On the standard asymptotic confidence ellipsoids of Wald

By

L. Le Cam

University of California, Berkeley

Technical Report No. 100
July 1987
(revised November 1987)

Research was supported by NSF Grant DMS-8701426

Department of Statistics
University of California
Berkeley, California

On the standard asymptotic confidence ellipsoids of Wald

By L. Le Cam

University of California, Berkeley
NSF Grant DMS-8701426

1. Introduction. In his famous paper of 1943 Wald proved asymptotic optimality properties for a variety of tests of simple or composite hypotheses. The tests are derived from a recipe that involves maximum likelihood estimates $\hat{\theta}_n$ and estimates Γ_n of the inverse of the covariance matrix of $\hat{\theta}_n$. One forms a chi-square type statistic $(\hat{\theta}_n - \theta)'\Gamma_n(\hat{\theta}_n - \theta)$ and reject those θ's for which the statistic is too large. For Γ_n Wald uses the Fisher information matrix J_θ evaluated the estimate $\hat{\theta}_n$ of θ. It has been noted by several authors that Wald's procedure can suffer from some unsatisfactory features. One defect, noted by Hauck and Donner (1977) is that, for fixed θ, a criterion of the type $(\theta - t)'J_t(\theta - t)$ can decrease as $|\theta - t|$ becomes large. Another feature, noted by Vaeth (1985) is that the results of the test procedure are not invariant under smooth one-to-one transformations of the parameter space. Vaeth also points out that the behavior described by Hauck and Donner can occur if one uses the criterion $(\theta - \hat{\theta})'J_{\hat{\theta}}(\theta - \hat{\theta})$ for the maximum likelihood estimate $\hat{\theta}$: There can be sequences $\hat{\theta}_n$ such that $(\theta - \hat{\theta}_n)'J_{\hat{\theta}_n}(\theta - \hat{\theta}_n)$ tends to zero for all fixed values of θ. An example of this, imitated from Vaeth, (1985), will be given in Section 3.

The purpose of the present paper is to propose a substitute for Wald's chi-square formula. It is as follows. Let $\{\mathbf{E}_n\}$ be a sequence of experiments $\mathbf{E}_n = \{P_{\theta,n};\ \theta \in \Theta_n\}$. Let $q_n^2(s,t) = -8\log\int\sqrt{dP_{s,n}\,dP_{t,n}}$. Our proposal is to use $q_n^2(s,t)$ as a measure of the separation of the parameter points s and t and to build confidence intervals of the type $\{\theta;\ q_n^2(\mathbf{T}_n,\theta) \le c_n(\theta)\}$ for suitably selected estimates $\mathbf{T}_n$.

The performance of the procedure depends, of course, on the choice of the estimates $\mathbf{T}_n$. For the situation described by Wald and for $\mathbf{T}_n$ equal to the maximum likelihood the conditions

$$q_n^2(\mathbf{T}_n,\theta) \le c \quad \text{and} \quad (\theta - \mathbf{T}_n)'J_{\mathbf{T}_n}(\theta - \mathbf{T}_n) \le c$$

can be shown to be locally equivalent in the following sense: Let θ_0 be the true value of the parameter. Then for sequences $\{\theta_n\}$ such that $q_n^2(\theta_n,\theta_0)$ remains bounded the difference

$$q_n^2(\hat{\theta}_n, \theta_n) - (\hat{\theta}_n - \theta_n)' J_{\hat{\theta}_n}(\hat{\theta}_n - \theta_n)$$

tends to zero in $P_{\theta_0, n}$ probability.

Since q_n^2 is intrinsically defined by the experiment E_n, it remains invariant under all one-to-one transformations of the parameter space.

The function q_n^2 is a monotone increasing function of the Hellinger distance h defined by $h^2(s,t) = \frac{1}{2}\int(\sqrt{dP_{s,n}} - \sqrt{dP_{t,n}})^2$. Since the Hellinger distance is closely related to the total variation norm distance, and even more closely connected to the square distance $k_n^2(s,t)$

$$= \frac{1}{2}\int \frac{(dP_{s,n} - dP_{t,n})^2}{d(P_{s,n} + P_{t,n})},$$

a decrease of $q_n^2(\theta, t)$ for fixed θ as $|\theta - t|$ increases is an indication that the distance $|\theta - t|$ or the parametrization is not well chosen. Thus if the Hauck-Donner phenomenon arises for q_n^2, one should check the parametrization.

The use of q_n^2 instead of a chi-square has some disadvantages. For one thing it may be harder to compute than the chi-square. However, leaving this aside, the main inconvenience is that to compute $q_n^2(\mathbf{T}_n, \theta)$ the value of the estimate $\mathbf{T}_n$ must lie in the range of definition of q_n^2. For a first draft of the present paper we had not paid sufficient attention to that requirement. The gap was pointed out by Yu-Lin Chang, who deserves my thanks. That there is a real difficulty is pointed out in Section 2 below. Attempts at remedying it are described in Section 5.

Otherwise the paper is organized as follows. Section 2 describes Gaussian shift experiments and the role there of the function we called q^2. It also describes heteroschedastic normal experiments pointing out that chi-square formulas are not readily defensible there. This same Section 2 reviews the results obtained in Wald's paper of 1943 or, more precisely, our version of them.

Section 3 gives details of an example analogous to the one considered by Vaeth (1985). The behavior of our substitute criterion appears satisfactory.

Section 4 is suggested by the heteroschedastic approximations that occur naturally in the framework used by Wald. It show that variations on the definitions of the chi-square type criteria can lead to very different answers.

Section 5 touches upon a number of different matters: the effect of lack of uniformity in the local convergence to Gaussian shift experiments, the need to use estimates $\mathbf{T}_n$ that take values outside the assumed parameter sets Θ_n and some possibilities for the extension of the domain of definition of q_n to cover such eventualities.

Section 6 is an aside on covariance stablizing transformations, An appendix gives the derivation of the formula for $q^2(s,t)$ in the heteroschedastic Gaussian case.

2. Local Asymptotic normality, an outline of the theory.

In this section we recall a few facts about approximations by Gaussian shift experiments. The facts are well known but presented in a manner that emphasizes the role of chi-square type expressions and the role of the function q^2 defined in the introduction.

Let Θ be an arbitrary set. An experiment $\mathbf{G} = \{G_\theta : \theta \in \Theta\}$ is called Gaussian, or Gaussian shift for precision, if it satisfies the following two conditions:

1) The measures G_θ; $\theta \in \Theta$ are mutually absolutely continuous,

2) Let $\Lambda(t,s) = \log \frac{dG_t}{dG_s}$. Then the stochastic process $t \rightarrow \Lambda(t,s)$, $t \in \Theta$ is a Gaussian process for the distribution induced by the measure G_s.

(Under condition (1) the choice of the point s does not matter).

Note that the definition of Gaussian given here does not refer to any particular algebraic or vectorial structure that may exist on Θ. The set Θ was not assumed to have any such structure. However if one is given a Gaussian experiment $\mathbf{G}$ on Θ one is also automatically given a map of Θ into a Hilbert space. To define it consider the process $X(t) = \Lambda(t,s) - E_s \Lambda(t,s)$. Let $M_0(\Theta)$ be the set of finite signed measures with finite support on Θ that are such that $\mu(\Theta) = 0$. Let $\|\mu\|^2$ to be the variance of the random variable $\int X(t)\mu(dt)$. These Gaussian variables generate a Hilbert space. One maps Θ into it by associating to θ the difference $\delta_\theta - \delta_s$ of the Dirac masses carried by θ and s. The square distance between θ and t becomes

$$\begin{aligned} \|\delta_\theta - \delta_t\|^2 &= E_s |X(\theta) - X(t)|^2 \\ &= -8 \log \int \sqrt{dG_\theta \, dG_t}, \end{aligned}$$

as can be readily checked.

If the measures G_θ had been given by a standard normal density with respect to the Lebesgue measure of the form

$$\frac{|\det \Gamma|^{1/2}}{(2\pi)^{k/2}} \exp\{-\frac{1}{2}(x-\theta)'\Gamma(x-\theta)\}$$

with Γ fixed, independent of θ, one would have $\|\delta_\theta - \delta_t\|^2 = (\theta - t)'\Gamma(\theta - t)$.

Note that the Gaussian family $\{G_\theta : \theta \in \Theta\}$ can be expanded to a Gaussian shift experiment $\{G_\mu : \mu \in M_0(\Theta)\}$ indexed by the entire linear space $M_0(\Theta)$, or by the Hilbert space that completes it. To do this define G_μ by $dG_\mu = \exp\{\int X(t)\mu(dt) - \frac{1}{2}\|\mu\|^2\} dG_s$.

All of this suggests that, if an experiment $E = \{P_\theta ; \theta \in \Theta\}$ is not exactly Gaussian shift but only approximately so, an appropriate measure of the separation of two parameter values s and t is the number $q^2(s,t) = -8\log \int \sqrt{dP_s\, dP_t}$.

The function $(s,t) \to q(s,t)$ is not necessarily a metric. In particular it does not necessarily satisfy the triangular inequality. However $|1 - \exp[-\frac{q^2}{8}]|^{1/2}$ is a metric on Θ since it is the Hellinger distance. For further use note that for the Gaussian case one has $\|\mu\|^2 = -\frac{1}{2}\int\int \|\delta_s - \delta_t\|^2 \mu(ds)\mu(dt)$. An attempt to extend the definition of q^2 to all of $M_0(\Theta)$ by taking the corresponding $\Gamma(\mu) = -\frac{1}{2}\int\int q^2(s,t)\mu(ds)\mu(dt)$ can fail because the form so defined need not be positive and because q^2 may take infinite values.

We have mentioned "approximately Gaussian experiments". A formal definition can be given as follows.

In Le Cam, (1964), we introduced a distance $\Delta(\mathbf{E},\mathbf{F})$ between two experiments $\mathbf{E} = \{P_\theta ; \theta \in \Theta\}$ and $\mathbf{F} = \{Q_\theta ; \theta \in \Theta\}$ indexed by the same set Θ. Except for technicalities, to say that $\Delta(\mathbf{E},\mathbf{F}) \le \varepsilon$ is to say that, as long as one uses only loss functions W bounded by zero and unity, any risk function available on one of the experiments can be matched within ε by a risk function available on the other experiment.

Definition 1. A sequence $\{\mathbf{E}_n\}$ of experiments $\mathbf{E}_n = \{P_{\theta,n} ; \theta \in V_n\}$ is asymptotically Gaussian shift if there are Gaussian shift experiments $\mathbf{G}_n = \{G_{\theta,n} ; \theta \in V_n\}$ such that $\Delta(\mathbf{E}_n, \mathbf{G}_n) \to 0$ as $n \to \infty$.

Note that we have written the parameter sets as V_n instead of Θ_n. This is because the kind of approximation required by Definition 1 will usually be possible only very locally, in "neighborhoods that shrink as $n \to \infty$", instead of on an entire parameter space. In fact, even under the very severe conditions imposed by Wald in his 1943 paper, Gaussian shift approximations are only possible locally unless the Fisher information matrices are constant. By contrast Wald shows that his conditions imply the possibility of a global approximation by heteroschedastic Gaussian experiments. An heteroschedastic Gaussian experiment — henceforth abbreviated HetG — is a family $\{H_\theta ; \theta \in \Theta\}$ in which H_θ is the distribution of a vector Y which is Gaussian but with expectation *and covariance* that depend on θ.

Let P be a Gaussian distribution with center θ and covariance matrix Γ^{-1} on $\mathbb{R}^k$. Let Q be another Gaussian measure with center t and covariance matrix K^{-1} on the same $\mathbb{R}^k$. The value of $q^2 = -8\log\int\sqrt{dPdQ}$ is easily seen to be

$$-2\log\det[I - (M^{-1}\Delta)^2] + (t-\theta)'[M - \Delta M^{-1}\Delta](t-\theta)$$

where M is the matrix $M = \frac{1}{2}(\Gamma + K)$ and $\Delta = \frac{1}{2}(\Gamma - K)$. (C. Kraft, 1955, where however the formula is misprinted. See the Appendix).

This formula shows that q^2 consists of two terms. One of them measures the difference between the centers θ and t and reduces to the chi-square formula if $\Gamma = K$. The other term involves the determinant $\det[I - (M^{-1}\Delta)^2]$. Examples can be readily constructed where $-2\log\det(I - M^{-1}\Delta M^{-1}\Delta)$ is very large but where both $(\theta - t)'\Gamma(\theta - t)$ and $(\theta - t)'K(\theta - t)$ are small. This is part of the reason for the misbehavior of chi-square type criterion in the approximations used by Wald.

We have already used the word "locally" to describe properties valid on small enough sets. In the sequel we shall use the word "locally" in the following manner: Let $\{\mathbf{E}_n\}$ be a sequence of experiments $\mathbf{E}_n = \{P_{\theta,n};\ \theta \in \Theta_n\}$. Let q_n^2 be the corresponding separation function defined by $q_n^2(s,t) = -8\log\int\sqrt{dP_{s,n}\,dP_{t,n}}$. A "local" property is one that is valid on certain specified sets of the type $V_n(\tau_n, b) = \{\theta : q_n^2(\theta, \tau_n) \le b\}$ for specified sequences $\{\tau_n\}$ and, usually, for any arbitrarily fixed value of b.

In this sense global approximability by a HetG does not always imply local approximability by Gaussian shift experiments. To get into details of possible behavior, let us look at Wald's conditions, or more precisely at some of the conclusions the conditions were presumably meant to entail. The wording will be imitated from Le Cam (1956) and not directly from Wald's 1943 paper for reasons that will appear later.

Consider a sequence $\{\mathbf{E}_n\}$ of experiments $\mathbf{E}_n = \{P_{\theta,n};\ \theta \in \Theta_n\}$ where it will be assumed that Θ_n is a subset of a k-dimensional real vector space $\mathbf{R}^k$. It will also be assumed that $\mathbf{E}_n$ comes with estimates T_n defined on it. The estimates T_n will be assumed to take values in $\mathbf{R}^k$. If they take values in Θ_n, they are called strict.

One of the first conclusions in Wald's paper concerns the case where T_n are the maximum likelihood estimates and where their behavior is describable as follows

(A) Let $\{\tau_n\}$ be an arbitrary sequence, $\tau_n \in \Theta_n$. These are non random matrices $M_{\tau_n, n}$ such that, if $q_n(\theta_n, \tau_n)$ remains bounded, then the distribution $\mathbf{L}\{M_{\tau_n, n}(T_n - \theta_n) | \theta_n\}$ tend to the standard k-dimensional normal $\mathbf{N}(0, I)$.

B(1) The T_n take their values in Θ_n

B(2) If $\{\tau_n\}$ is as in (A) then, given $\varepsilon > 0$, there is a $b = b(\varepsilon)$ and an $N = N(\varepsilon)$ such that $n \ge N(\varepsilon)$ implies $P_{\tau_n, n}\{q_n^2(T_n, \tau_n) > b\} < \varepsilon$

(C) The T_n are asymptotically sufficient in the following sense: There are other families $\{Q_{\theta,n};\ \theta \in \Theta_n\}$ of probability measures, defined on the same σ-fields as the $P_{\theta,n}$, such that:

(i) For $\{Q_{\theta,n}; \theta \in \Theta\}$ the statistics T_n are sufficient (exactly so!).

(ii) $\sup_\theta \{\|P_{\theta,n} - Q_{\theta,n}\|; \theta \in \Theta_n\} \to 0$ for the total variation distance between measures.

D) Let $F_{\theta,n}$ be the distribution of T_n for $P_{\theta,n}$. There are Gaussian distributions $G_{\theta,n}$ centered at θ and with the following properties:

There are Markov kernels K_n' and K_n'' such that $\sup_\theta \|F_{\theta,n} - K_n' G_{\theta,n}\|$ and $\sup_\theta \|G_{\theta,n} - K_n'' F_{\theta,n}\|$ tend to zero as $n \to \infty$.

Finally the K_n' and K_n'' represent small distortions in the following sense. Let $\Gamma_{\theta,n}$ be the inverse of the covariance matrix of $G_{\theta,n}$, (assumed to exist).

E) For every sequence $\{\tau_n\}$, $\tau_n \in \Theta_n$, every $b < \infty$ and every $\varepsilon > 0$, let $B_n(t,\varepsilon) = \{x : (x-t)'\Gamma_{\tau_n,n}(x-t) < \varepsilon\}$ then $\sup_t K_n'[B_n^c(t,\varepsilon) \mid t]$; $(t-\tau_n)'\Gamma_{\tau_n,n}(t-\tau_n) \le b\}$ tends to zero as $n \to \infty$ and similarly for K_n''.

The reader should note that the conjunction of the properties (A) to (E) is very restrictive. One reason for this is the requirement that all convergence properties hold uniformly on the entire sets Θ_n. Another, perhaps less visible reason, is that the estimate T_n must be *strict*. That is they must take their values in Θ_n. A combination of (A) and (B) or (A) and (D) will usually force the points of Θ_n to be interior to the closure of Θ_n and even more. There is also some doubt that the sufficiency property of (C) can hold uniformly unless the space Θ_n is very special, for instance the whole of $\mathbf{R}^k$. In that case the uniformity of convergence does not usually hold.

On this particular point Wald's arguments seem to contain a gap. He works with a fixed subset $\Theta_n = \Theta$ of $\mathbf{R}^k$ but fails to specify what kind of set it may be. He proceeds as if the maximum likelihood estimates were, except for uniformly negligible probabilities, roots of the maximum likelihood equations. This cannot be at the usual kind of boundary points. Some of the problems in Wald's paper are avoided in Le Cam (1956) by two devices. The first is to allow estimates T_n that satisfy asymptotic normality and sufficiency requirements but take values outside of Θ_n. The second is to relax the uniformity requirements for the convergence properties. Both devices create problems of their own as we shall see.

The properties called (A) to (E) above were not stated in this form by Wald. He does not use our function q_n at all but assumes existence of an underlying Euclidean norm that has special properties and can be used to define what is "local". He does use a property that is very close to the part of our (D) (E) that involves the kernels K_n' but not the part relative to the kernels K_n''. Instead of our sufficiency property (C) he uses a set transformation that attaches to each measurable set in the observation space a set in the maximum likelihood space that has, uniformly in θ, almost the same probability. This seems asking for too much. One can have sufficiency of a subfield **B** of the observation σ-field **A** even though the conditional expectation of an indicator I_A is not close to an indicator I_B. We have not checked the validity of Wald's Lemma 1 but can readily conceive of situations where the properties (A) to (E) are satisfied but the conclusion of Wald's Lemma 1 is doubtful.

The misbehavior of Wald's chi-square criterion noted by Hauck and Donner (1977) and by Vaeth (1985) can readily be understood by looking at the form of the function q_n^2 for the approximating HetG families. The lack of invariance is another matter. It is simply this. Write θ and T_n as functions of other variables, say $\theta = \phi(\xi)$ and $T_n = \phi(S_n)$. The differences $T_n - \theta$ will be approximated by $\dot{\phi}(\xi)(S_n - \xi)$ for small values of $S_n - \xi$. this will allow a form of local asymptotic invariance. However, for $S_n - \xi$ large, the differences $T_n - \theta$ may have no relation with $\dot{\phi}(\xi)(S_n - \xi)$.

This means that if the properties (A) to (E) are satisfied by an estimate T_n that is a maximum likelihood estimate and is uniquely defined, one need not worry about the lack of invariance of Wald's chi-square for very small deviations $T_n - \theta$ but that large deviations are a totally different matter. The fact that trouble occurs in such a simple transformation as the passage from the usual Binomial $\binom{n}{k} p^k (1-p)^{n-k}$ to its "natural" exponential family form $\binom{n}{k} e^{\theta k}[1 + e^{\theta}]^{-n}$ is a bit dispiriting perhaps. However note that the conditions (A) to (E), or Wald's original assumptions are clearly not satisfied by the Binomial family, with $p \in (0,1)$. They are not satisfied either for p restricted to an interval $[a,b] \subset (0,1)$ (whether the interval is taken closed or open). One can make the conditions hold by taking an interval $[a,b]$, with $0 < a < b < 1$ if one removes the restriction (B1) that the estimate T_n be strict, that is take values in $[a,b]$ only.

Some relations between the properties (A) to (E) and our proposal to use tests and confidence sets based in q_n^2 are as follows.

Proposition 1. *Let conditions (A) to (E) be satisfied. For* τ_n *as in condition* (A), *let* $K_{\tau_n,n} = M_{\tau_n,n}' M_{\tau_n,n}$. *Then for all* θ_n *such that* $q_n^2(\theta_n, \tau_n)$ *stays bounded the difference*

$$q_n^2(T_n, \theta_n) - (T_n - \theta_n)' K_{\tau_n,n} (T_n - \theta_n)$$

tends to zero in $P_{\theta_n,n}$ *probability.*

Proof. Let θ_n and t_n be points of Θ_n such that both $q_n(\theta_n, \tau_n)$ and $q_n(t_n, \tau_n)$ remain bounded. Consider the binary experiments $\mathbf{B}_n = \{P_{\theta_n,n}, P_{t_n,n}\}$, $\mathbf{B}_n' = \{F_{\theta_n,n}, F_{t_n,n}\}$ and $\mathbf{B}_n'' = \{G_{\theta_n,n}, G_{t_n,n}\}$ where $\mathbf{B}_n'$ and $\mathbf{B}_n''$ are from condition (D). We claim that the distances between these three experiments tend to zero. For $(\mathbf{B}_n, \mathbf{B}_n')$ this follows from condition (C). For $(\mathbf{B}_n', \mathbf{B}_n'')$ this is exactly the statement of condition (D) restricted to the pairs (θ_n, t_n).

Now introduce the half space distance $\| F_{\theta,n} - G_{\theta,n} \|_h = \sup_H | F_{\theta,n}(H) - G_{\theta,n}(H) |$ taken over all half-spaces $H \subset \mathbf{R}^k$. This distance is invariant under all affine transformations. Condition (E) implies that $\sup_\theta \| F_{\theta,n} - G_{\theta,n} \|_h \to 0$. If $q_n(\theta_n, \tau_n)$ and $q_n(t_n, \tau_n)$ remain bounded, condition (A) says that both $\mathbf{L}[M_{\tau_n,n}(T_n - \theta_n) | \theta_n]$ and $\mathbf{L}[M_{\tau_n,n}(T_n - t_n) | t_n]$ tend to the same $\mathbf{N}(0,I)$ limit. Thus the half space distance between $\mathbf{L}[T_n - \theta_n | \theta_n]$ and $\mathbf{L}[T_n - t_n | t_n]$ tends to zero. Recenter both $G_{\theta_n,n}$ and $G_{t_n,n}$ at zero, getting new normal distributions $G^*_{\theta_n,n}$ and $G^*_{t_n,n}$. It follows from the above that $\| G^*_{\theta_n,n} - G^*_{t_n,n} \|_h \to 0$. However this implies that the L_1-norm $\| G^*_{\theta_n,n} - G^*_{t_n,n} \|$ also tends to zero. Thus the experiments $\mathbf{B}_n'' = \{G_{\theta_n,n}, G_{t_n,n}\}$ are asymptotically equivalent to experiments $\mathbf{B}_n''' = \{G_{\theta_n,n}', G_{t_n,n}'\}$ where $G_{\theta_n,n}'$ and $G_{t_n,n}'$ are normal distributions centered respectively at θ_n and at t_n but with the same covariance matrices. The same argument applies to pairs such as (θ_n, τ_n). The asymptotic equivalence of these experiments implies that the difference between the affinity numbers $\rho_n = \int \sqrt{dP_{\theta_n,n} dP_{t_n,n}}$, $\rho_n' = \int \sqrt{dF_{\theta_n,n} dF_{t_n,n}}$ and so forth up to ρ_n''' all tend to zero. Similarly the difference between affinities $\rho_n = \int \sqrt{dP_{\theta_n,n} dP_{\tau_n,n}}$ and the corresponding number for pairs $(G_{\theta_n,n}' G_{\tau_n,n}')$ will tend to zero. It follows that since $q_n(\theta_n, \tau_n)$ and $q_n(t_n, \tau_n)$ remain bounded the numbers $q_n(\theta_n, t_n)$ also remain bounded. But then the difference between $\log \int \sqrt{dP_{\theta_n,n} dP_{t_n,n}}$ and $\log \int \sqrt{dG_{\theta_n,n}' dG_{t_n,n}'}$ must also tend to zero. It follows easily that $q_n^2(\theta_n, t_n) - (\theta_n - t_n)' K_{\tau_n,n}(\theta_n - t_n)$ must tend to zero. The result claimed for $q_n^2(T_n, \theta_n)$ follows then from condition (B). Hence the Proposition.

Although the above proof is devious the result is hardly surprising. Note that the proof made use of all the conditions (A) to (E). It is perhaps more surprising that Proposition 1 admits a partial converse.

Proposition 2. *Assume that conditions (A) and (B) hold and that, with the notation of Proposition 1,*

$$q_n^2(\theta_n, \tau_n) - (\theta_n - \tau_n)' K_{\tau_n,n}(\theta_n - \tau_n)$$

tends to zero for all pairs such that $q_n^2(\theta_n, \tau_n)$ *remains bounded. Then conditions (C) to (E) and also satisfied.*

Proof. Condition (A) and the approximation property for the q_n function imply that in sets of the type $V_n = \{\theta : q_n^2(\theta, \tau_n) \le b\}$ the statistics T_n are asymptotically sufficient and distinguished. A proof to that effect can be carried out as in Le Cam (1977). Condition (B) allows carrying out a patchwork argument as in Le Cam (1986), Chapter 11, Theorem 3. This gives (C). Statement (D) and (E) can then be obtained by a method similar to the method used in Le Cam (1986) Chapter 11, Section 8.

It follows from the above propositions that, under the conditions (A) to (E), one can asymptotically treat $q_n^2(T_n, \tau_n)$ as if it was, under τ_n, a central chi-square, just as would be the case for $(T_n - \tau_n)' K_{\tau_n, n}(T_n - \tau_n)$. If the distributions are induced by τ_n then $q_n^2(T_n, \theta_n)$ will behave as a non-central chi-square, as is the case for $(T_n - \theta_n)' K_{\tau_n, n}(T_n - \theta_n)$, at least if $q_n(\theta_n, \tau_n)$ remains bounded. If $q_n(\theta_n, \tau_n)$ tends to infinity condition (B) implies that $q_n(T_n, \theta_n)$ will tend to infinity in $P_{\tau_n, n}$ probability. This can be used as support of our proposal to use $q_n^2(T_n, \theta)$ to build confidence sets, or for testing purposes, instead of Wald's chi-square. The phenomenon noted by Hauck and Donner (1977) and Vaeth (1985) would still be possible for $(T_n - \theta_n)' K_{\theta_n, n}(T_n - \theta_n)$. That can tend to zero for suitable sequences θ_n while $q_n^2(T_n, \theta_n) \to \infty$. However, by condition (B), $q_n^2(T_n, \theta_n)$ cannot stay bounded in $P_{\tau_n, n}$ probability unless $q_n^2(\theta_n, \tau_n)$ remains bounded. Thus the two criteria may behave differently. The one based on q_n^2 seems more satisfactory if (B) holds.

There are, however, certain difficulties. One of them arises from the fact that Wald's conditions, and our conditions (A) to (E), are extremely restrictive. Part of this is due to the insistence on uniformity of the convergence on the entire Θ_n. To judge this appropriately one should note that most asymptotic papers do not bother about uniformity of convergence, much less about bounds. However, uniformity of convergence is something that happens "as n tends to infinity" and is therefore of little interest to the practitioner, except perhaps as psychological reassurance.

It is probably too much, in the present state of the art to ask for usable, computable bounds. Yet it is often feasible to look at auxiliary estimates and associated confidence regions that can be trusted to limit the possible range of values of the parameter θ to rather small subsets of the initial set Θ_n. If, in that restricted range, the approximations implied by our conditions (A) to (E) can be verified to hold reasonably, then we may proceed.

To repeat, suppose that 1) you have evidence that the model $\{P_{\theta_n, n}; \theta \in \Theta_n\}$ can fit the observations adequately and 2) you have some auxiliary "robust" estimate θ_n^* with known variability that says that the true parameter value must with high probability lie in a certain small subset $A_n \subset \Theta_n$. Then the validity of (A) to (E) on the whole of Θ_n is of little relevance. What matters is the adequacy of the approximations on A_n.

Thus, if one has just taken appropriate precautions and if auxiliary confidence sets that can be trusted say that one is in a region where the chi-square approximation to the distribution of $q_n^2(T_n, \theta)$ holds adequately, one may feel justified in using it.

A different kind of difficulty arises from the fact that to use $q_n^2(T_n, \theta)$ the value of T_n must be in the range where q_n^2 is defined. As argued above, it may be rather difficult near boundary points to check the validity of both (A) and (B1). It is often easier to dispense with (B1) and either extend the domain of definition of q_n^2 or replace the estimates T_n that would satisfy (A) to (E) except (B) by some other estimate T_n' for which the distribution of $q_n^2(T_n', \theta)$ is no longer chi-square but something that can be evaluated according to the geometry of the situation.

Further elaboration on this will be found in Section 5 below.

One obvious problem in using q_n^2 is that one has to evaluate it. In this respect let us note the following results, already imbedded in the proof of Proposition 1

Proposition 3. *Let the conditions (A) to (E) be satisfied. Let* G_n *be the heteroschedastic gaussian experiment of condition (D) and let* $g_n^2(s,t) = -8\log\int\sqrt{dG_{s,n}\,dG_{t,n}}$. *Then, for every* $\varepsilon > 0$ *there is an* $N(\varepsilon)$ *such that for* $n \geq N(\varepsilon)$ *and all pairs* (s,t) *one has either*

$$|q_n^2(s,t) - g_n^2(s,t)| < \varepsilon \quad \textit{or}$$

$$\min[q_n^2(s,t),\, g_n^2(s,t)] > 1/\varepsilon.$$

Indeed the difference between the affinities tend to zero uniformly as $n \to \infty$.

3. An example of M. Vaeth.

This section refers to the paper by M. Vaeth (1985) and in particular to the example discussed pages 205-206. Actually we shall not use the exact formulation of Vaeth but a simpler one that exhibits the same phenomenon but in terms of "exponential integrals" instead of Bessel functions.

For a fixed k let $f_k(x,\theta)$ be the density

$$f_k(x,\theta) \;=\; \frac{e^{-\theta x}}{F_k(\theta)}\,\frac{1}{x^k}; \quad x \geq 1,\; \theta > 0,$$

with respect to the Lebesgue measure on $[1,\infty)$. Here $F_k(\theta) = \int_1^\infty e^{-\theta x}\frac{1}{x^k}\,dx$ is the "exponential integral of order k" usually denoted $E_k(\theta)$. We shall use F_k instead of E_k to avoid possible confusion with expectations.

For such a family the following relations hold:

1) $F_{k+1}(\theta) = \frac{1}{k}\,[e^{-\theta} - \theta F_k(\theta)]$

2) $E_\theta X = \dfrac{F_{k-1}(\theta)}{F_k(\theta)}$

3) $E_\theta X^2 = \dfrac{F_{k-2}(\theta)}{F_k(\theta)}$

The maximum likelihood estimate $\hat\theta$ is the solution of the equation

$$X \;=\; \frac{F_{k-1}(\hat\theta)}{F_k(\hat\theta)} \;=\; E_{\hat\theta} X,$$

at least for $k \leq 3$. For $k > 3$ the range of $E_\theta X$ is limited. One has $E_\theta X \leq \frac{k-1}{k-2}$. Hence, for $X > \frac{k-1}{k-2}$ the m.l.e. $\hat\theta$ is equal to zero. Otherwise, if $k < 3$, the m.l.e. coincides with the estimate obtained by the method of moments.

The phenomenon discussed by Vaeth is as follows. Consider the parametrization by $\beta(\theta) = E_\theta X$ so that X is the m.l.e. of $\beta(\theta)$.

To test the hypothesis that $\theta = \theta_1$ or to build confidence intervals, Wald suggests the use of the expression $\frac{X - \beta(\theta_1)}{\hat\sigma}$ where $\hat\sigma$ is the m.l.e. of the standard deviation of X. For values of k such that $1 \leq k \leq 2$ this expression tends to zero as X tends to infinity. Thus large values of X, which tend to indicate values of θ close to zero, are held compatible with any value of θ. For $k > 1$, $1 < k \leq 2$ this is not too disturbing since the sequences $\{f_k(\cdot,\theta)\}$ and $\{f_k(\cdot,\theta_1)\}$ are contiguous as $\theta \to 0$. In fact $f_k(x,\theta)$ tends to $f_k(x,0) = \frac{k-1}{x^k}$, $x \geq 1$. For $k < 1$ the phenomenon in question does not occur: the coefficient of variation of X stays finite as $\theta \to 0$.

This can be easily checked by using the classical expansions of $F_k(\theta)$ for θ near zero. They can be found, for instance, in Abramowitz and Stegun (1964) or can be derived directly.

For k = 1 the m.l.e. of $\beta(\theta)$ is X itself. It has a variance

$$\operatorname{Var} X = \frac{F_{-1}(\theta)}{F_1(\theta)} - \left[\frac{F_0(\theta)}{F_1(\theta)}\right]^2 .$$

Now $F_0(\theta) = e^{-\theta}/\theta$ and for θ tending to zero $F_1(\theta)$ behaves like $-\log\theta - \gamma$ where γ is Euler's constant $\gamma = -\int_0^\infty e^{-y}\log y\,dy \sim .57$. Thus for small θ the variance of X will behave like

$$\frac{F_{-1}(\theta)}{F_1(\theta)} = \left\{\frac{1}{\theta^2}e^{-\theta} + \frac{1}{\theta}e^{-\theta}\right\}\frac{1}{F_1(\theta)}$$

$$\sim \frac{1}{\theta^2|\log\theta|}.$$

The maximum likelihood equation shows that, for large X, the m.l.e. $\frac{1}{\hat\theta}$ behaves like $X\log X$ so that the estimated standard deviation of X is of the order

$$\frac{1}{\hat\theta}\frac{1}{\sqrt{|\log\hat\theta|}} \sim X\sqrt{\log X},$$

hence the behavior of the criterion $\dfrac{X-\beta(\theta_1)}{\hat\sigma}$.

There is nothing particularly surprising about this fact. As $\theta \to 0$ the distribution of X is far from normal. Its expectation and standard deviation are poor indications of location and spread. For instance the median of X behaves like $1/\sqrt{\theta}$ while $E_\theta X$ behaves like $\dfrac{1}{\theta|\log\theta|}$. The α^{th} quantiles behave like $\theta^{-\alpha}$. The distribution of X cannot be "normalized" by a change of location and scale. The observed misbehavior of Wald's criterion extends to some other expressions. For instance if one uses an estimate $\bar\theta$ obtained by putting X equal to its median and then estimate the spread of the distribution by an interquartile range computed at $\bar\theta$ the resulting ratio will also tend to zero as $X \to \infty$.

All the arguments used above in this section use only one observation. If one has n independent identically distributed observations $X_1, X_2, \ldots, X_n$ their average $\overline{X}_n$ will still be the maximum likelihood estimate of $E_\theta X$. As explained by Vaeth (1985) the misbehavior noted for one observation persists for every value of n. Now let us see how the functions $q_n^2(T_n, \theta)$ of Section 2 can behave. Here $q_n^2(s,t) = nq^2(s,t)$ where q is the function computed for one observation only.

The argument of Section 2 depend on finding estimates T_n that are well behaved and in particular satisfy the condition (B) of Section 2. One can easily show that, here, the m.l.e. $\hat{\theta}_n$ satisfies condition (B) even though, as we shall see, it does not satisfy the other conditions of Section 2 uniformly on $\Theta = (0, \infty)$.

Proposition 4. *Let* $\{p_\theta;\ \theta \in (0,\infty)\}$ *be an exponential family of rank one in an arbitrary parametrization. Then for* n *independent identically distributed observations and for the m.l.e.* $\hat{\theta}_n$ *one has*

$$P_{\theta,n}\{q_n^2(\hat{\theta}_n, \theta) \geq 8z\} \leq 2e^{-z}$$

for all $z > 0$.

Proof. An exponential family in its natural parametrization has the form

$$p_\theta(dx) = \exp\{\theta x - A(\theta)\}\mu(dx)$$

for some measure μ. Thus

$$q^2(\theta,t) = 8\{\frac{1}{2}[A(\theta) + A(t)] - A[\frac{\theta + t}{2}]\}.$$

Since A is a convex function, for θ fixed $q^2(\theta,t)$ increases as $|\theta - t|$ increases. Consider any particular $t > \theta$ and the test based on n observations that minimizes the sum of probabilities of error for θ and t. This sum of probabilities of error is $\|P_{\theta,n} \wedge P_{t,u}\| \leq \exp\{-\frac{1}{8}q_n^2(\theta,t)\}$. However, by concavity of the logarithm of likelihood ratios, if the test in question rejects t, the test of θ against $t' > t$ will also reject t'. Thus, except for probability at most $\exp\{-\frac{1}{8}q_n^2(\theta,t)\}$ for $P_{\theta,n}$, one will reject all $t' \geq t$. The same argument applies to values $s < \theta$. Hence the result, since the inequality $q_n^2(\hat{\theta}_n, \theta) \geq 8z$ is invariant under all one to one reparametrizations.

Note the 8z in the expression in curly brackets of Proposition 4. If $q_n^2(\hat{\theta}_n, \theta)$ was actually chi-square one could replace it by 2 z for the same bound on the probabilities. Part of the loss can be attributed to the passage from $\|P_{\theta,n} \wedge P_{t,n}\|$ to Hellinger affinities but part may just be due to the fact that, here, nothing much is known about the distribution of $\hat{\theta}_n$ or $q_n^2(\hat{\theta}_n, \theta)$.

In the present specific example one can obtain a variety of results about the asymptotic behavior of q_n^2. Of course, if θ is kept fixed, independent of n, the variables $\sqrt{n}[\bar{X}_n - E_\theta(X)]$ will be asymptotically normal and, $\hat{\theta}_n$ being the m.l.e., $q_n^2(\hat{\theta}_n, \theta)$ will be asymptotically χ_1^2. If, on the contrary, the true θ is a θ_n that depends on n and tends to zero, the behavior of $q_n^2(\hat{\theta}_n, \theta)$ can be very different from chi-square. To investigate what can happen consider two sequences $\{s_n\}$ and $\{t_n\}$ both tending to zero and such that $s_n > t_n$. For n observations the Hellinger transform of the pair $\{P_{s_n,n}, P_{t_n,n}\}$ has a logarithm of the form

$$\phi_n(\alpha) = n \log \int [f(x, s_n)]^{1-\alpha} [f(x, t_n)]^{\alpha} dx$$
$$= n \{\log F[(1-\alpha) s_n + \alpha t_n] - (1-\alpha) \log F(s_n) - \alpha \log F(t_n)\}$$

where $F(v)$ is the exponential integral $F(v) = F_1(v) = \int_1^{\infty} e^{-vx} \frac{1}{x} dx.$

For small z it has the expansion

$$F(z) = |\log z| - \gamma + \sum_{j=1}^{\infty} a_j z^j,$$

and its logarithm has the expansion

$$\log F(z) = \log|\log z| + \log[1 - \frac{\gamma}{|\log z|} + \frac{1}{|\log z|} \Sigma a_j z^j].$$

Let us first look at the $\log|\log|$ term in the expansion. For $\phi_n(\alpha)$ they give a first term

$$\omega_n(\alpha) = n\{\log|\log(1-\alpha) s_n + \alpha t_n| - (1-\alpha)\log|\log s_n| - \alpha \log|\log t_n|\}.$$

To investigate the behavior of this we shall assume $t_n < s_n$ and let $t_n = (1-\xi_n) s_n$, $0 < \xi_n < 1$. Then $(1-\alpha) s_n + \alpha t_n = s_n[1 - \alpha \xi_n]$ and $|\log[(1-\alpha) s_n + \alpha t_n]|$ $= |\log s_n| - \log(1 - \alpha \xi_n)$.

Similarly $|\log t_n| = |\log s_n| - \log(1-\xi_n) = |\log s_n| \left\{1 - \frac{\log(1-\xi_n)}{|\log s_n|}\right\}$. This yields

$$\omega_n(\alpha) = n \left\{\log[1 - \frac{1}{|\log s_n|} \log(1 - \alpha \xi_n)] - \alpha \log[1 - \frac{\log(1-\xi_n)}{|\log s_n|}]\right\}.$$

To study this it is convenient to introduce the notation

$$\delta_n = \frac{1}{|\log s_n|}$$

so that

$$\omega_n(\alpha) = n\{\log[1 - \delta_n \log(1 - \alpha \xi_n)] - \alpha \log[1 - \delta_n \log(1 - \xi_n)]\}.$$

We shall distinguish three cases:

Case A, $n \delta_n \to \infty$. Then, for $\omega_n(\alpha)$ to stay bounded, ξ_n must tend to zero. In such a case one has $-\log(1 - \alpha \xi_n) \sim \alpha \xi_n + \frac{1}{2} \alpha^2 \xi_n^2$ and $-\log(1 - \xi_n) \sim \xi_n + \frac{1}{2} \xi_n^2$ and

$\omega_n(\alpha)$ behaves like

$$n \left\{ \log [1 + \delta_n(\alpha\xi_n + \frac{1}{2}\alpha^2\xi_n^2)] - \alpha \log [1 + \delta_n(\xi_n + \frac{1}{2}\xi_n^2)] \right\}.$$

Expanding the logarithms once more, one sees that the terms in $\delta_n \xi_n$ cancel. The expression remains bounded if $n\delta_n\xi_n^2$ remains bounded. If $n\delta_n\xi_n^2 \to \sigma^2$ the term $\omega_n(\alpha)$ tends to $\frac{1}{2}\sigma^2[\alpha^2 - \alpha]$. This is the logarithm of the Hellinger transform for a Gaussian experiment.

This suggests looking at a family $\mathbf{F}_n = \{Q_{\lambda,n}\}$ where $Q_{\lambda,n}$ is $P_{\theta,n}$ with a θ taken equal to $s_n + \lambda s_n \sqrt{\frac{|\log s_n|}{n}}$ with λ restricted so that $\theta > 0$. It can be shown that the experiments $\mathbf{F}_n$ converge to a Gaussian shift experiment linearly indexed by λ. Thus, with this parametrization, the corresponding $q_n^2(\hat{\lambda}_n, \lambda)$ will still behave asymptotically as chi-square, with one degree of freedom.

Case B, $n\delta_n \to b$, finite, positive. Then $\omega_n(\alpha)$ can stay bounded for values ξ_n that stay away from zero and unity. If $\xi_n \to \xi$, $0 < \xi < 1$ then $\omega_n(\alpha)$ tends to $-b[\log(1-\alpha\xi) - \alpha\log(1-\xi)]$. This shows that, under $P_{s_n,n}$, the distribution of $\log \frac{dP_{t_n,n}}{dP_{s_n,n}}$ tends to a shifted gamma distribution. The sequences are contiguous.

Case C, $n\delta_n \to 0$. In this case it is possible to let ξ_n tend to unity in such a way that $n\delta_n \log(1-\xi_n)$ stays bounded. If $-n\delta_n\log(1-\xi_n)$ tends to a limit b then $\omega_n(\alpha) \to -b\alpha$. This is the log Hellinger transform for a pair (Q_0, Q_1) where the part of Q_1 that is dominated by Q_0 has a constant density equal to e^{-b}. The part of Q_1 that is Q_0 singular has mass $1 - e^{-b}$. This implies that the sequence $\{P_{s_n,n}\}$ is contiguous to $\{P_{t_n,n}\}$ but the reverse is not true. Here $q_n^2(s_n, t_n)$ tends to 4 b and $\| P_{s_n,n} \wedge P_{t_n,n} \|$ tends to e^{-b}.

In the above derivations we have used only the log log term in the expansion of log F. However, it is easy to check that the other terms tend to zero.

In all cases, the logarithm of likelihood ratio $\Lambda_n = \log \frac{dP_{t_n,n}}{dP_{s_n,n}}$ has the form $\Lambda_n = a_n\overline{X}_n + b_n$ where a_n and b_n are constants and where $\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$. Since $\overline{X}_n$ is the maximum likelihood estimate of its expectation $\beta(\theta) = E_\theta \overline{X}_n$ the expression $q_n^2(\hat{\theta}_n, \theta)$ can also be written in terms of $\overline{X}_n$ and β as, say $\overline{q}_n^2(\overline{X}_n, \beta)$. Since, by Proposition 4, $\overline{q}_n^2(\overline{X}_n, \beta_n)$ remain bounded in $P_{\theta,n}$ probability no matter what $\beta_n = \beta(\theta_n)$ does, one can approximate $\overline{q}_n^2(\overline{X}_n, \beta_n)$ by the expressions used above for $-8\,\omega_n(1/2)$.

In case B this leads to an approximation of the type

$$-8\,b\log 2 + 8\,b\log\left\{\sqrt{\frac{\overline{X}_n}{\beta_n}} + \sqrt{\frac{\beta_n}{\overline{X}_n}}\right\}.$$

Since $\dfrac{\overline{X}_n - c_n}{\beta_n}$ is approximately distributed as a gamma variable with exponent b, this does not seem to behave like a chi-square. For case C the variables $\overline{q}_n^2(\overline{X}_n, \beta_n)$ seem to behave in the same manner as $\dfrac{n}{\log\beta_n}\log\dfrac{\overline{X}_n}{\beta_n}$. By Proposition 4 this must stay bounded in $P_{\theta_n, n}$ probability for $\beta_n = E_{\theta_n}X$. If θ_n is replaced by a $t_n = (1-\xi_n)\theta_n$ such that $\dfrac{n}{\log\beta_n}\log(1-\xi_n) \to b$, then $\dfrac{n}{\log\beta_n}\log\dfrac{\overline{X}_n}{\beta_n}$ will have for $P_{t_n, n}$ a distribution with a mass $1 - e^{-b}$ tending to infinity. This should be taken into account in the construction of confidence intervals.

4. Some heteroschedastic gaussian cases.

As seen in Section 2 heteroschedastic gaussian experiments occur routinely in asymptotic theory. In fact the conditions (A) to (E) of Section 2 provide for a situation where the experiments $E_n = \{P_{\theta,n};\ \theta \in \Theta_n\}$ are such that $\Delta(E_n, G_n) \to 0$ for the heteroschedastic experiment G_n of condition (D). For this reason we shall study here the behavior of some heteroschedastic gaussian experiments. However, for simplicity we shall only use parameters θ that run through an interval of the real line, say $\Theta = [a, \infty)$ where a is a large positive number.

Let X be a normal variable whose distribution depends on a parameter $\theta \in [a, \infty)$. Assume that, given θ, the variable X has expectation θ and a variance $\sigma^2(\theta) = \frac{1}{\gamma(\theta)}$.

The family so obtained defines an affinity

$$\rho(s,t) = \left\{\frac{4\gamma(s)\gamma(t)}{[\gamma(s)+\gamma(t)]^2}\right\}^{1/4} \exp\{-\frac{1}{4}\frac{\gamma(s)\gamma(t)}{\gamma(s)+\gamma(t)}|t-s|^2\}$$

yielding

$$\begin{aligned} q^2(s,t) &= -8\log\rho(s,t) \\ &= 2\log\left\{\frac{[\gamma(s)+\gamma(t)]^2}{4\gamma(s)\gamma(t)}\right\} + 2\frac{\gamma(s)\gamma(t)}{\gamma(s)+\gamma(t)}|t-s|^2. \end{aligned}$$

We shall be interested in situations where γ is a smooth decreasing function that tends to zero rapidly as $\theta \to \infty$. For a first example let us take $\gamma(\theta) = e^{-2\theta}$. Then

$$q^2(s,t) = 4\log\cosh(t-s) + \frac{e^{-(s+t)}}{\cosh(t-s)}(t-s)^2,$$

indicating that, for s and t large, the main contribution to $q^2(s,t)$ will arise from the first term. This is the term that takes into account the difference between the variances at s and t.

The negative of the logarithm of the likelihood function is

$$\frac{1}{2}(X-\theta)^2 e^{-2\theta} + \theta = \frac{1}{2}[X - 1/2\log v]^2 v^{-1} + \frac{1}{2}\log v,$$

where v is the variance $v = e^{2\theta}$. The maximum likelihood equation is

$$\exp\{2\hat\theta\} = (X-\hat\theta) + (X-\hat\theta)^2.$$

In terms of the variance v this becomes

$$\hat v = (X - \frac{1}{2}\log\hat v) + (X - \frac{1}{2}\log\hat v)^2,$$

showing that, for $|X|$ large, $\hat v$ will behave approximately like $X^2 + X$. Approximate solution of the likelihood equation shows that, for $|X|$ large, $\hat{\theta}$ (restricted to (a, ∞), a large) behaves approximately like $\log|X|$. Some standard methods of constructing confidence intervals can lead to very different results. The standard "equal tails" intervals with probability of coverage near .955 would be given by inverting the inequalities $\theta - 2e^{2\theta} \le X \le \theta + 2e^{2\theta}$. However, for $X \ge -(1 + \log 2)$ the lower barrier is ineffective. The resulting intervals would be half infinite, of the form $[c(X), \infty)$. A similar phenomenon occurs for X negative but $|X|$ large.

If, on the contrary, one uses intervals of the type $X - 2\hat{\sigma} \le \theta \le X + 2\hat{\sigma}$ where $\hat{\sigma}$ is estimated then the intervals would take the form $[a, c_1(X)]$ with an ineffective bounding for small values. For instance if one estimate θ by $\log|X|$ for $|X| \ge 1$ one would estimate $\hat{\sigma}$ by $|X|$ and get intervals of the type $X - 2|X| \le \theta \le X + 2|X|$. The lower bound is always negative and therefore ineffective since we assume $\theta \ge a$ with $a > 0$, large. As $|X| \to \infty$ these intervals produce an instance of the Hauck-Donner-Vaeth phenomenon. They accept any finite value of θ.

For confidence intervals based on the function q the situation is different. Let us take some estimate $\tilde{\theta}$. If $\tilde{v} = \sigma^2(\tilde{\theta})$ is large, the main contribution to $q(t, \tilde{\theta})$ will be $4 \log\cosh(t - \tilde{\theta})$. Thus the intervals will be given approximately by an inequality of the type $\{\theta : \cosh(\theta - \tilde{\theta}) \le e^{b/4}\}$, that is $|\theta - \tilde{\theta}| \le \cosh^{-1}(e^{b/4}) = c$.

For simplicity, let us use the crude estimate $\tilde{v} = X^2$ so that $\tilde{\theta} = \frac{1}{2} \log X^2 = \log|X|$ with $|X|$ assumed ≥ 1. Then we have intervals equivalent to $|\theta - \log|X|| \le c$. For θ very large these intervals have a probability of coverage about equal to $P\{\log|\xi| \le c\}$ for a ξ with a $N(0,1)$ distribution. Thus one can consider intervals obtained from a value c of the order of $\log 2$. Note that these have a fixed length as $|X| \to \infty$. This is in sharp contrast with the intervals obtained from chi-square type formula.

The situation described above is rather extreme in that one would not expect to encounter very often observations X whose standard deviation is an exponential function of their expectation. However, the same kind of analysis can be carried out for a variety of other cases.

One could object that, for the normal family $N(\theta, e^{2\theta})$ used above, there is nothing "asymptotic". However, the same kind of analysis will apply for each fixed n to a family of the type $\{N(\theta, \frac{1}{n} e^{2\theta}) ; \theta \ge 1\}$. This shows that for many densities $f(x, \theta)$ that are sufficiently smooth functions of θ, the analysis will apply for n i.i.d. observations $X_1, \ldots, X_n$ provided that the Fisher information decreases exponentially fast as $\theta \to \infty$.

5. The choice of estimates and the range of q_n.

As noted in Section 2, the use of confidence sets of the form $\{\theta : q_n^2(T_n, \theta) \le b\}$ must rely on an appropriate choice of estimates T_n. To avoid various complex problems, let us consider only the case where we have product measures $P_{\theta,n}$; $\theta \in \Theta_n$ obtained from n independent identically distributed observations, $X_1, \ldots, X_n$. A prime candidate for T_n is the maximum likelihood estimate $\hat{\theta}_n$. However, one cannot limit oneself to $\hat{\theta}_n$ for many reasons. The m.l.e. $\hat{\theta}_n$ is known to misbehave for very many common families $\{P_{\theta,n}\}$, for instance for families obtained from mixtures. In 1973, 1975 and 1986 we proposed estimates T_n obtained by a complex construction. Another construction was proposed in Birgé (1983). In both cases the estimates T_n have the following desirable property. Let $h^2(s,t) = \frac{1}{2}\int(\sqrt{dp_s} - \sqrt{dp_t})^2$ where p_θ is the distribution of individual X_j under θ. Then $E_\theta\, n\, h^2(\theta, T_n) \le CD(\tau_n)$ where C is a universal constant and where D is a metric dimension function for Θ_n and h, evaluated at some suitable number τ_n selected roughly so that $C_1\, n\tau_n^2 = D(\tau_n)$ for a coefficient C_1 approximately equal to 64. Since the relation between q_n and h is given by $q_n^2(s,t) = -8\, n \log[1 - h^2(s,t)]$ one sees that for sets Θ_n that have bounded metric dimension one will be able to assert that $q_n^2(T_n, \theta)$ will remain bounded in $P_{\theta,n}$ probability uniformly in θ. Thus one may contemplate the use of such estimates.

The arguments of Le Cam (1986), Chapter 16, also give bounds for the probabilities $P_{\theta,n}\{n h^2(T_n, \theta) \ge c\}$ that can be converted into inequalities for $P_{\theta,n}[q_n^2(T_n, \theta) \ge b\}$. However this gives only inequalities and not approximate values for the probabilities in question.

The arguments of Le Cam (1986), Chapter 16, can be applied to i.i.d. sequences $X_{1,n}, \ldots, X_{n,n}$ whose individual distributions, say $p_{\theta,n}$, are the same for all $X_{j,n}$ $j = 1, \ldots, n$, but may be allowed to depend on n. In some cases (see Le Cam (1986), Chapters 16 and 17) one can use estimates based on empirical cumulatives to the same effect but with the same or similar difficulties relative to the approximate evaluation of probabilities. Since Wald's construction was intended for the case of asymptotically normal families, it would be pleasant to find estimates T_n that are such that $q_n^2(T_n, \theta)$ is not only bounded in $P_{\theta,n}$ probability uniformly for $\theta \in \Theta_n$ but also asymptotically distributed as a chi-square. As already noted in Section 2 the combination of the conditions (A) (= asymptotic normality for T_n and (B1) (that $T_n \in \Theta_n$) is not likely to be satisfied at boundary points of the closure of Θ_n.

In such cases one may want to extend the domain of definition of the function q_n. There are situations where an extension is immediate. For an instance consider the binomial $\mathbf{B}(n,p)$ with p restricted to lie in an interval $[a,b] \subset (0,1)$. There is then some $\varepsilon > 0$ such that $0 < a - \varepsilon < b + \varepsilon < 1$. One can then let the estimate T_n take values in $(a - \varepsilon, b + \varepsilon)$ and use the function q_n relative to the binomial.

Such extensions raise the question of whether one can extend the definition of the family $\{P_{\theta,n}; \theta \in \Theta_n\}$ to a larger set $\Theta_n^* \supset \Theta_n$ and let T_n take values in Θ_n^*. The problem is essentially the same as the following: Let ϕ be a map from $\Theta \subset \mathbf{R}^k$ to a Hilbert space $\mathbf{X}$. Let $\Theta^* \supset \Theta$, when can one extend ϕ to Θ^* in a reasonable way and what is a reasonable way? In many cases the map will be uniformly continuous on Θ. Then it extends to a uniformly continuous map in the closure $\overline{\Theta}$ of Θ. To extend it further one could take for $\theta \notin \Theta$ a $\tau(\theta) \in \overline{\Theta}$ closest to θ and let $\phi(\theta)$ be $\phi[\tau(\theta)]$. However, this need not yield a continuous map. Other possibilities need to be investigated.

Another form of extension is discussed in some detail in Le Cam (1986), Chapter 11, It is as follows: One ignores totally the Euclidean structure of Θ_n including the fact that $\mathbf{R}^k$ is a vector space. Instead one uses a space $\mathbf{M}_0(\Theta_n)$ defined as the space of finite signed measures that are such that $\mu(\Theta_n) = 0$ and have finite support on Θ_n. For each $\theta \in \Theta_n$ one defines on $\mathbf{M}_0(\Theta_n)$ a positive quadratic form $\Gamma_{\theta,n}$ obtained from a "best fitting" Gaussian shift experiment to $\mathbf{E}_{n,\theta} = \{P_{t,n}; q_n^2(t,\theta) \le b_n\}$. Thus, for each θ one obtains an approximation $\Gamma_{\theta,n}(\delta_t - \delta_s)$ to $q_n^2(t,s)$ for $q_n^2(t,\theta) \le b_n$ and $q_n^2(s,\theta) \le b_n$. This approximation is then extended to all of $\mathbf{M}_0(\Theta_n)$. Under suitable circumstances (described in Le Cam (1986) Chapter 11) that involve finite dimensional restrictions locally on $\mathbf{M}_0(\Theta_n)$ normed by $\to \sqrt{\Gamma_{\theta,n}(\mu)}$ and existence of appropriate estimates, one can form centerings Z_n with values in the space $\mathbf{M}_1(\Theta_n)$ of finite signed measures μ with finite support such that $\mu(\Theta_n) = 1$. These will be asymptotically normally distributed and $\Gamma_{\theta,n}(Z_n - \delta_\theta)$ will be asymptotically distributed like a chi-square. Use of the expressions $\Gamma_{\omega,n}(Z_n - \delta_\theta)$ where ω is an estimate of θ can lead to the phenomenon described by Vaeth (1985) and Hauck and Donner (1977), but here one can take precautions since the estimates Z_n and ω need not be or cannot be the same. This suggests, however, that it may be preferable to use the formula given in Section 2 for the heteroschedastic Gaussian case. Here it would take a form of the type

$$g_n^2(\lambda,\mu) = -2\log\det[I - (M^{-1}\Delta)^2] + \overline{\Gamma}_n(\lambda - \mu)$$

where M corresponds to $\frac{1}{2}(\Gamma_{\omega,n} + \Gamma_{\theta,n})$ and Δ to $\frac{1}{2}(\Gamma_{\omega,n} - \Gamma_{\theta,n})$. The $\overline{\Gamma}_n$ would correspond to $(M - \Delta M^{-1}\Delta)$. This can be made to have a perfectly well defined meaning with operators on the Hilbert space completion of $\mathbf{M}_0(\Theta_n)$. (See Appendix).

There are occasionally other possibilities. For instance in the i.i.d. case of observations $X_1, \ldots, X_n$ with individual distributions $\{p_\theta : \theta \in \Theta_n\}$ or $\{p_{\theta,n}; \theta \in \Theta_n\}$, one can use the square distance $H_n^2(s,t) = n\,h_n^2(s,t)$ with $h_n^2 = 1 - \int \sqrt{dp_{s,n}\,dp_{t,n}}$. One can then use on $M_0(\Theta_n)$ the quadratic form Γ_n defined by $\Gamma_n(\mu) = -4\int\int H_n^2(s,t)\,\mu(ds)\,\mu(dt)$. According to Le Cam (1986) Chapter 11, Section 6 page 265 s.q.q. this gives a positive global substitute for the local $\Gamma_{\theta,n}$. Then one can use approximation of $\mathbf{E}_n = \{P_{\theta,n};\ \theta \in \Theta_n\}$ by a Gaussian shift experiment in which the Hauck-Donner-Vaeth cannot occur. Note however that this procedure imbeds Θ_n into a Hilbert space that can be infinite dimensional.

It is tempting to use instead of the Γ_n defined above a Γ_n^* defined by $\Gamma_n^*(\mu) = -\frac{1}{2}\int\int q_n^2(s,t)\,\mu(ds)\,\mu(dt)$. However, there are difficulties due to the fact that $\Gamma_n^*(\mu)$ may be negative. Also, if q_n can take infinite values, $\Gamma_n^*(\mu)$ may fail to be defined. Under the conditions (A) to (E) of Section 2 the use of Γ_n appears to be safe. However since part of the purpose of introducing the centers Z_n and the quadratic form Γ_n is to avoid the restriction (B1) of Section 2, one should consider conditions weaker than (A) to (E). Here is a possible set of conditions. The square norm Γ_n defines a metric or pseudo-metric on $M_0(\Theta_n)$. For measures m carried by $M_0(\Theta_n)$ (or its completion) let $\|m\|_{D_n}$ be the corresponding dual Lipschitz norm

(A^*) For any sequence $\{\tau_n\}$, $\tau_n \in \Theta_n$, if $\Gamma_n(\delta_{\theta_n}, \delta_{\tau_n})$ remains bounded, then the dual Lipschitz distance between $L(Z_n - \delta_{\theta_n} | \theta_n)$ and a centered normal distribution $G_{\theta_n,n}$, with inverse covariance form Γ_n tends to zero as $n \to \infty$.

(B^*) For sequences $\{\tau_n\}$ and $\{\theta_n\}$ as above $\Gamma_n[Z_n - \delta_{\theta_n} | \theta_n]$ stays bounded in $P_{\theta_n,n}$ probability.

(C^*) There is an integer k with the following property: Let $\{\tau_n\}$ be as in (A^*). Then there are linear subspaces $H_n(\tau_n) \subset M_0(\Theta_n)$ of linear dimension at most equal to k, such that, for any sequence θ_n such that $\Gamma_n(\delta_{\theta_n} - \delta_{\tau_n})$ stays bounded, there are elements $\mu_n \in H_n(\tau_n)$ with the following properties:

1) $\Gamma_n(\delta_{\theta_n} - \delta_{\tau_n} - \mu_n]$ tends to zero as $n \to \infty$

2) The L_1-norms $\|\mu_n\|$ stay bounded.

If these conditions are satisfied the distance $\Delta(\mathbf{E}_n, \mathbf{G}_n)$ between $\mathbf{E}_n = \{P_{\theta,n};\ \theta \in \Theta\}$ and the Gaussian shift experiment attached to Γ_n (with parameters restricted to δ_θ, $\theta \in \Theta_n$) will tend to zero.

The conditions given above yield uniformity of all convergences, as needed for Wald's properties (A) to (E) described in Section 2. There are many examples that would satisfy (A) to (E) but with uniformity of convergence only on certain subsets of the parameter spaces Θ_n. This create problems, but, at least in certain cases, the problems can be more of an aesthetic mathematical nature than a real practical impediment.

The real problem is that the convergences occur "as n tends to infinity" while the practical questions occur for fixed finite n. What seems to be a reasonable procedure is as follows: Proceed in two steps. First narrow down the range of possible values of θ by a trustworthy method that may be far from optimal. Then check that in the range in question the approximations by Gaussian shift experiments implicit in (A^*), (B^*), (C^*) above are reasonably good.

To give an example let us take Vaeth situation as described in Section 3. That is, assume that one observes $X_1, \ldots, X_n$ independent and all distributed according to a density

$$f(x,\theta) = \frac{e^{-\theta x}}{F_1(\theta)} \frac{1}{x}, \quad x \geq 1, \ \theta > 0.$$

Let us assume that n is very substantial, say n = 1600. Then one can construct an empirical cumulative distribution function by $H_n(x) = \frac{1}{n} \sum_{j=1}^{n} I(X_j \leq x)$. Let $H(x,\theta)$ be the cumulative corresponding to $f(\cdot,\theta)$.

Define a restricted range for θ by the recipe $R_n = \{\theta : \sqrt{n} \sup_x |H_n(x) - H(x,\theta)| \leq 4\}$. For any θ_0 the probability under $P_{\theta_0,n}$ that R_n does not contain θ_0 is of the order of 10^{-13}. This does not tend to zero as n tends to infinity but is close enough to zero for many and perhaps most practical purposes. For n = 1600 this limits θ to a range $R_n = \{\theta : \sup_X |H_n(x) - H(x,\theta)| \leq \frac{1}{400}\}$.

For the densities $f(x,\theta)$ in question the Kolmogorov distance is just one half of the L_1 distance between densities. Since this is larger than the square Hellinger distance one sees that two values s and t that belong to R_n must be such that $h^2(s,t) = \frac{1}{2} \int |\sqrt{f(x,s)} - \sqrt{f(x,t)}|^2 dx \leq \frac{1}{200}$ giving for the one observation value of $-8 \log [1 - h^2(s,t)] = q^2(s,t)$ a bound approximately equal to 1/25. This restricts θ to an interval $[l_n, r_n]$ that cannot be very long unless r_n happens to be very large.

The first thing to do is to check that R_n is not empty. If R_n is empty, as will happen often, the modelling by the densities $f(\cdot,\theta)$ is not credible and one should rethink the whole problem anew. If R_n is not empty and if one is sufficiently convinced of the validity of the model, one might want to try the inequality of Section 3 that says that for the maximum likelihood estimate $\hat{\theta}_n$ one has $P_{\theta,n}\{n q^2(\hat{\theta}_n, \theta) \geq 222\} \leq 10^{-12}$. Here this may not narrow the possible range of θ materially.

The next step would be to check whether on R_n a Gaussian shift approximation seems appropriate. Now one can try Gaussian shift approximations whose random term is linear in the parameter θ itself or in the expectation $\beta(\theta) = E_\theta X$ or in any other parametrization that seem suitable. To check the reasonableness of a linear parametrization in θ one would just have to check that in the range $\theta \in R_n$ the variance $\mathrm{Var}_\theta X$ does not change too much. Since we have not defined "reasonable" or "too much" this cannot be taken to make precise mathematical sense, but the idea should be clear.

One may be tempted to use the parametrization proposed by E. Mammen (1985). For this, let l_n be the left extremity of R_n and let r_n be the right one.

One could use the parameter $\xi(t) = q(t, l_n)$ for $t \in [l_n, r_n]$. This would not affect at all the approximability of a pair $(P_{s,n}, P_{t,n})$ by a Gaussian experiment $[G_{s,n}, P_{t,n}]$ but would affect the approximability of $\{P_{t,n}; t \in [l_n, r_n]\}$ by a Gaussian experiment whose log likelihood has a leading random term linear in $\xi(t)$ instead of t. It also would affect the construction of chi-square type statistics. Unless one uses our proposed $n q^2(\hat{\theta}_n, \theta)$ formula, one may want to look at expression of the type $n(\hat{\theta}_n - \theta)^2 J(\hat{\theta}_n)$, for the Fisher information J computed on one observation, or the formula $n[\overline{X}_n - \beta(\theta)]^2 / \mathrm{Var}_{\hat{\theta}_n} X$ or a formula of the type $n[q(\hat{\theta}_n, l_n) - q(\theta, l_n)]^2 C(\hat{\theta}_n)$ with an appropriate coefficient $C(\hat{\theta}_n)$.

In the final analysis there does not seem to be any possibility of escaping a check on the reasonableness of the chi-square distribution approximation for the expression retained. However it needs to be checked only for θ in the range R_n.

6. Reduction to the homoschedastic case.

We have just seen, in Section 5 above, that one can occasionally imbed the parameter space Θ_n into a Hilbert space and obtain approximations by Gaussian shift experiments. This is what is done as a matter of routine, when $\Theta_n \subset \mathbb{R}$, by variance stabilizing transformations. Then Θ_n remains a subset of the line $\mathbb{R}$. Vaeth (1985) shows that such transformations will prevent the misbehavior of Wald's criterion at least for exponential families.

When Θ_n is a subset of $\mathbb{R}^k$, $k > 1$, the situation is vastly more complex.

Under the conditions (A) to (E) of Section 2 one can approximate the experiments $E_n = \{P_{\theta,n}; \theta \in \Theta_n\}$ by heteroschedastic Gaussian ones, say $G_n = \{G_{\theta,n}; \theta \in \Theta_n\}$. Such an approximation yields for each θ a quadratic form on $\mathbb{R}^k$, using as a matrix the inverse covariance matrix at θ. Thus we obtain on Θ_n a structure of a Riemannian manifold. A transformation that leads to a Gaussian shift experiment amounts to an isometric imbedding of this Riemannian space into a Euclidean space $\mathbb{R}^m$. Now there are theorems that indicate the possibility of such embedding: E. Cartan says that such manifolds of dimension k can locally be imbedded in $\mathbb{R}^m$ for $m = \frac{1}{2} k(k+1)$. This is a local result. There is a global result of Nash that says that the entire manifold, if it is of class C^3, can be isometrically embedded in $\mathbb{R}^m$ for $m = \frac{1}{2}[3k^3 + 14k^2 + 11k]$. (See for instance J.T. Schwartz (1969) page 43 and J. Dieudonné (1971) volume 4 page 341.)

Unfortunately, the dimension m of the imbedding space is very much larger than that of the original $\Theta_n \subset \mathbb{R}^k$. Thus Θ_n becomes a very thin subset of $\mathbb{R}^m$.

The problem of stabilization of covariances is also related to the problem of selection of a ''best'fitting'' local Gaussian shift experiment. E. Mammen (1985) argues that, for one-dimensional exponential families around a point θ_0, the optimal choice is one in which the expectation $\xi(t)$ of the normal approximation is such that

$$q_n^2(\theta_0, t) = |\xi(t) - \xi(\theta_0)|^2$$

(if the standard deviation is taken equal to unity). We do not know of extension of such results to multidimensional families, but they may involve the same kind of difficulty as the embedding of Riemannian manifolds described above.

When feasible, the Hilbert space imbedding by $\Gamma_n(\mu) = -4 \int\int H_n^2(s,t)\,\mu(ds)\,\mu(dt)$ described above seems simpler, although perhaps not optimal.

7. Conclusion.

An examination of Wald's asymptotic tests or confidence intervals shows that they can exhibit distressing behavior. This is partly due to differential approximations that replace the expression $q^2(s,t) = -8 \log \int \sqrt{dP_s\, dP_t}$ by approximations that are quadratic in $(s-t)$. That can be justified only within the confines of sets where a good approximation to the log likelihood of the type

$$\log \frac{dP_t}{dP_s} \sim (t-s)W - \frac{1}{2}(t-s)'\Gamma(t-s)$$

holds for a constant matrix Γ. For larger sets one needs to take into account the variability of the matrices Γ.

Using instead confidence sets of the form $\{\theta : q_n^2(T_n, \theta) \leq c\}$ one can avoid many of the difficulties provided suitable estimates T_n can be constructed. If asymptotic normality of the estimates cannot be achieved with T_n that take values in the parameter space, palliatives are sometimes available.

One should also refrain from expecting too much from asymptotics. In the case discussed by Hauck and Donner there were a total of 455 observations. The analysis used effectively classifies them into 512 different boxes. This does not allow enough observations per box to induce faith in the applicability of asymptotic theory. It also can result in confusion over what the data try to say and what they are made to say by features that have been inserted, inadvertently perhaps, in the mathematical formulas. The use of multiple chi-square like tests under such circumstances can be hard to defend, for obvious reasons and for the added reason that the tests in question are based on a formula that does not measure adequately the difference between the postulated distributions at different values of the parameters.

Appendix
The affinity for two Gaussian measures

In Section 2 and Section 4 we have used a formula for the Hellinger affinity $\int \sqrt{dG_1 \, dG_2}$ between two Gaussian measures that may have different means and covariance matrices. The formula goes back at least to C.H. Kraft (1955). However, in the published version of Kraft's thesis it was misprinted and is barely recognizable. The derivation of the formula is not difficult but its implications are many. Thus we shall present here two derivations. One is intended for finite dimensional situations. The other is meant for arbitrary dimensions.

Let P and Q be two gaussian measures on a finite dimensional vector space $\mathbf{R}^k$. Assume that P has expectation θ and that the inverse of its covariance matrix is Γ so that it has a density

$$\frac{(\det \Gamma)^{1/2}}{(2\pi)^{k/2}} \exp\{-\frac{1}{2}(x-\theta)'\Gamma(x-\theta)\}$$

with respect to the Lebesgue measure of $\mathbf{R}^k$. Similarly, let Q have expectation t and inverse covariance matrix K.

Multiplying the square root of the densities will yield, in the exponent, a quadratic form that, except for the coefficient $(-\frac{1}{4})$, is equal to

$$(x-\theta)'\Gamma(x-\theta) + (x-t)'K(x-t).$$

To reduce this to a tractable form assume that Γ and K are both invertible and introduce a centering v by

$$(\Gamma + K)v = \Gamma\theta + Kt.$$

Then the above quadratic becomes

$$(x-v)'(\Gamma+K)(x-v) + (v-\theta)'\Gamma(v-\theta) + (v-t)'K(v-t).$$

The term $(v-\theta)$ can be expressed as $(v-\theta) = (\Gamma+K)^{-1}K(t-\theta)$. Similarly $(v-t) = (\Gamma+K)^{-1}\Gamma(\theta-t)$.

Let M be the matrix $M = \frac{1}{2}(\Gamma+K)$ and let $\Delta = \frac{1}{2}(\Gamma-K)$ so that $\Gamma = M + \Delta$ and $K = M - \Delta$. Then $(v-\theta)$ takes the form $(v-\theta) = \frac{1}{2}(t-\theta) - \frac{1}{2}M^{-1}\Delta(t-\theta)$ and $v - t$ is obtained by changing Δ to $-\Delta$. Then the sum of the two terms of the quadratic that do not involve $(x-v)$ will yield a quadratic in $(\theta - t)$ with a matrix $\frac{1}{4}\{(I-M^{-1}\Delta)'\Gamma(I-M^{-1}\Delta) + (I+M^{-1}\Delta)K(I+M^{-1}\Delta)\}$. Direct computation shows this to be equal to $\frac{1}{2}[M - \Delta' M^{-1}\Delta]$. Thus the term in the exponent of the square root of the product of the densities is

$$-\frac{1}{4}(x-v)'(\Gamma+K)(x-v)-\frac{1}{8}(t-\theta)'[M-\Delta' M^{-1}\Delta](t-\theta).$$

Now integrate out the variable x. One will get a term equal to $[\det\frac{(\Gamma+K)}{2}]^{-1/2}$ multiplied by $\exp\{-\frac{1}{8}(t-\theta)'[M-\Delta' M^{-1}\Delta](t-\theta)\}$. Combining this with the determinantal coefficients of the densities yields an affinity equal to

$$\frac{[(\det\Gamma)(\det K)]^{1/4}}{[\det M]^{1/2}}\exp\{-\frac{1}{8}(t-\theta)'[M-\Delta' M^{-1}\Delta](t-\theta)\}$$

and the result quoted in Section 2 follows by using the fact that the determinant of a product of two matrices (square of same order) is the product of the determinants. This shows that $\int\sqrt{dP\,dQ}$ is a product of two terms. One of them involves differences between expectations in the form

$$\exp\{-\frac{1}{8}(t-\theta)'[M-\Delta' M^{-1}\Delta](t-\theta)\}.$$

The other involves only the inverse of the covariance matrices in the form

$$\{\det[I-M^{-1}\Delta)^2]\}^{1/4}.$$

For many uses this determinantal form is not convenient. A form using the covariances themselves can also be used. Let $A=\Gamma^{-1}$ be the covariance matrix of P and let $B=K^{-1}$ be the corresponding matrix for Q. The determinantal term in $\int\sqrt{dP\,dQ}$ can also be written

$$[\det A]^{-1/4}[\det B]^{-1/4}\{\det[\frac{A^{-1}+B^{-1}}{2}]\}^{-1/2}.$$

Its fourth power is

$$\{[\det A][\det B]\det\{[\frac{A^{-1}+B^{-1}}{2}]^2\}\}^{-1} = \frac{\det AB}{\det(\frac{A+B}{2})^2}.$$

Write $S=\frac{1}{2}(A+B)$, $D=\frac{1}{2}(A-B)$. Then $A=S+D$, $B=S-D$ and $AB=S^2-D^2$. Thus the determinantal term can also be written as fourth root of

$$\det[I-(S^{-1}D)^2].$$

This form is particularly convenient for passage to infinite dimensions. To look at such a case, take a vector space V over the real numbers. Let X and Y be two processes $v \to X(v)$ and $v \to Y(v)$ indexed by V and linear in v. Assume that $X(v)$ is Gaussian with expectation zero and variance $E|X(v)|^2=\tilde{A}(v)$. Similarly let

$E\,Y(v) = 0$ and $E\,|\,Y(v)\,|^2 = \tilde{B}(v)$. These are squares of Hilbertian seminorms on V. The processes X and Y yield distributions that can be represented by measures P and Q on the algebraic dual of V.

If V contains sequences $\{v_n\}$ such that $\tilde{A}(v_n) + \tilde{B}(v_n)$ stays bounded away from zero but $\min[\tilde{A}(v_n), \tilde{B}(v_n)] \to 0$ then P and Q are obviously disjoint. Thus if $\int \sqrt{dP\,dQ} = \rho(P,Q) > 0$ the two seminorms $\tilde{A}^{1/2}$ and $\tilde{B}^{1/2}$ must be equivalent in the sense that there exists numbers a, b, $0 < a \le b < \infty$ such that $a\tilde{A}(v) \le \tilde{B}(v) \le b\tilde{A}(v)$ for all v. This shows that there will be no loss of generality in assuming that if $\tilde{S} = \frac{1}{2}(\tilde{A} + \tilde{B})$ then $\tilde{S}(v) = 0$ only at $v = 0$. Also one can assume V complete for the norm $\tilde{S}^{1/2}$ so that $(V, \tilde{S}^{1/2})$ is a Hilbert space. Let **X** be the dual of V for the norm $\tilde{S}^{1/2}$. It is clear that the inner product $[\cdot|\cdot]_A$ defined by $\tilde{A}$ on V can be represented as $[u|v]_A = \langle u, Av \rangle$ where the bracket $\langle u, x \rangle$ is the evaluation of the linear function $x \in \mathbf{X}$ at $u \in V$ and A is a linear map of V onto **X** such that $\langle u, Av \rangle = \langle v, Au \rangle$. Similarly, the inner product corresponding to B can be written $\langle u, Bv \rangle$ and the inner product defined by S is $\langle u, Sv \rangle$ where S is the canonical identification of the Hilbert space $(V, \tilde{S}^{1/2})$ with its dual. The inverse S^{-1} of that identification map sends **X** onto V. We shall also denote the norms of $(V, \tilde{S}^{1/2})$ and of its dual by the symbols $\|v\|$, so that $\|v\|^2 = \tilde{S}(v)$.

Consider then a finite dimensional subspace H of V. For the processes X and Y restricted to H we have distributions P_H and Q_H. It is clear that $\int \sqrt{dP_H\,dQ_H} \ge \int \sqrt{dP\,dQ}$.

Let Π be the orthogonal projection of V onto H in the Hilbert space $(V, \tilde{S}^{1/2})$. Let Π^t be the transpose of Π on the dual space **X** of V. One can show that $\Pi^t S\Pi = S\Pi$. Indeed, let $H^0 = \{x : \langle v, x \rangle = 0, \text{ all } v \in H\}$ be the polar of H in **X**. For any $y \in \mathbf{X}$ one has $\langle (1 - \Pi)v, y \rangle = \langle v, (1 - \Pi)^t y \rangle$. Thus if $v \in H$ $\langle v, (I - \Pi)^t y \rangle = 0$ and therefore $(I - \Pi)^t y \in H^0$. The defining relation for H^0 can also be written $\langle \Pi v, x \rangle = 0 = \langle v, \Pi^t x \rangle$ for all $v \in V$. This means that if $x \in H^0$ then $\Pi^t x = 0$. Therefore $H^0 = (I - \Pi)^t \mathbf{X}$.

Now take any $v \in V$ and consider $\langle w, S\Pi v \rangle = \langle w, \Pi^t S\Pi v \rangle + \langle W, (1 - \Pi)^t S\Pi v \rangle$. The second term on the right is equal to $\langle (I - \Pi)w, S\Pi v \rangle = [(I - \Pi)w | \Pi v]$ where $[\cdot|\cdot]$ is the inner product corresponding to $\tilde{S}$ on V. Thus $\langle w, S\Pi v \rangle = \langle w, \Pi^t S\Pi v \rangle$ for all v, implying $S\Pi = \Pi^t S\Pi$ and $SH = \Pi^t \mathbf{X}$. Let $[\cdot|\cdot]_A$ be the inner product defined on V by $\tilde{A}$. The map A from V to **X** is such that $\langle u, Av \rangle = [u|v]_A$ for all pairs (u, v) of elements of V. By the same argument there is also a map A_H from H to the space $H' = SH = \Pi^t \mathbf{X}$ such that $\langle u, A_H v \rangle = [u|v]_A$ for all pairs (u,v) of elements of H.

This gives $< u, A_H v > = < u, Av >$ for all such pairs. Equivalently $< \Pi u, A_H v > = < \Pi u, Av >$ for all $u \in V$ and $v \in H$. Therefore $A_H v = \Pi^t Av$. Defining B_H in a similar manner, we get a difference $D_H = \frac{1}{2}(A_H - B_H)$.

(Note that AH need not be in $SH = H'$. An example is given by the matrices $A = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$ on the plane). Considering P_H and Q_H as measures on $H' = SH$ the formula derived above shows that $[\rho(P_H, Q_H)]^4 = \det[I - (S^{-1} D_H)^2] \geq \rho^4(P, Q) = \beta$, say. Let m be the dimension of H and let $\lambda_1, \lambda_2, \ldots, \lambda_m$ be the eigenvalues of $S^{-1} D_H$. Since $\prod_{j=1}^{m} (1 - \lambda_j^2) \geq \beta$ and since $|\lambda_j| \leq 1$ one must have $0 \leq \lambda_j^2 \leq 1 - \beta$ and $\sum_{j=1}^{m} \lambda_j^2 \leq -\log \beta$. However $\sum_{j=1}^{m} \lambda_j^2 = \sum_{j=1}^{m} \| S^{-1} D_H u_j \|^2$ for any orthogonal base $\{u_j; j = 1, \ldots, m\}$ of H. since S^{-1} preserves the norms, this also means that $\sum_{j=1}^{m} \| D_H u_j \|^2 \leq -\log \beta$.

This is true for any H and any orthogonal sequence $\{u_1, u_2, \ldots, u_m, \ldots\}$ in V yielding $\sum_{j=1}^{\infty} \| D u_j \|^2 \leq -\log \beta$. Thus, if P and Q are not disjoint, D and $S^{-1} D$ are Hilbert-Schmidt operators and $(S^{-1} D)^2$ is an operator with finite trace.

A consequence of this is that one can find in V a basis $\{u_j; j \in J\}$ that is orthogonal for the norm $\tilde{S}^{1/2}$ and also for the norms $\tilde{A}^{1/2}$ and $\tilde{B}^{1/2}$. To obtain it, let u_1 be such that $\tilde{A}(u_1)$ is maximized subject to $\| u_1 \| \leq 1$. This is the same problem as maximizing $\frac{1}{2}[\tilde{A}(u) - \tilde{B}(u)]$ subject to $\frac{1}{2}[\tilde{A}(u) + \tilde{B}(u)] \leq 1$. Equivalently again, subject to the same condition, we are to maximize $< u, Du >$. Since D is a compact operator, there does exist a u_1 that achieves the maximum. If $u_1, \ldots, u_n$ have been determined, one selects u_{n+1} to maximize $\tilde{A}(u)$ subject to $\| u \| \leq 1$ among those u's that are orthogonal to $u_1, \ldots, u_n$. For this basis $S^{-1}D$ is represented by a diagonal matrix and there is no difficulty in writing the determinant of $I - (S^{-1} D)^2$ as a product $\prod_j \{(1 - \lambda_j^2); j \in J\} = \rho^4(P, Q)$.

A corollary of the above is that if $\rho(P, Q) > 0$ then P and Q are mutually absolutely continuous. From this it is easy to derive the Hájek-Feldman theorem: Two Gaussian measures are either mutually absolutely continuous or disjoint.

References

1. Abramowitz, M. and Stegun, I.A. (1964). *Handbook of Mathematical Functions* National Bureau of Standard Applied Mathematics Series. Vol 55.

2. Birgé, L. (1983). "Approximation dans les espaces métriques et théorie de l'estimation." *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **65**, p. 181-237.

3. Dieudonné, J. (1971). *Elements d'analyse,* vol 4. Gauthier-Villars, Paris 411 pages.

4. Hauck, W.W. and Donner, A. (1977). "Wald's test as applied to hypotheses in logit analysis." *J. Amer. Statistical Assoc.* **72**, p. 851-853. Corrigendum (1980) **75**, p. 482.

5. Kraft, C.H. (1955). "Some conditions for consistency and uniform consistency of statistical procedures". *U.C. Pub. Statistics* vol. 2 #6 pp. 125-142.

6. Le Cam, L. (1956). "On the asymptotic theory of estimation and testing hypotheses." *Proc. 3rd Berkeley Symp. Math. Stat. and Prob.* **1**, pp. 129-156.

7. Le Cam, L. (1960). "Locally asymptotically normal families of distributions." *Univ. of California Publ in Statistics* **3**, pp. 37-98.

8. Le Cam, L. (1964). "Sufficiency and approximate sufficiency." *Ann. Math. Stat.* **35**, pp.1419-1455.

9. Le Cam, L. (1975). "On local and global properties in the theory of asymptotic normality of experiments". *Stochastic Processes and Related Topics* M.L. Puri editor - Academic Press pp. 13-54.

10. Le Cam, L. (1977). "On the asymptotic normality of estimates." Proc. of the Symposium to honor J. Neyman. *Pânstw. Wydawn. Nauk.* Warsaw pp. 203-217.

11. Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag. xxvi + 742 page.

12. Mammen, E. (1985). "Local optimal Gaussian approximation of an exponential family". *Heidelberg preprint* 299.

13. Mantel N. (1987). "Understanding Wald's test for exponential families." *The American Statistician* **41**, pp. 147-149.

14. Schwartz, J.T. (1969). *Nonlinear functional analysis*. Gordon and Breach N.Y. 236 pages.

15. Vaeth, (1985). "On the use of Wald's test in exponential families." *Inter. Statistical Review*. **53**, pp. 199-214.

16. Wald, A. (1943). "Tests of statistical hypotheses concerning several parameters when the number of observations is large." *Trans. Amer. Math. Soc.* **54**, pp. 426-482.

TECHNICAL REPORTS

Statistics Department

University of California, Berkeley

1. BREIMAN, L. and FREEDMAN, D. (Nov. 1981, revised Feb. 1982). How many variables should be entered in a regression equation? Jour. Amer. Statist. Assoc., March 1983, 78, No. 381, 131-136.

2. BRILLINGER, D. R. (Jan. 1982). Some contrasting examples of the time and frequency domain approaches to time series analysis. Time Series Methods in Hydrosciences, (A. H. El-Shaarawi and S. R. Esterby, eds.) Elsevier Scientific Publishing Co., Amsterdam, 1982, pp. 1-15.

3. DOKSUM, K. A. (Jan. 1982). On the performance of estimates in proportional hazard and log-linear models. Survival Analysis, (John Crowley and Richard A. Johnson, eds.) IMS Lecture Notes - Monograph Series, (Shanti S. Gupta, series ed.) 1982, 74-84.

4. BICKEL, P. J. and BREIMAN, L. (Feb. 1982). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. Ann. Prob., Feb. 1982, 11. No. 1, 185-214.

5. BRILLINGER, D. R. and TUKEY, J. W. (March 1982). Spectrum estimation and system identification relying on a Fourier transform. The Collected Works of J. W. Tukey, vol. 2, Wadsworth, 1985, 1001-1141.

6. BERAN, R. (May 1982). Jackknife approximation to bootstrap estimates. Ann. Statist., March 1984, 12 No. 1, 101-118.

7. BICKEL, P. J. and FREEDMAN, D. A. (June 1982). Bootstrapping regression models with many parameters. Lehmann Festschrift, (P. J. Bickel, K. Doksum and J. L. Hodges, Jr., eds.) Wadsworth Press, Belmont, 1983, 28-48.

8. BICKEL, P. J. and COLLINS, J. (March 1982). Minimizing Fisher information over mixtures of distributions. Sankhyā, 1983, 45, Series A, Pt. 1, 1-19.

9. BREIMAN, L. and FRIEDMAN, J. (July 1982). Estimating optimal transformations for multiple regression and correlation.

10. FREEDMAN, D. A. and PETERS, S. (July 1982, revised Aug. 1983). Bootstrapping a regression equation: some empirical results. JASA, 1984, 79, 97-106.

11. EATON, M. L. and FREEDMAN, D. A. (Sept. 1982). A remark on adjusting for covariates in multiple regression.

12. BICKEL, P. J. (April 1982). Minimax estimation of the mean of a mean of a normal distribution subject to doing well at a point. Recent Advances in Statistics, Academic Press, 1983.

14. FREEDMAN, D. A., ROTHENBERG, T. and SUTCH, R. (Oct. 1982). A review of a residential energy end use model.

15. BRILLINGER, D. and PREISLER, H. (Nov. 1982). Maximum likelihood estimation in a latent variable problem. Studies in Econometrics, Time Series, and Multivariate Statistics, (eds. S. Karlin, T. Amemiya, L. A. Goodman). Academic Press, New York, 1983, pp. 31-65.

16. BICKEL, P. J. (Nov. 1982). Robust regression based on infinitesimal neighborhoods. Ann. Statist., Dec. 1984, 12, 1349-1368.

17. DRAPER, D. C. (Feb. 1983). Rank-based robust analysis of linear models. I. Exposition and review.

18. DRAPER, D. C. (Feb 1983). Rank-based robust inference in regression models with several observations per cell.

19. FREEDMAN, D. A. and FIENBERG, S. (Feb. 1983, revised April 1983). Statistics and the scientific method, Comments on and reactions to Freedman, A rejoinder to Fienberg's comments. Springer New York 1985 Cohort Analysis in Social Research, (W. M. Mason and S. E. Fienberg, eds.).

20. FREEDMAN, D. A. and PETERS, S. C. (March 1983, revised Jan. 1984). Using the bootstrap to evaluate forecasting equations. J. of Forecasting, 1985, Vol. 4, 251-262.

21. FREEDMAN, D. A. and PETERS, S. C. (March 1983, revised Aug. 1983). Bootstrapping an econometric model: some empirical results. JBES, 1985, 2, 150-158.

22. FREEDMAN, D. A. (March 1983). Structural-equation models: a case study.

23. DAGGETT, R. S. and FREEDMAN, D. (April 1983, revised Sept. 1983). Econometrics and the law: a case study in the proof of antitrust damages. Proc. of the Berkeley Conference, in honor of Jerzy Neyman and Jack Kiefer. Vol I pp. 123-172. (L. Le Cam, R. Olshen eds.) Wadsworth, 1985.

24. DOKSUM, K. and YANDELL, B. (April 1983). Tests for exponentiality. Handbook of Statistics, (P. R. Krishnaiah and P. K. Sen, eds.) 4, 1984.

25. FREEDMAN, D. A. (May 1983). Comments on a paper by Markus.

26. FREEDMAN, D. (Oct. 1983, revised March 1984). On bootstrapping two-stage least-squares estimates in stationary linear models. Ann. Statist., 1984, 12, 827-842.

27. DOKSUM, K. A. (Dec. 1983). An extension of partial likelihood methods for proportional hazard models to general transformation models. Ann. Statist., 1987, 15, 325-345.

28. BICKEL, P. J., GOETZE, F. and VAN ZWET, W. R. (Jan. 1984). A simple analysis of third order efficiency of estimates. Proc. of the Neyman-Kiefer Conference, (L. Le Cam, ed.) Wadsworth, 1985.

29. BICKEL, P. J. and FREEDMAN, D. A. Asymptotic normality and the bootstrap in stratified sampling. Ann. Statist. 12 470-482.

30. FREEDMAN, D. A. (Jan. 1984). The mean vs. the median: a case study in 4-R Act litigation. JBES, 1985 Vol 3 pp. 1-13.

31. STONE, C. J. (Feb. 1984). An asymptotically optimal window selection rule for kernel density estimates. Ann. Statist., Dec. 1984, 12, 1285-1297.

32. BREIMAN, L. (May 1984). Nail finders, edifices, and Oz.

33. STONE, C. J. (Oct. 1984). Additive regression and other nonparametric models. Ann. Statist., 1985, 13, 689-705.

34. STONE, C. J. (June 1984). An asymptotically optimal histogram selection rule. Proc. of the Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer (L. Le Cam and R. A. Olshen, eds.), II, 513-520.

35. FREEDMAN, D. A. and NAVIDI, W. C. (Sept. 1984, revised Jan. 1985). Regression models for adjusting the 1980 Census. Statistical Science, Feb 1986, Vol. 1, No. 1, 3-39.

36. FREEDMAN, D. A. (Sept. 1984, revised Nov. 1984). De Finetti's theorem in continuous time.

37. DIACONIS, P. and FREEDMAN, D. (Oct. 1984). An elementary proof of Stirling's formula. Amer. Math Monthly, Feb 1986, Vol. 93, No. 2, 123-125.

38. LE CAM, L. (Nov. 1984). Sur l'approximation de familles de mesures par des familles Gaussiennes. Ann. Inst. Henri Poincaré, 1985, 21, 225-287.

39. DIACONIS, P. and FREEDMAN, D. A. (Nov. 1984). A note on weak star uniformities.

40. BREIMAN, L. and IHAKA, R. (Dec. 1984). Nonlinear discriminant analysis via SCALING and ACE.

41. STONE, C. J. (Jan. 1985). The dimensionality reduction principle for generalized additive models.

42. LE CAM, L. (Jan. 1985). On the normal approximation for sums of independent variables.

43. BICKEL, P. J. and YAHAV, J. A. (1985). On estimating the number of unseen species: how many executions were there?

44. BRILLINGER, D. R. (1985). The natural variability of vital rates and associated statistics. Biometrics, to appear.

45. BRILLINGER, D. R. (1985). Fourier inference: some methods for the analysis of array and nonGaussian series data. Water Resources Bulletin, 1985, 21, 743-756.

46. BREIMAN, L. and STONE, C. J. (1985). Broad spectrum estimates and confidence intervals for tail quantiles.

47. DABROWSKA, D. M. and DOKSUM, K. A. (1985, revised March 1987). Partial likelihood in transformation models with censored data.

48. HAYCOCK, K. A. and BRILLINGER, D. R. (November 1985). LIBDRB: A subroutine library for elementary time series analysis.

49. BRILLINGER, D. R. (October 1985). Fitting cosines: some procedures and some physical examples. Joshi Festschrift, 1986. D. Reidel.

50. BRILLINGER, D. R. (November 1985). What do seismology and neurophysiology have in common? - Statistics! Comptes Rendus Math. Rep. Acad. Sci. Canada, January, 1986.

51. COX, D. D. and O'SULLIVAN, F. (October 1985). Analysis of penalized likelihood-type estimators with application to generalized smoothing in Sobolev Spaces.

52. O'SULLIVAN, F. (November 1985). A practical perspective on ill-posed inverse problems: A review with some new developments. To appear in Journal of Statistical Science.

53. LE CAM, L. and YANG, G. L. (November 1985, revised March 1987). On the preservation of local asymptotic normality under information loss.

54. BLACKWELL, D. (November 1985). Approximate normality of large products.

55. FREEDMAN, D. A. (December 1985, revised Dec. 1986). As others see us: A case study in path analysis. Prepared for the Journal of Educational Statistics.

56. LE CAM, L. and YANG, G. L. (January 1986). Distinguished Statistics, Loss of information and a theorem of Robert B. Davies.

57. LE CAM, L. (February 1986). On the Bernstein - von Mises theorem.

58. O'SULLIVAN, F. (January 1986). Estimation of Densities and Hazards by the Method of Penalized likelihood.

59. ALDOUS, D. and DIACONIS, P. (February 1986). Strong Uniform Times and Finite Random Walks.

60. ALDOUS, D. (March 1986). On the Markov Chain simulation Method for Uniform Combinatorial Distributions and Simulated Annealing.

61. CHENG, C-S. (April 1986). An Optimization Problem with Applications to Optimal Design Theory.

62. CHENG, C-S., MAJUMDAR, D., STUFKEN, J. & TURE, T. E. (May 1986, revised Jan 1987). Optimal step type design for comparing test treatments with a control.

63. CHENG, C-S. (May 1986, revised Jan. 1987). An Application of the Kiefer-Wolfowitz Equivalence Theorem.

64. O'SULLIVAN, F. (May 1986). Nonparametric Estimation in the Cox Proportional Hazards Model.

65. ALDOUS, D. (JUNE 1986). Finite-Time Implications of Relaxation Times for Stochastically Monotone Processes.

66. PITMAN, J. (JULY 1986, revised November 1986). Stationary Excursions.

67. DABROWSKA, D. and DOKSUM, K. (July 1986, revised November 1986). Estimates and confidence intervals for median and mean life in the proportional hazard model with censored data.

68. LE CAM, L. and YANG, G.L. (July 1986). Distinguished Statistics, Loss of information and a theorem of Robert B. Davies (Fourth edition).

69. STONE, C.J. (July 1986). Asymptotic properties of logspline density estimation.

71. BICKEL, P.J. and YAHAV, J.A. (July 1986). Richardson Extrapolation and the Bootstrap.

72. LEHMANN, E.L. (July 1986). Statistics - an overview.

73. STONE, C.J. (August 1986). A nonparametric framework for statistical modelling.

74. BIANE, PH. and YOR, M. (August 1986). A relation between Lévy's stochastic area formula, Legendre polynomials, and some continued fractions of Gauss.

75. LEHMANN, E.L. (August 1986, revised July 1987). Comparing Location Experiments.

76. O'SULLIVAN, F. (September 1986). Relative risk estimation.

77. O'SULLIVAN, F. (September 1986). Deconvolution of episodic hormone data.

78. PITMAN, J. & YOR, M. (September 1987). Further asymptotic laws of planar Brownian motion.

79. FREEDMAN, D.A. & ZEISEL, H. (November 1986). From mouse to man: The quantitative assessment of cancer risks.

80. BRILLINGER, D.R. (October 1986). Maximum likelihood analysis of spike trains of interacting nerve cells.

81. DABROWSKA, D.M. (November 1986). Nonparametric regression with censored survival time data.

82. DOKSUM, K.J. and LO, A.Y. (November 1986). Consistent and robust Bayes Procedures for Location based on Partial Information.

83. DABROWSKA, D.M., DOKSUM, K.A. and MIURA, R. (November 1986). Rank estimates in a class of semiparametric two-sample models.

84. BRILLINGER, D. (December 1986). Some statistical methods for random process data from seismology and neurophysiology.

85. DIACONIS, P. and FREEDMAN, D. (December 1986). A dozen de Finetti-style results in search of a theory.

86. DABROWSKA, D.M. (January 1987). Uniform consistency of nearest neighbour and kernel conditional Kaplan - Meier estimates.

87. FREEDMAN, D.A., NAVIDI, W. and PETERS, S.C. (February 1987). On the impact of variable selection in fitting regression equations.

88. ALDOUS, D. (February 1987, revised April 1987). Hashing with linear probing, under non-uniform probabilities.

89. DABROWSKA, D.M. and DOKSUM, K.A. (March 1987, revised November 1987). Estimating and testing in a two sample generalized odds rate model.

90. DABROWSKA, D.M. (March 1987). Rank tests for matched pair experiments with censored data.

91. DIACONIS, P and FREEDMAN, D.A. (March 1987). A finite version of de Finetti's theorem for exponential families, with uniform asymptotic estimates.

92. DABROWSKA, D.M. (April 1987, revised September 1987). Kaplan-Meier estimate on the plane.

92a. ALDOUS, D. (April 1987). The Harmonic mean formula for probabilities of Unions: Applications to sparse random graphs.

93. DABROWSKA, D.M. (June 1987). Nonparametric quantile regression with censored data.

94. DONOHO, D.L. & STARK, P.B. (June 1987). Uncertainty principles and signal recovery.

95. RIZZARDI, F. (Aug 1987). Two-Sample t-tests where one population SD is known.

96. BRILLINGER, D.R. (June 1987). Some examples of the statistical analysis of seismological data. To appear in *Proceedings, Centennial Anniversary Symposium, Seismographic Stations, University of California, Berkeley.*

97. FREEDMAN, D.A. and NAVIDI, W. (June 1987). On the multi-stage model for cancer.

98. O'SULLIVAN, F. and WONG, T. (June 1987). Determining a function diffusion coefficient in the heat equation.

99. O'SULLIVAN, F. (June 1987). Constrained non-linear regularization with application to some system identification problems.

100. LE CAM, L. (July 1987, revised Nov 1987). On the standard asymptotic confidence ellipsoids of Wald.

101. DONOHO, D.L. and LIU, R.C. (July 1987). Pathologies of some minimum distance estimators.

102. BRILLINGER, D.R., DOWNING, K.H. and GLAESER, R.M. (July 1987). Some statistical aspects of low-dose electron imaging of crystals.

103. LE CAM, L. (August 1987). Harald Cramér and sums of independent random variables.

104. DONOHO, A.W., DONOHO, D.L. and GASKO, M. (August 1987). Macspin: Dynamic graphics on a desktop computer.

105. DONOHO, D.L. and LIU, R.C. (August 1987). On minimax estimation of linear functionals.

106. DABROWSKA, D.M. (August 1987). Kaplan-Meier estimate on the plane: weak convergence, LIL and the bootstrap.

107. CHENG, C-S. (August 1987). Some orthogonal main-effect plans for asymmetrical factorials.

108. CHENG, C-S. and JACROUX, M. (August 1987). On the construction of trend-free run orders of two-level factorial designs.

109. KLASS, M.J. (August 1987). Maximizing $E \max_{1 \le k \le n} S_k^+ / ES_n^+$: A prophet inequality for sums of I.I.D. mean zero variates.

110. DONOHO, D.L. and LIU, R.C. (August 1987). The "automatic" robustness of minimum distance functionals.

111. BICKEL, P.J. and GHOSH, J.K. (August 1987). A decomposition for the likelihood ratio statistic and the Bartlett correction — a Bayesian argument.

112. BURDZY, K., PITMAN, J.W. and YOR, M. (September 1987). Some asymptotic laws for crossings and excursions.

113. ADHIKARI, A. and PITMAN, J. (September 1987). The shortest planar arc of width 1.

114. RITOV, Y. (September 1987). Estimation in a linear regression model with censored data.

115. BICKEL, P.J. and RITOV, Y. (September 1987). Large sample theory of estimation in biased sampling regression models I.

116. RITOV, Y. and BICKEL, P.J. (September 1987). Unachievable information bounds in non and semiparametric models.

117. RITOV, Y. (October 1987). On the convergence of a maximal correlation algorithm using alternating projections.

118. ALDOUS, D.J. (October 1987). Meeting times for independent Markov chains.

119. HESSE, C.H. (October 1987). An asymptotic expansion for the mean of the passage-time distribution of integrated Brownian Motion.

120. DONOHO, D. (October 1987). Geometrizing rates of convergence, II.

121. BRILLINGER, D.R. (October 1987). Estimating the chances of large earthquakes by radiocarbon dating and statistical modelling. To appear in *Statistics a Guide to the Unknown*.

122. ALDOUS, D., FLANNERY, B. and PALACIOS, J.L. (November 1987). Two applications of urn processes: The fringe analysis of search trees and the simulation of quasi-stationary distributions of Markov chains.

Copies of these Reports plus the most recent additions to the Technical Report series are available from the Statistics Department technical typist in room 379 Evans Hall or may be requested by mail from:

Department of Statistics
University of California
Berkeley, California 94720

Cost: $1 per copy.