

and this for $i = 1 , 2$.

Then for every $\gamma \in \Gamma$ and $\theta \in \Theta$ one has

$$\lim_\nu |\gamma \rho_{\nu,2} - \gamma \rho_{\nu,1}| P_{\theta,\nu} = 0 .$$

Proof. For each ν and each finite subset S of Θ , let $\alpha(\nu,S)$ be the deficiency $\delta[\mathfrak{F}(S),\mathcal{E}_\nu(S)]$ of the experiments restricted to S. There are transitions $T_{\nu,S}$ from $L(\mathfrak{F})$ to $L(\mathcal{E}_\nu)$ such that

$$\tfrac{1}{2}||T_{\nu,S} \, Q_\theta - P_{\theta,\nu}|| \leq \alpha(\nu,S)$$

for all $\theta \in S$. Let $\sigma(\nu,S,i)$ be the element $\sigma(\nu,S,i) = \rho_{\nu,i} \, T_{\nu,S}$ of $\mathcal{D}(\mathfrak{F},\Gamma)$.

~~In the remainder of this proof, limits will be taken as both ν and S increase. For the $\sigma(\nu,S,i)$ this means that we consider limits along the filter whose base are the sets~~

~~$$\{\sigma(\nu',S',i) \; ; \; \nu' \geq \nu \; , \; S' \supset S \; , \; S' \text{ finite} \} .$$~~

~~Let σ_i be a cluster point of this filter.~~ Let V be a special loss function smaller than W and let $||V||_S$ be

$$||V||_S = \sup\{|V_\Theta(z)| \ , \ \Theta \in S \ , \ z \in Z\} .$$

Then

$$V_\Theta \, \rho_{\nu,i} \, T_{\nu,S} \, Q_\Theta \preceq V_\Theta \, \rho_{\nu,i} \, P_{\Theta,\nu} + ||V||_S \, \alpha(\nu,S) .$$

Take limits along ordered set (ν,S,V,ℓ) with $(\nu',S',V',\ell') \succ (\nu,S,V,\ell)$ iff $\nu' \geq \nu$, $S' \supset S$, $V' \geq V$, $\ell' \geq \ell$ in the space where $V \leq W$, $||V||_S \, \alpha(\nu,S) < \frac{1}{\ell}$

It follows that

$$\limsup V_\Theta \, \sigma(\nu,S,i) Q_\Theta \preceq f(\Theta)$$

and consequently $V_\Theta \sigma_i Q_\Theta \preceq f(\Theta)$. Since σ was admissible and unique this implies $\sigma_i = \sigma$. Equivalently, $\sigma(\nu,S,i) \to \sigma$ pointwise on $\Gamma \times L(\mathfrak{V})$.

One can also introduce a procedure $\rho_{\nu,3} = \frac{1}{2}[\rho_{\nu,1} + \rho_{\nu,2}]$ and the corresponding $\sigma(\nu,S,3)$. For the same reason $\sigma(\nu,S,3) \to \sigma$.

It has been assumed that σ is nonrandomized. Thus, for every $\gamma \in \Gamma$ one has $\gamma^2\sigma = (\gamma\sigma)^2$ in the dual $M(\mathfrak{V})$ of $L(\mathfrak{V})$. Note that $\gamma^2 \, \sigma(\nu,S,i) - [\gamma \, \sigma(\nu,S,i)]^2 \geq 0$. This yields

$$\begin{aligned} 0 &\leq [\gamma\sigma(\nu,S,i) - \gamma\sigma]^2 \\ &\leq \gamma^2 \sigma(\nu,S,i) + \gamma^2 \sigma - 2[\gamma\sigma(\nu,S,i)]\gamma\sigma \ . \end{aligned}$$

Since $\gamma^2\sigma(\nu,S,i) \to \gamma^2\sigma$ and $\gamma \, \sigma(\nu,S,i) \to \gamma \, \sigma$ for the weak topology $w[M(\mathfrak{V}) \ , \ L(\mathfrak{V})]$ one concludes that the right side of the above inequality tends to zero for this same weak topology. However, since it is nonnegative, it must then <u>tend to zero in measure</u>. It follows that both $\gamma^2(\nu,S,i)$

and $[\gamma\,\sigma(\nu,S,i)]^2$ converge in measure to the same limit $\gamma^2\sigma = (\gamma\sigma)^2$.

One can also write

$$\gamma^2\,\sigma(\nu,S,i) - [\gamma\,\sigma(\nu,S,i)]^2$$
$$= \{\gamma^2\,\rho_{\nu,i}\,T_{\nu,S} - (\gamma\,\rho_{\nu,i})^2\,T_{\nu,S}\}$$
$$+ \ \{(\gamma\,\rho_{\nu,i})^2\,T_{\nu,S} - [\gamma\,\rho_{\nu,i}\,T_{\nu,S}]^2\}\ .$$

In this identity both brackets on the right are nonnegative. Thus both of them must tend to zero in measure, and thus for $i = 1, 2, 3$.

Let φ_ν be the difference $\varphi_\nu = \frac{1}{2}[\rho_{\nu,2} - \rho_{\nu,1}]$, and consider the identity

$$[\gamma\,\rho_{\nu,3}]^2 + [\gamma\,\varphi_\nu]^2 = \frac{1}{2}\,[\gamma\,\rho_{\nu,1}]^2 + [\gamma\,\rho_{\nu,2}]^2\}\ .$$

Apply the transformation $T_{\nu,S}$ to both sides of the identity and pass to the limit. Since all the $[\gamma\,\rho_{\nu,i}]^2\,T_{\nu,S}$ converge in measure to $\gamma^2\sigma$ the remaining term $\gamma\varphi_\nu{}^2$ tends to zero. Explicitely $[\gamma\,\varphi_\nu]^2\,T_{\nu,S}\,Q_\theta \rightarrow 0$. However

$$[\gamma\,\varphi_\nu]^2\,P_{\theta,\nu} \ \leq\ [\gamma\,\varphi_\nu]^2\,T_{\nu,S}\,Q_\theta + ||\gamma||\ \alpha(\nu,S)\ .$$

Hence $[\gamma\,\varphi_\nu]^2\,P_{\theta,\nu}$ and a fortiori $|\gamma\,\varphi_\nu|P_{\theta,\nu}$ must tend to zero for each $\theta \in \Theta$. This completes the proof of the theorem.

As can be seen, the argument relies heavily on the fact that the limit σ is nonrandomized. We do not know what can occur when this condition is not met.

It should be obvious that Proposition 1 and Theorem 1 could be applied to situations which are remote from the usual Gaussian limits. However, because of its special importance we shall now illustrate their application to the Gaussian case, limiting ourselves to a strictly minimal elaboration.

Section 3. Gaussian shift limits.

Consider the k-dimensional vector space R^k, and an experiment $\mathcal{G} = \{G_t ; t \in R^k\}$ indexed by R^k. Such an experiment will be called homogeneous if the measures which constitute it are mutually absolutely continuous.

One can say that $\mathcal{G}$ is a Gaussian shift experiment linearly indexed by R^k if it is homogeneous and if the logarithms of likelihood ratios $\Lambda(t) = \log (dG_t/dG_0)$ have the form $\Lambda(t) = tS - \frac{1}{2} tKt'$ for some statistic S and some matrix K.

There are, of course, other kinds of experiments in which the observations have Gaussian distributions. In the present formula, the covariance system does not depend on the parameters and the expectations are linear functions of $t \in R^k$. These are the only Gaussian experiments considered here.

In the formula $\Lambda(t) = tS - \frac{1}{2}tKt'$, if $tKt' = 0$ the measures G_0 and G_t are the same. One could then reduce the dimension of the indexing space R^k. Also, when K is nonsingular, a simple change of variables will reduce K to the identity matrix.

Taking this into account it will be convenient to define a standard Gaussian shift experiment as follows.

It will be assumed that R^k carries a Euclidean norm, denoted $|\cdot|$. With the corresponding notation for inner product and transpose, the standard Gaussian shift experiment of $(R^k, |\cdot|)$ is an homogeneous experiment $\mathscr{G} = \{G_t ; t \in R^k\}$ such that

$$\log \frac{d\, G_t}{d\, G_0} = t\, S - \tfrac{1}{2}|t|^2$$

for a certain statistic S which takes values in the dual $(R^k)'$ of R^k.

It is often necessary to consider restrictions of $\mathscr{G}$ to subsets of R^k. Here, we shall only consider restrictions to subsets $C \subset R^k$ which are convex cones, with a nonempty interior and with a vertex at the origin of R^k.

For such a cone one can say that $\{G_t\ , t \in C\}$ is the standard Gaussian shift experiment restricted to C if it is homogeneous and if the equivalence classes

$$L(t) = \log \left(\frac{d\, G_t}{d\, G_0} \right) + \tfrac{1}{2}|t|^2$$

are such that

$$L(s+t) = L(s) + L(t)$$

for all pairs $(s,t) \in C \times C$.

It is true that the experiment $\mathscr{G}$ is not entirely specified by the above description. However, its <u>type</u> is entirely determined. One version of $\mathcal{G}$ is given by the Gaussian measures

$$G_t = n(t,I) \text{ on } R^k .$$

Many statistical problems lead to the following situation. One has a sequence , or net $\{\mathcal{E}_\nu\}$ of experiments $\mathcal{E}_\nu = \{P_{\theta,\nu} ; \alpha \in A_\nu\}$ indexed by sets A_ν . For each ν , there is a map $\alpha \rightsquigarrow \xi_\nu(\alpha)$ of A_ν into the Euclidean space $(R^k , 1.1)$ and a distinguished point $\alpha_{o,\nu} \in A_\nu$. For simplicity of notation we shall write o instead of $\alpha_{o,\nu}$ and assume that $\xi_\nu(\alpha_{o,\nu}) = 0$ in R^k.

In such a case one çan define sets $A_\nu(c)$ by the relation

$$A_\nu(c) = \{\alpha ; \alpha \in A_\nu , |\xi_\nu(\alpha)| \leq c \} ,$$

and let $\mathcal{E}_{\nu,c}$ be the experiment $\mathcal{E}_{\nu,c} = \{P_{\alpha,\nu} ; \alpha \in A_\nu(c)\}$. One can also construct experiments $\mathcal{G}^*_{\nu,c} = \{Q_{\alpha,\nu} ; \alpha \in A_\nu(c)\}$ in which $Q_{\alpha,\nu}$ is the Gaussian measure G_t , $t = \xi_\nu(\alpha)$ of the standard Gaussian shift experiment of $(R^k , |\cdot|)$.

Definition. The net $\{\mathcal{E}_\nu\}$ satisfies condition (G) for the maps $\{\xi_\nu\}$ if for every $c \in (0,\infty)$ the distances $\Delta(\mathcal{E}_{\nu,c} , \mathcal{G}^*_{\nu,c})$ tend to zero.

The image $\xi_\nu(A_\nu)$ of A_ν is some subset B_ν of R^k . It may vary rather arbitrarily. The following assumption is intended to insure that B_ν stabilizes itself and is not too small.

Assumption 1. *There is a convex cone* C *with vertex at the origin of* R^k *and with a nonempty interior satisfying the following requirement: Let* $U(c)$ *be the ball of radius* c *centered at the origin of* R^k. *Then the Hausdorff distance between* $B_\nu \cap U(c)$ *and* $C \cap U(c)$ *tends to zero as* ν *increases.*

When Assumption 1 is satisfied the asymptotic normality requirement (G) can be replaced by a variety of equivalent statements. Two possibilities are as follows.

Denote $\Lambda_\nu(\alpha)$ the logarithm of the density $\frac{d\,P_{\alpha,\nu}}{d\,P_{o,\nu}}$ of that part of $P_{\alpha,\nu}$ which is dominated by $P_{o,\nu}$.

If S_ν is a Euclidean valued statistic available on $\mathcal{E}_\nu$, let $\mathcal{L}[S_\nu | P_{\alpha,\nu}]$ be its distribution for the measure $P_{\alpha,\nu}$.

Requirement A. *There are statistics* S_ν *available on* $\mathcal{E}_\nu$, *taking values in* R^k *and such that when* $|\xi_\nu(\alpha_\nu)|$ *stays bounded, the difference*

$$\mathcal{L}(S_\nu \mid P_{\alpha_\nu,\nu}) - \mathcal{N}(\xi_\nu(\alpha_\nu), I)$$

tends to zero vaguely. Also, for bounded $\{\xi_\nu(\alpha_\nu)\}$, *the* *deficiency of the binary experiment* $\{\mathfrak{N}(0,I), \mathfrak{N}(\xi_\nu(\alpha_\nu), I)\}$ *with respect to* $(P_{0,\nu}, P_{\alpha_\nu,\nu})$ *tends to zero.*

Requirement B. *If* $\{\xi_\nu(\alpha_\nu)\}$ *stays bounded then* $\{P_{0,\nu}\}$ *and* $\{P_{\alpha_\nu,\nu}\}$ *are contiguous. Furthermore, letting* $L_\nu(\alpha) = \Lambda_\nu(\alpha) + \frac{1}{2}|\xi_\nu(\alpha_\nu)|^2$, *if* $\{\xi(\alpha_{\nu,i})\}$, $i = 1, 2, 3$ *stays bounded and* $\xi_\nu(\alpha_{\nu,3}) - [\xi_\nu(\alpha_{\nu,1}) + \xi_\nu(\alpha_{\nu,2})] \to 0$ *then*

$$L_\nu(\alpha_{\nu,3}) - [L_\nu(\alpha_{\nu,1}) + L_\nu(\alpha_{\nu,2})]$$

tends to zero in probability.

The three conditions (G) (A) and (B) are related as follows.

Lemma 1. *One has always* (A) => (G) => (B) . *When Assumption* 1 *is satisfied all three are equivalent.*

The proof, which is not difficult but somewhat long, will be given elsewhere.

When (G) and Assumption 1 are satisfied, one can replace the experiments $\mathcal{E}_\nu$ indexed by A_ν by other experiments $\mathfrak{F}_\nu$ indexed by C itself according to the following scheme. Let $\Delta_{\nu,n}$ be the distance $\Delta_{\nu,n} = \Delta[\mathcal{E}_{\nu,n}, \mathcal{G}^*_{\nu,n}]$ and let $\beta_\nu = \sum_n 2^{-n} \Delta_{\nu,n}$. If there is a ν_0 such that $\nu \geq \nu_0$ implies $\beta_\nu = 0$ one can, from ν_0 on, let F $F_{t,\nu} = P_{\alpha,\nu}$ if there is an α such that $\xi_\nu(\alpha) = t$. For other values of t one can define $F_{t,\nu}$ by the obvious relation involving likelihood ratios. If there is no ν_o such that $\nu \geq \nu_o$ implies $\beta_\nu = 0$, one can take for $\alpha_\nu(t)$ any point $\alpha_\nu(t) \in A_\nu$ such that

$$|t - \xi_\nu(\alpha_\nu(t))| \leq \beta_\nu + \inf_\alpha\{|t - \xi_\nu(\alpha)| \; ; \alpha \in A_\nu\} .$$

One lets $F_{t,\nu}$ be the measure $P_{\alpha_\nu(t),\nu}$. This construction gives experiments $\mathfrak{F}_\nu$ which enjoy the following property.

Lemma 2. *Let Assumption* 1 *and* (G) *be satisfied*. *Let* $\mathfrak{F}_{\nu,c} = \{F_{t,\nu} \; ; t \in C , |t| \leq c \}$ *and* $\mathcal{G}_{\nu,c} = \{G_t \; ; t \in C , |t| \leq c \}$. *Then for all* $c \in (0,\infty)$ *the distances* $\Delta(\mathfrak{F}_{\nu,c}, \mathcal{G}_{\nu,c})$ *tend to zero as* ν *increases*.

The above conditions (G) or (A) are related to the (LAN) assumption of [3] and the corresponding assumption of [2] as follows. The paper [2] uses a stronger condition, requiring a kind of uniformity which is pleasant, but happens to leave out several interesting cases. Using the notation of the present paper, Hájek requires in [2] that there be statistics S_ν such that

$$\sup_\alpha \{ |\Lambda_\nu(\alpha) - \xi_\nu(\alpha)S_\nu + \tfrac{1}{2}|\xi_\nu(\alpha)|^2| \; ; \alpha \in A_\nu(c) \}$$

converges to zero in probability. It is however clear from the context that such a strong requirement was used only for expository convenience.

On the contrary the (LAN) assumption of [3] would correspond essentially to the weak convergence of $\mathfrak{F}_\nu$ to $\mathcal{G}$. This is weaker than our assumption (G). The difference is not negligible, but not extreme, as can be seen using the results of D. Lindae [7].

We have kept here a formulation which may be deemed more restrictive than absolutely necessary because the metrizable nature of the convergence used lends itself more easily to statements which involve approximations instead of limits.

Let us consider now some of the implications of (G). The following are only meant by way of illustration. For instance we shall use repeatedly quadratic loss functions. However, if as Hájek did in [3], one proves for the Gaussian case and any specified loss function a unique admissibility result, then Theorem 1 can be applied readily.

For the notation, let us agree that if T_ν is any statistic available on $\mathcal{E}_\nu$ the symbol $\mathcal{L}(T_\nu|t)$ denotes the distribution of T_ν for the measure $F_{t,\nu}$ involved in Lemma 2.

Also S_ν will be any statistic available on $\mathcal{E}_\nu$, taking values in R^k and having an asymptotically normal distribution as provided for in Lemma 1 and Requirement A.

<u>Proposition</u> 2. <u>Let Assumptions</u> 1 <u>and</u> (G) <u>be satisfied, for a</u>

cone C _which is the entire Euclidean space_ (R^k, 1.1) _with_ $k \leq 2$. _Let_ T_ν _be any statistic available on_ $\mathcal{E}_\nu$. _Suppose that for all_ $t \in C$ _and all_ $b \in (0,\infty)$ _one has_

$$\lim_\nu \sup \int \min\{b, |T_\nu - t|^2\} dF_{t,\nu} \leq k .$$

Then

$$\lim_\nu \int \min\{1, |T_\nu - S_\nu|\} dP_{\alpha_\nu,\nu} \to 0$$

for all $\{\alpha_\nu\}$ _such that_ $\sup_\nu |\xi_\nu(\alpha_\nu)| < \infty$.

Proof. For the lattice Γ take the space of bounded continuous functions on $Z = R^k$. Let W be the quadratic loss function $W_\theta(z) = |\theta - z|^2$. It has been shown by C. Blyth [8] for $k = 1$ and by C. Stein [9] for $k = 2$ that for the Gaussian family $G_t = \mathfrak{N}(t,I)$ the observed vector X is the unique estimate such that $\int |X-t|^2 dG_t \leq k$ for all $t \in R^k$. If $V \leq W$ is a special loss function it is smaller than $\min(b,W)$ for some b. The assumptions made allow the application of Theorem 1 with the result that

$$\int | \gamma(T_\nu) - \gamma(S_\nu) | dF_{t,\nu} \to 0$$

for all $t \in R^k$ and all $\gamma \in \Gamma$. This implies the result as stated.

In the present case one could carry out a direct argument using the strict convexity of the square loss function. The argument is very simple, but it would not extend to the loss functions used by Hájek in [3]. Note that in any case, the results involve expected square deviations for limiting distributions instead of the possibly larger limits of expected square deviations.

One should not expect the result of Proposition 2 to be valid for $k \geq 3$ since then the Gaussian vector X is not an admissible estimate for usual loss functions. However, for $k \geq 3$ and for the quadratic loss, there are still functions $X \rightsquigarrow \varphi(X)$ which are uniquely admissible and almost surely continuous. Let then $r(t) = \int |\varphi(X) - t|^2 dG_t$ and suppose that Assumption 1, with $C = R^k$, and Assumption (G) are satisfied. If $T_\nu \in \mathcal{D}(\mathcal{E}_\nu, \Gamma)$ and if for every $b \in (0,\infty)$ and $t \in R^k$ one has

$$\lim_\nu \sup \int \min[b, |T_\nu - t|^2] dF_{t,\nu} \leq r(t) ,$$

then

$$\lim_\nu \int \min[1, |T_\nu - \varphi(S_\nu)|] dF_{t,\nu} = 0 .$$

As an example where C is not necessarily the entire space R^k, consider the following situation, which occurs for instance in Neyman's theory of $C(\alpha)$ tests [10]. In this case C is a half-space whose boundary is some hyperplane H. The problem is to test H against the interior of C.

Assume that X has the Gaussian distribution $G_t = \mathfrak{n}(t,I)$. Let Y be the projection of X on the orthogonal complement of H . For any fixed $a \in (-\infty, +\infty)$, the test which rejects H if $Y \geq a$ has a certain probability of error η_0 for $t \in H$, and a probability of error $\eta(t)$ for $t \in C \setminus H$. Let φ be the indicator of the set $\{Y \geq a\}$. Suppose that Assumption 1 and Assumption (G) hold for the present C. For each ν , let φ_ν be a test available on $\mathcal{E}_\nu$. Assume that if $t \in H$ then $\lim_\nu \sup \int \varphi_\nu dF_{t,\nu} \leq \eta_0$ and if $t \in H \setminus C$ then $\lim_\nu \sup \int (1-\varphi_\nu) dF_{t,\nu} \leq \eta(t)$. These conditions imply that

$$\lim_\nu \int |\varphi_\nu - \varphi(S_\nu)| \, d F_{t,\nu} \to 0 .$$

Indeed φ is admissible for the Gaussian experiment. It is uniquely defined by its power function, since $\{G_t ; t \in C\}$ is complete. Also the tests $\varphi(S_\nu)$ have the appropriate limiting behavior, since φ is almost everywhere continuous.

Finally, here is an example of application of Proposition 1 .

<u>Proposition</u> 3. <u>Let Assumption</u> 1 <u>hold with</u> $C = R^k$. <u>Suppose that</u> (G) <u>holds and that</u> $\varepsilon > 0$ <u>is a given number. Then there is a</u> ν_0 , <u>a finite set</u> $S \subset \Theta$ <u>and a</u> $b \in (0,\infty)$ <u>such that</u> $\nu \geq \nu_0$ <u>implies</u> $\sup_{t \in S} \int \min\{b, |T_\nu - t|^2\} \, d F_{t,\nu} > k - \varepsilon$

<u>for every estimate</u> $T_\nu \in \mathcal{D}(\mathcal{E}_\nu, \Gamma)$.

This is a direct application of Proposition 1 using the fact that the minimax risk for the quadratic loss function is equal to k . It resembles the statement of the first part of Theorem 4.1 of Hájek in [3] or the statement of P. Huber in [11].

The application of Proposition 1 makes it clear that the pair (S,b) depends only on k and ε . The element ν_0 depends on the rate of convergence of the $\mathcal{E}_\nu$. There is no dependence on the estimates T_ν . This was not made clear in [3] even though a careful reading of Hájek proof would suggest that the fact was known to him.

Section 4. Independent identically distributed observations.

The asymptotic normality requirements of Section 3, and a fortiori the weaker (LAN) requirement of Hájek in [3] are satisfied in a large variety of cases. Examples can be found in Hájek [12], in the regression models of Hájek and Šidák [13] as well as in the papers [14] , [15] in which independence under the hypothesis is used to deduct results valid under dependent alternatives. However, it appears that the standard independently identically distributed case is still not completely understood. It is about this particular case that Hájek expressed the wish that "someone should write it properly." As will be seen below, we are still far from a definitive theory.

By the standard i.i.d. case will be meant the following. One has a fixed parameter set Θ and a family $\{p_\theta ; \theta \in \Theta\}$ of probability measures on a σ-field $\mathcal{A}$ carried by a set $\mathcal{X}$. For each integer n, one takes the direct products $P_\theta^n = \overset{n}{\underset{j=1}{X}} p_\theta$ on the direct products $(\mathcal{X}^n, \mathcal{A}^n)$ of the pair $(\mathcal{X}, \mathcal{A})$. The family $\{p_\theta ; \theta \in \Theta\}$ stays fixed and limits are taken as $n \to \infty$.

In words, one takes n independent identically distributed observations and let their number n tend to infinity.

For simplicity we shall assume that $p_s = p_t$ implies

$s = t$, so that θ can be considered just another name for p_θ . In addition, it will be assumed that a particular $\theta_0 \in \Theta$ has been singled out for special attention. Instead of p_{θ_0} we shall write simply p .

Situations where the individual observations are still independent, but with distributions which depend both on n and on the label of the individual observation, can also be handled as indicated in [16] and [13], but we shall not consider these situations here except to say that some of the results used below are specializations of results available in the more general cases.

Also we consider here only local problems, as indicated in the Introduction. This means that the sets of interest are of the type $\{\theta : \frac{1}{2}||P_\theta^n - P_{\theta_0}^n|| < 1 - \varepsilon\}$ for some $\varepsilon > 0$.

An equivalent and more convenient formulation uses an Hilbert space representation as follows. Let $h(s,t)$ be the Hellinger distance defined by

$$h^2(s,t) = \tfrac{1}{2} \int (\sqrt{dp_s} - \sqrt{dp_t})^2 .$$

With this metric Θ can be identified with a subset $\tilde{\Theta}$ of a Hilbert space H . It will be convenient to assume that the image of θ_0 is the origin of H . The image of $t \in \Theta$ will be denoted $\tilde{t}$.

The sets of interest are then the sets of the type

$B(\frac{c}{\sqrt{n}}) = \tilde{\Theta} \cap H(\frac{c}{\sqrt{n}})$, the symbol $H(b)$ denoting the ball of radius b centered at the origin of H .

It will also be convenient to use representations of H , or subspaces of H, by spaces of square integrable functions as follows. Let $\Theta_0 \subset \Theta$ be a part of Θ such that $\theta_0 \in \Theta_0$ and such that $\{p_\theta : \theta \in \Theta_0\}$ is dominated. There is then a probability measure μ such that $\mu(A) = 0$ if and only if $p_\theta(A) = 0$ for all $\theta \in \Theta_0$. One can take Radon-Nikodym derivatives $f_\theta^2 = dp_\theta/d\mu$ with $f_\theta \geq 0$. These f_θ belong to the space $L_2(\mu)$ and the map $\tilde{t} = (1/\sqrt{2})(f_t - f)$ with $f = dp/d\mu$ is an affine isometry of the set $\{f_\theta : \theta \in \Theta_0\}$ of $L_2(\mu)$ into H.

The following condition is analogous to condition (G) of Section 3. The two differ in that the set $B(c/\sqrt{n})$ used here is not exactly the same thing as the $A_n(c\sqrt{2})$ of condition (G) .

Condition (G'). *There are maps* $\{\xi_n\}$ *to a Euclidean space* $(R^k, |\cdot|)$ *such that* $\xi_n(\theta_0) = 0$ *and such that for every* $c \in (0,\infty)$ *the distance between the experiment* $\mathcal{E}_{n,c} = \{P_\theta^n ; \theta \in B(c/\sqrt{n})\}$ *and the Gaussian shift experiment*

$$\mathcal{G}^*_{n,c} = \{\mathcal{N}(\xi_n(\theta), I_k) ; \theta \in B(c/\sqrt{n})\}$$

tends to zero as $n \to \infty$.

In view of the results of [1] and [3] it is of interest to know when (G') holds. A possible answer to this question is provided by specialization of statements to be found in [16]. It will be convenient to use a condition weaker than (G') as follows.

Let us call the experiments $\mathcal{E}^n = \{P^n_\theta ; \theta \in \Theta\}$ <u>pairwise asymptotically normal</u> (at θ_0) if for every $c \in (0,\infty)$ and evry sequence $\{s_n\}$, $\tilde{s}_n \in B(c/\sqrt{n})$ there are pairs of Gaussian distributions $\mathcal{P}_n = [\mathfrak{N}(0,\sigma^2_n) , \mathfrak{N}(a_n,\sigma^2_n)] , \sigma_n \in (0,\infty]$ such that the distance $\Delta[\mathcal{P}_n , (P^n_0 , P^n_{s_n})]$ tends to zero as $n \to \infty$.

Decompose each p_t as a sum $p_t = p'_t + p''_t$ of a part p'_t which is dominated by p and a part p''_t which is singular with respect to p . Let $\text{Sglr}(t)$ be the mass of the singular part p''_t . Write the Radon-Nikodym density of p'_t in the form $dp'_t/dp = [1 + Y(t)]^2$ with $1 + Y(t) \geq 0$ considered as random variable for the distribution induced by p .

Finally, let K be the covariance kernel

$$K(s,t) = h^2(s,\theta_0) + h^2(t,\theta_0) - h^2(s,t) .$$

<u>Lemma</u> 3. <u>The experiment</u> $\mathcal{E}^n$ <u>are pairwise asymptotically normal at</u> θ_0 <u>if and only if the following two conditions hold</u> :

(G1) *If* $\sup_n ||\sqrt{n}\,\tilde{s}_n|| < \infty$, *then*

$$\lim_n n\,\mathrm{Sglr}(s_n) = 0 .$$

(G2) *If* $\sup_n ||\sqrt{n}\,\tilde{s}_n|| < \infty$, *then*

$$\lim_n n\,E|Y(s_n)|^2 I\{|Y(s_n) > \varepsilon\} = 0 .$$

For maps $\{\xi_n\}$, the condition (G') is equivalent to the combination of (G1) (G2) and of (G3) as follows

(G3) *If* $\sup_n \{||\sqrt{n}\,\tilde{s}_n|| + ||\sqrt{n}\,\tilde{t}_n||\} < \infty$, *then*

$$\lim_n [n\,K(s_n,t_n)] - \frac{1}{4}\,\xi_n(s_n)[\xi_n(t_n)]' = 0 .$$

Proofs of statements which imply Lemma 3 can be found in [16] , pages 91-95.

In this Lemma 3 the only condition which involves the maps ξ_n or dimensionality considerations is (G3). We shall return to it later. For the present note that the Lindeberg condition (G2) and (G1) are independent requirements, even in the presence of (G3). Also, the combination (G1) (G2) (G3) does not imply in any way that the sequences $\sqrt{n}\,\tilde{s}$ are strongly relatively compact in the Hilbert space H. Cases where they are must be deemed very special even though the conditions of differentiability in quadratic mean of [12] and [4] imply such a behavior. The relatively common occurrence of these cases lends some interest to the following two lemmas. We recall that $\tilde{s}$ is the image of $s \in \Theta$ in the Hilbert space H.

Lemma 4. *Let* $\{t_n\}$ *be a sequence of elements of* Θ *such that* $\sqrt{n}\,\tilde{t}_n$ *converges to a limit* t *in the Hilbert space* H. *Then the corresponding random variables* $Y(t_n)$ *satisfy the Lindeberg condition* (G2).

Proof. The only part of H involved here is the closed linear span of the sequence $\tilde{t}_n$. Thus, there is no loss of generality in assuming that $\{p_\theta : \theta \in \Theta\}$ is dominated. In this case take a probability measure μ equivalent to $\{p_\theta : \theta \in \Theta\}$. Form the space $L_2(\mu)$ and let $f_n \geq 0$ be that element of $L_2(\mu)$ defined by the density $f_{t_n}^2 = dp_{t_n}/d\mu$. Similarly, let $f^2 = dp/d\mu$. Let J be the indicator of the set $\{x : x \in \mathcal{X}, f^2(x) > 0\}$. One passes from $L_2(\mu)$ to H by the isometry $\tilde{t} = (1/\sqrt{2})(f_t - f)$. Multiplication by J is an orthogonal projection of $L_2(\mu)$ onto a subspace L $L_2(\mu')$ with $d\mu' = J\,d\mu$. Thus, if $\sqrt{n}\,\tilde{t}_n \to t$, the images $\sqrt{n}\,J(f_{t_n} - f)$ converge in $L_2(\mu')$.

The map $g \rightsquigarrow gf^{-1}$ is an isometry of $L_2(\mu')$ onto $L_2(p)$. Therefore $Z_n = \sqrt{n}\ J(f_{t_n} - f)f^{-1}$ converges in $L_2(p)$. This convergence implies uniform integrability of the Z_n^2. Explicitely, for each $\varepsilon > 0$ there is a

$b \in (0,\infty)$ such that

$$\int Z_n^2 \, I\{|Z_n| > b\} \, dp < \varepsilon ,$$

for every n. In terms of the random variables $Y(t_n)$ this can also be written

$$n \, E \, |Y(t_n)|^2 \, I\{|Y(t_n)| > b/\sqrt{n}\} < \varepsilon .$$

This condition is stronger than the Lindeberg condition (G2).

The preceding Lemma 4 is essentially a version of Lemma 4 in [4].

The next Lemma is intended to replace Proposition 1 of [4] which asserts that differentiability in quadratic mean implies a condition like (G1) almost everywhere. The proof given there is awkward. It is also garbled by inadvertent substitution of indicators J where $I - J$ should have been written. The following argument is a version of an argument transmitted to us by J. Hájek.

Lemma 5. *Let* s_n *and* t_n *be two elements of* Θ. *Assume that* $\sqrt{n}\,\tilde{s}_n$ *converges in* H *to a limit* s *and that* $\sqrt{n}\,\tilde{t}_n$ *converges to* $t = -\beta s$ *with* $\beta > 0$. *Then* $\{s_n\}$ *satisfies condition* (G1).

Proof. Since only sequences are involved one can assume, as in Lemma 4, that $\{p_\theta : \theta \in \Theta\}$ is equivalent to a probability measure μ. In the space $L_2(\mu)$ the assumptions of the lemma say that there is a $\varphi \in L_2(\mu)$ for which

i) $\int |\sqrt{n}\,(f_{s_n}-f) - \varphi|^2 \, d\mu \to 0$

ii) $\int |\sqrt{n}\,(f_{t_n}-f) + \beta\varphi|^2 \, d\mu \to 0$.

Let Z_1 be the indicator of the set $\{x : \varphi(x) < 0 , f(x) = 0\}$ and let Z_2 be the indicator of $\{x : \varphi(x) \geq 0 , f(x) = 0\}$ so that $Z = Z_1 + Z_2$ is the indicator of $\{x : f(x) = 0\}$. The first relation yields

$$\int Z_1 |\sqrt{n} \;\; f_{s_n} - \varphi|^2 \, d\mu \to 0$$

However $Z_1 |\sqrt{n}\, f_{s_n} - \varphi|^2 \geq n\, Z_1 f_{s_n}^2 + Z_1 \varphi^2$. Thus $n \int Z_1 f_{s_n}^2 \, d\mu \to 0$ and $\int Z_1 \varphi^2 \, d\mu = 0$. Similarly the second relation implies that $\int Z_2 \varphi^2 \, d\mu = 0$. It follows that $\int Z \varphi^2 \, d\mu = 0$. Returning to relation (i) one sees that

$$\int Z |\sqrt{n}\,(f_{s_n}-f) - \varphi|^2 \, d\mu = \int Z |\sqrt{n}\, f_{s_n}|^2 \, d\mu$$

must also tend to zero. Hence the result.

A fairly common case in which the preceding lemmas apply directly, is the situation where Θ itself is a subset of a k-dimensional space R^k and where the map from Θ to $\tilde{\Theta}$ is differentiable. This can also be described as follows.

Suppose $\Theta \subset R^k$ and let ζ be a stochastic process whose covariance is given by $E\,\zeta(s)\,\zeta(t) = \rho(s,t) = \int\sqrt{dp_s dp_t}$.

Take $\theta_0 \in \Theta$ equal to the origin of R^k. The process ζ is differentiable in quadratic mean at $\theta_0 = 0$ if there is a vector $\zeta^{\cdot} = (\zeta_1^{\cdot}, \cdots, \zeta_k^{\cdot})$ of random variables such that

$$\lim_{\varepsilon \to 0} \sup_{|t|<\varepsilon, t\in\Theta} E\left| \frac{1}{|t|}[\,[\,\zeta(t) - \zeta(0)] - t(\zeta^{\cdot})'\,]\,\right|^2 = 0 .$$

A ray $\{\alpha u \,;\, \alpha \geq 0\}$ of R^k is said to belong to the contingent of Θ at zero if there are elements $t_n \in \Theta$ such that $t_n \to 0$ and $t_n|t_n|^{-1} \to u$.

Proposition 3. Assume that the process ζ is differentiable in quadratic mean at zero. Then conditions (G2) and (G3) are satisfied. If in addition the contingent of Θ at zero contains straight lines $\{\alpha u_j \,;\, \alpha \in (-\infty, +\infty)\}$ where the u_j form a basis of R^k, then condition (G1) is also satisfied.

This is an immediate consequence of Lemmas 4 and 5. We shall return to differentiability in quadratic mean in Section 5. For the present, note that it is not enough to assume that the process $t \rightsquigarrow Y(t)$ of Lemma 3 is differentiable in quadratic mean. An example can be constructed as follows. For $\theta \in (-\frac{1}{2}, \frac{1}{2})$ let q_θ be the Gaussian distribution $\mathfrak{N}(\theta, 1)$ on the line. Let s be a probability measure which is singular with respect to Lebesgue measure. Define measures p_θ by $p_\theta = (1-\theta^2)q_\theta + \theta^2 s$, It is easily checked that for this family the process $t \rightsquigarrow Y(t)$ is differentiable in quadratic mean at zero. Also the process $\zeta(t)$ admits *one sided* derivatives in quadratic mean. However the singular masses $\mathrm{Sglr}(t)$ are exactly equal to t^2 so that (G1) fails to hold. Returning to the general case where Θ had no structure of its own, note that the condition (G3) expresses the fact that the maps $\frac{1}{2}\zeta_n$ are approximate isometries for the convex hulls of the sets $B(c/\sqrt{n})$ in H.

This suggests that (G3) can be expressed as a condition of a purely geometrical nature on the sets $B(c/\sqrt{n})$. The following condition (C_n) is of such a character. It implies (G3) and is implied by a combination of (G3) and of a condition slightly weaker than the Assumption 1 of Section 3.

Condition (C_n). *For each integer* n, *there is a* k-*dimensional linear subspace* H_n *of* H *and a convex cone* $C_n \subset H_n$ *which has its vertex at the origin of* H_n. *For every* $b \in (0,\infty)$ *the Hausdorff distance between* $H(b) \cap C_n$ *and* $H(b) \cap (\sqrt{n}\,\tilde{\Theta})$ *tends to zero as* $n \to \infty$. *In addition the solid angle spanned by* C_n *in* H_n *stays bounded away from zero.*

As mentioned above, it is easily verified that condition (C_n) is equivalent to (G3) reinforced by an assumption of nondegeneracy similar to Assumption 1, Section 3.

It is also easily verified that the sets C_n must satisfy the following slow variation requirement. Let (m_1, m_2) be a fixed pair of integers. Then, for any b, the Hausdorff distance between $H(b) \cap (C_{nm_1})$ and $H(b) \cap (C_{nm_2})$ must tend to zero.

This slow variation does not imply in any way that the C_n have a limit. The case where they can be taken fixed and all equal to a given C is closely related to the condition of differentiability in quadratic mean discussed above. It must be regarded as a very special case.

Several examples which satisfy (G') with variable cones C_n have been known for a long time. One may mention for instance the shift densities $[1 - |x-\theta|]^+$ or $C \exp\{ - |x-\theta|^{\frac{1}{2}}\}$ on the line. See [16] pages 108-111 and [17].

Unfortunately a theory describing which sequences $\{C_n\}$ may arise, and when, is not yet available.

Section 5. Differentiability in quadratic mean.

As mentioned in Section 4, the situation covered by Condition (C), that is condition (C_n) with cones C_n all identical to one given C is closely related to the situation where Θ is itself a subset of R^k and where the process $t \leadsto \xi(t)$ is differentiable in quadratic mean.

The present section is an elaboration on conditions which imply this differentiability in quadratic mean and are often easily verifiable.

Consider first the case where Θ is an open interval of the real line.

In many statistical problems the measures p_θ are given by specifying their densities with respect to a common dominating measure μ. Explicitely one gives a function $(x,\theta) \leadsto f(x,\theta)$ defined on $\mathcal{X} \times \Theta$ such that $f(x,\theta) \geq 0$ and $\int f(x,\theta)\mu(dx) = 1$ for each $\theta \in \Theta$.

It is common practice to take derivatives with respect to θ for each fixed x and to call Fisher information the number $i(\theta) = \int (1/f(x,\theta))[\partial/\partial\theta\ f(x,\theta)\ ]^2\, d\mu$. Unfortunately the literature is rather vague about what happens when there are points θ possibly dependent on x at which the ordinary derivatives fail to exist.

This author tried to clarify the situation in [4]. However, the statements made at the end of [4] could be interpreted to mean that one must check directly the absolute continuity of the images $\theta \rightsquigarrow \sqrt{f(x,\theta}$. It was shown by Hájek in [3] that one can instead check absolute continuity of the maps $\theta \rightsquigarrow f(x,\theta)$ themselves.

The following argument gives a precise result applicable to many situations. Instead of absolute continuity we shall use Lusin's condition (N). A function φ from an interval I of the line to the line satisfies Lusin's condition (N) if for every set $S \subset I$ which has Lebesgue measure zero, the image $\varphi(S)$ of S also has Lebesgue measure zero. This condition is very much weaker than absolute continuity.

For any function φ defined on an interval I of $\mathbb{R} = (-\infty, +\infty)$ we shall use a derivative $\dot{\varphi}$ as follows :

$\dot{\varphi}(t)$ is the derivative $\varphi'(t)$ of φ at t if this derivative exists and is finite. Otherwise $\dot{\varphi}(t)$ is zero .

Suppose that ψ maps an interval I into an interval J and that φ is real valued defined on J .

The composed function $g = \varphi \cdot \psi$ has a "derivative" $\dot{g}$ as above. It has also a "chain rule" derivative $\bar{g}$ which can be written $\bar{g}(t) = \dot{\varphi}[\psi(t)]\,\dot{\psi}(t)$. In other words $\bar{g}(t)$ is zero unless both derivatives exist and are finite.

The following lemma asserts that, under suitable restrictions, the function g is the indefinite integral of $\bar{g}$.

Lemma 6. *Let* φ *be a function from an interval* I *to an interval* J *and let* ψ *be a function from* J *to the line. Assume that* φ *and* ψ *are continuous and that they both satisfy Lusin's condition* (N) .

Suppose also that φ *admits a finite derivative almost everywhere on* J *and let* $g = \varphi \cdot \psi$. *Then for any interval* $[a,b] \subset I$ *one has*

$$|g(b) - g(a)| \le \int_a^b |\overline{g}(t)|\, dt .$$

Proof. Let $s = \psi(t)$ be a point of J at which the derivative $\varphi'(s)$ exists, is finite *and not zero*. At such a point one must have $\overline{g}(t) = \dot{g}(t)$. Let S be the subset of J where φ' exists but is zero and let T be the subset where the derivative does not exist or has infinite absolute value.

Denote the Lebesgue measure by λ . We have assumed that $\lambda(T) = 0$. Thus, since φ satisfies (N) one has $\lambda[\varphi(T)] = 0$.

According to Theorem 6.5 page 227 of [18] , one has also $\lambda[\varphi(S)] = 0$. Thus, one has $\dot{g}(t) = \overline{g}(t)$, except for some set of values of t whose image by the map $g = \varphi \cdot \psi$ has Lebesgue measure zero. The map g itself satisfies condition (N). The desired result is then a consequence of Theorem 6.9 page 281 of [18].

With this we can return to the Fisher information. For this purpose, let $(\mathcal{X}, \mathcal{A})$ carry a positive measure μ and let f defined on $\mathcal{X} \times I$, $I \subset \mathbb{R}$ be such that $f(x,t) \geq 0$ and $\int f(x,t)\mu(dx) = 1$. Let $g(x,t) = \sqrt{f(x,t)}$. For each fixed x define a derivative $\bar{g}(x,t)$ as follows. If $f(x,t) > 0$ and if the derivative $f'(x,t) = df(x,t)/dt$ exists and is finite then

$$\bar{g}(x,t) = \frac{1}{2}\frac{f'(x,t)}{\sqrt{f(x,t)}} . \quad \text{Otherwise} \quad \bar{g}(x,t) = 0 .$$

Let $s(t) \geq 0$ be defined by

$$s^2(t) = \int |\bar{g}(x,t)|^2 \mu(dx) .$$

The map which associates to t the equivalence class of $g(x,t)$ in the space $L_2(\mu)$ will be denoted $t \leadsto G(t)$.

Proposition 4. *Assume that for almost every* $x \in \mathcal{X}$, *the map* $t \leadsto f(x,t)$ *is continuous and satisfies condition* (N). *Assume also that for every compact interval* $[a,b] \subset I$ *one has* $\int_a^b s(t)\,dt < \infty$. *Then the map* $t \leadsto G(t)$ *is strongly differentiable at almost all points of* I. *For ~~about~~ almost all* t *the function* $x \leadsto \bar{g}(x,t)$ *is in the class* $G'(t)$. *Furthermore* G *is an indefinite Bochner-Lebesgue integral of its derivative* $G'(t)$.

Proof. The continuity of $t \rightsquigarrow f(x,t)$ implies that f is jointly measurable and that the family of probability measures p_t with densities $dp_t = f(\cdot,t)d\mu$ is dominated by a probability measure. The value $s^2(t)$ does not depend on which dominating measure is used. Thus it is enough to prove the result assuming that μ itself is a probability measure.

The derivative $\bar{g}$ is also jointly measurable. By Fubini's theorem one may write

$$\int \{ \int_a^b | \bar{g}(x,t) | \, dt \} \, \mu(dx) \leq \int_a^b s(t) \, dt \; .$$

From this follows that for almost all x we have $\int_a^b | \bar{g}(x,t) | \, dt < \infty$ .. One concludes from Theorem 7.7 page 285 of [18] and Lemma 5 that for almost all x the map $t \rightsquigarrow g(x,t)$ is absolutely continuous and such that

$$g(x,b) - g(x,a) = \int_a^b \bar{g}(x,t) \, dt \; .$$

To proceed, take a fixed element u of $L_2(\mu)$ and let

$$\begin{aligned} \langle u, G(t) \rangle &= \int u(x) g(x,a) \mu(dx) \\ &\quad + \int \{ u(x) \int_a^t \bar{g}(x,t) dt \} \, \mu(dx) \\ &= \int u(x) \, g(x,a) \mu(dx) \\ &\quad + \int_a^t \{ \int u(x) \, \bar{g}(x,\tau) \mu(dx) \} \, d\tau \; . \end{aligned}$$

This equation shows that G is the indefinite Pettis integral of $\overline{g}$. However since $\overline{g}$ is jointly measurable and since its norm is integrable, the Pettis integral is also a Bochner integral. Hence the result.

The foregoing proposition can be complemented by a remark as follows. Let us say that θ is a Lebesgue point of s if $s(\theta) < \infty$ and

$$\lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_{-\varepsilon}^{+\varepsilon} |s(\theta + \tau) - s(\theta)| \, d\tau = 0 \, .$$

The following Corollary is analogous to Proposition 7 of [4]. However here we have a special choice of "derivatives" $\overline{g}$. This replaces the $\sigma(t)$ of [4] by $s(t)$ and allows removal of the extra assumption of differentiability in measure.

Corollary. *Assume that* μ *is finite and that for almost all* x *the map* $t \leadsto f(x,t)$ *is continuous and satisfies condition* (N) . *Let* θ *be a Lebesgue point of* $t \leadsto s(t)$. *Then the map* $t \leadsto G(t)$ *is differentiable at* $t = \theta$ *and the derivative* $G'(\theta)$ *is the class of* $\overline{g}(x,\theta)$.

Proof. Let φ_τ be the function $\varphi_\tau(x) = \frac{1}{\tau}[g(x,\theta+\tau) - g(x,\theta)]$. Write $\overline{g}(x)$ for $\overline{g}(x,\theta)$ and let A be the subset of on which $\lim_{\tau \to 0} \varphi_\tau(x) = \overline{g}(x)$. The definition of $\overline{g}$ implies that

$$v^2 = \liminf_{\tau} |\varphi_\tau|^2 \geq |\overline{g}|^2 \, .$$

An application of Fatou's lemma shows that

$$\liminf_{\tau} \int |\varphi_\tau|^2 d\mu \geq \int v^2 \, d\mu \geq \int |\overline{g}|^2 d\mu \, .$$

The fact that Θ is a Lebesgue point of $t \rightsquigarrow s(t)$ implies as in [4], that

$$\limsup_{\tau} \int |\varphi_\tau|^2 \, d\mu \leq s^2(\Theta) = \int |\overline{g}|^2 \, d\mu \, .$$

Let ψ_τ^2 be the function

$$\psi_\tau^2 = \inf_{|\varepsilon| \leq \tau} |\varphi_\varepsilon|^2 \, .$$

The above relations imply that

$\int [\, |\varphi_\tau|^2 - |\psi_\tau|^2 \,] \, d\mu$ tends to zero as $\tau \to 0$. In addition $|\psi_\tau|^2$ increases to a certain function v^2 which is equal to $|\overline{g}|^2$ on A. Since $\int v^2 d\mu = \int |\overline{g}|^2 d\mu$ and since $\overline{g}$ is equal to zero outside A , we conclude that v^2 itself is almost everywhere zero on A^c . Therefore $\int_{A^c} |\varphi_\tau|^2 d\mu$ tends to zero as $\tau \to 0$. It follows that $\varphi_\tau \to \overline{g}$ in measure and that $\int |\varphi_\tau|^2 d\mu \to \int |\overline{g}|^2 d\mu$. The usual argument shows then that $\int |\varphi_\tau - \overline{g}|^2 d\mu$ must tend to zero. Hence the result.

The above statements admit a variety of converses. For instance, suppose that the map $t \rightsquigarrow G(t)$ admits almost everywhere on I a weak derivative $G'(t)$. Assume also that $t \rightsquigarrow G(t)$ is the indefinite Bochner integral of G'. Then there is a version $g(x,t)$ of $G(t)$ such that $t \leadsto g(x,t)$ is absolutely continuous for almost all x and to which all the preceding arguments can be applied.

Note that in this one dimensional case differentiability in quadratic mean, and a moderate amount of integrability leads immediately to densities $f(x,t)$ which are continuous in t for each x. We are indebted to R. M. Dudley for several examples showing that differentiability in quadratic mean does not necessarily imply continuity of the trajectories $t \leadsto f(x,t)$ when the integrability conditions are removed.

However, one should specially note that continuity of trajectories is not at all implied by differentiability in quadratic mean as soon as the parameter space Θ is allowed to have more than one dimension. This is true even for shift families under rather severe integrability restrictions. For instance, in R^2, one can take, with respect to Lebesgue measure, a density f which is extremely smooth except that near $x = 0$ it behaves like $\log \log(1/|x|)$. The family $f(x-\Theta)$ satisfies all due requirements of differentiability in quadratic mean, but the trajectories are not continuous. Similarly, for R^3 one can take densities f which are smooth, except that near zero they behave like $|x|^{-\frac{1}{2}}$.

These examples, communicated to the author by J. B. H. Kemperman, show that although differentiability in quadratic mean entails a wealth of statistical consequences it does not in any way imply proper behavior of certain standard techniques such as the maximum likelihood technique. It is to be hoped that the mystique surrounding these techniques will eventually fade.

References

[1] J. Hájek - "A characterisation of limiting distributions of regular estimates." Zeit. f. Wahrscheinlichkeitstheorie u.v. Gebiete. Vol. 14 (1970) pp. 323-330.

[2] J. Hájek - "Limiting properties of likelihoods and inference." Foundations of Statistical Inference. Godambe and Sprott, editors, Holt-Rinehart-Winston Toronto 1971.

[3] J. Hájek - "Local asymptotic minimax and admissibility in estimation." Proceedings 6th Berkeley Symposium on Math. Stat. and Prob. Vol. 1 (1972) pp. 175-194.

[4] L. Le Cam - "On the assumptions used to prove asymptotic normality of maximum likelihood estimates." Ann. Math. Stat. Vol. 41 (1970) pp. 802-828.

[5] L. Le Cam - "Limits of experiments." Proc. 6th Berkeley Symp. on Math. Stat. and Prob. Vol. 1 1972 pp.

[6] L. Le Cam - "Sufficiency and approximate sufficiency." Ann. Math. Stat. Vol. 35 (1964) pp. 1419-1455.

[7] D. Lindae - "Distributions of likelihood ratios and convergence of experiments." Unpublished Ph.D. Thesis, University of California, Berkeley 1972.

[8] C. R. Blyth - "On minimax statistical decision procedures and their admissibility." Ann. Math. Stat. Vol. 52 (1951) pp. 22-42.

[9] C. Stein - "Inadmissibility of the usual estimate for the mean of a normal distribution." Proc. 3rd Berkeley Symposium on Math. Stat. and Prob. Vol. 1 (1956) pp. 197-206.

[10] J. Neyman - "Optimal asymptotic tests of composite statistical hypotheses." The Harald Cramér Volume, Wiley, N. Y. (1959) pp. 213-234.

[11] P. Huber - "Strict efficiency excludes super efficiency." (Abstract) Ann. Math. Stat. Vol. 37 (1966) p. 1425.

[12] J. Hájek - "Asymptotically most powerful rank order tests." Ann. Math. Stat. Vol. 53 (1962) 1124-1147.

[13] J. Hájek and Z. Šidák - Theory of rank tests C.S.A.V. Prague 1967.

[14] G. L. Yang - "Contagion in stochastic models for epidemics." Ann. Math. Stat. Vol. 39 (1968) 1863-1889.

[15] R. H. Traxler - "On tests for trend in renewal Processes." Ph.D. Thesis, University of Calif, Berkeley 1974.

[16] L. Le Cam - Theórie asymptotique de la décision statistique - Presses de l'Université de Montreál 1969.

[17] Michel Woodroofe - "Maximum likelihood estimation of a translation parameter of a truncated distribution. Ann. Math. Stat. Vol. 43 (1972) pp. 113-122.

[18] S. Saks - Theory of the Integral. Warsaw 1937 - Dover reprint 1964.