DEFINING THE CURVATURE OF A STATISTICAL PROBLEM (WITH APPLICATIONS TO SECOND ORDER EFFICIENCY)

BY BRADLEY EFRON

Stanford University

Statisticians know that one-parameter exponential families have very nice properties for estimation, testing, and other inference problems. Fundamentally this is because they can be considered to be "straight lines" through the space of all possible probability distributions on the sample space. We consider arbitrary one-parameter families $\mathscr{F}$ and try to quantify how nearly "exponential" they are. A quantity called "the statistical curvature of $\mathscr{F}$" is introduced. Statistical curvature is identically zero for exponential families, positive for nonexponential families. Our purpose is to show that families with small curvature enjoy the good properties of exponential families. Large curvature indicates a breakdown of these properties. Statistical curvature turns out to be closely related to Fisher and Rao's theory of second order efficiency.

1. Introduction. Suppose we have a statistical problem involving a one-parameter family of probability density functions $\mathscr{F} = \{f_\theta(x)\}$. Statisticians know that if $\mathscr{F}$ is an exponential family then standard linear methods will usually solve the problem in neat fashion. For example, the locally most powerful test of $\theta = \theta_0$ versus $\theta > \theta_0$ is uniformly most powerful in an exponential family. The maximum likelihood estimator for θ is a sufficient statistic in an exponential family, and achieves the Cramér–Rao lower bound if we have chosen the right function of θ to estimate.

In this paper we consider arbitrary one-parameter families $\mathscr{F}$ and try to quantify how nearly "exponential" they are. A quantity γ_θ called "*the statistical curvature of* $\mathscr{F}$ *at* θ" is introduced such that γ_θ is identically zero if $\mathscr{F}$ is exponential and greater than zero, for at least some θ values, otherwise.

Our purpose is to show that families with small curvature enjoy, nearly, the good statistical properties of exponential families. Large curvature indicates a breakdown of this favorable situation. For example, if γ_{θ_0} is large, the locally most powerful test of $\theta = \theta_0$ versus $\theta > \theta_0$ can be expected to have poor operating characteristics. Similarly the variance of the maximum likelihood estimator (MLE) exceeds the Cramér–Rao lower bound in approximate proportion to $\gamma_{\theta_0}^2$. (See Sections 8 and 10.)

For nonexponential families the MLE is not, in general, a sufficient statistic. How much information does it lose, compared with all the data x? The answer

Received March 1974; revised May 1975.

AMS 1970 *subject classifications.* 62B10, 62F20.

Key words and phrases. Curvature, exponential families, Cramér Rao lower bound, locally most powerful tests, Fisher information, second order efficiency, deficiency, maximum likelihood estimation.

can be expressed in terms of γ_θ^2. This theory goes back to Fisher (1925) and Rao (1961, 1962, 1963). They attempted to show that if $\mathcal{F}$ is a one-parameter subset of the k-category multinomial distributions, indexed say by the vector of probabilities $f_\theta(x) = P_\theta(X \in \text{category } x)$, $x = 1, 2, \cdots, k$, the following result holds: let $\mathbf{i}_\theta$ be the Fisher information in an independent sample of size n from f_θ, $\mathbf{i}_\theta^{\hat{\theta}}$ the Fisher information in the maximum likelihood estimator $\hat{\theta}(x_1, x_2, \cdots, x_n)$ based on that sample, and i_θ the Fisher information in a sample of size one (so $\mathbf{i}_\theta = ni_\theta$). Then

$$\lim_{n\to\infty} (\mathbf{i}_\theta - \mathbf{i}_\theta^{\hat{\theta}}) = i_\theta \left\{ \frac{\mu_{02} - 2\mu_{21} + \mu_{40}}{i_\theta^2} - 1 - \frac{\mu_{11}^2 + \mu_{30}^2 - 2\mu_{11}\mu_{30}}{i_\theta^3} \right\} \tag{1.1}$$

where

$$\mu_{hj} \equiv E_\theta \left(\frac{\dot{f}_\theta(x)}{f_\theta(x)} \right)^h \left(\frac{\ddot{f}_\theta(x)}{f_\theta(x)} \right)^j, \tag{1.2}$$

the dot indicating differentiation with respect to θ. Moreover, for any other consistent, efficient estimator $T(x_1, x_2, \cdots, x_n)$ the asymptotic loss of information $\lim_{n\to\infty} (\mathbf{i}_\theta - \mathbf{i}_\theta^T)$ is equal or greater than the right side of (1.1). Rao has coined the term "*second order efficiency*" for this property of the MLE which gives it a preferred place in the class of "first order efficient" estimators T, those which satisfy the weaker condition $\lim_{n\to\infty} \mathbf{i}_\theta^T / \mathbf{i}_\theta = 1$.

It turns out that the unpleasant looking bracketed term in (1.1) equals γ_θ^2. This leads to a straightforward geometrical "proof" of (1.1). The quotes are necessary here since, as the counter-example of Section 9 shows, the result is actually not true for multinomial families. However, the difficulty arises only because of the discrete nature of the multinomial, and can be overcome by dealing with less lumpy distributions. More importantly, a similar result of Rao's for squared error estimation risk holds even for the multinomial, as discussed in Section 10.

Under our definition an exponential family has zero curvature everywhere so in some sense it is a "straight line through the space of possible probability distributions." (This is intuitively plausible since linear methods, that is, methods based on linear approximations to the log likelihood function, tend to work perfectly in exponential families. The fact that locally most powerful tests are uniformly most powerful is an example of this.) We will make this notion precise by considering families $\mathcal{F}$ which are subsets of multi-parameter exponential families. If the subset is a straight line in the natural parameter space of the bigger family then $\mathcal{F}$ is a one-parameter exponential family. If the subset is a curved line through the natural parameter space then $\mathcal{F}$ is not exponential, and it turns out that the statistical curvature exactly equals the ordinary geometric curvature of the line, the rate of change of direction with respect to arc-length. For the sake of exposition we actually start with this latter definition in Section 3 and show in Section 5 how it leads to a sensible definition of statistical curvature in the general case.

There are really two halves to this paper. Sections 3–7 introduce the notion of statistical curvature, Sections 8–10 apply curvature to hypothesis testing, partial sufficiency, and estimation. Section 2 consists of a brief review of the notion of the geometrical curvature of a line.

2. Curvature. If $Y = Y(X)$ defines a curved line $\mathscr{L}$ in the (X, Y) plane then

$$\gamma_X = \left[\frac{(Y'')^2}{[1 + (Y')^2]^3}\right]^{\frac{1}{2}} \tag{2.1}$$

is defined to be the curvature of $\mathscr{L}$ at X, where $Y' \equiv dY/dX$, $Y'' \equiv d^2Y/dX^2$ are assumed to exist continuously in a neighborhood of the value X where the curvature is being evaluated. In particular if $Y' = 0$ then $\gamma_X = |Y''|$. An exercise in differential calculus shows that γ_X is the rate of change of direction of $\mathscr{L}$ with respect to arc-length along the curve. The "*radius of curvature*", $\rho_X \equiv 1/\gamma_X$, is the radius of the circle tangent to $\mathscr{L}$ at (X, Y) whose Taylor expansion about (X, Y) agrees up to the quadratic term with that of $\mathscr{L}$. Struik (1950) is a good elementary reference for curvature and related concepts.

The concept of curvature extends to curved lines in Euclidean k-space, E^k, say $\mathscr{L} = \{\eta_\theta, \theta \in \Theta\}$, where Θ is an interval of the real line. For each θ, η_θ is a vector in E^k whose componentwise derivatives with respect to θ we denote $\dot{\eta}_\theta \equiv (\partial/\partial\theta)\eta_\theta$, $\ddot{\eta}_\theta \equiv (\partial^2/\partial\theta^2)\eta_\theta$. These derivatives are assumed to exist continuously in a neighborhood of a value of θ where we wish to define the curvature. Suppose also that a $k \times k$ symmetric nonnegative definite matrix $\mathbf{\Sigma}_\theta$ is defined continuously in θ. Let M_θ be the 2×2 matrix, with entries denoted $\nu_{20}(\theta)$, $\nu_{11}(\theta)$, $\nu_{02}(\theta)$ as shown, defined by

$$M_\theta \equiv \begin{pmatrix} \nu_{20}(\theta) & \nu_{11}(\theta) \\ \nu_{11}(\theta) & \nu_{02}(\theta) \end{pmatrix} \equiv \begin{pmatrix} \dot{\eta}_\theta' \mathbf{\Sigma}_\theta \dot{\eta}_\theta & \dot{\eta}_\theta' \mathbf{\Sigma}_\theta \ddot{\eta}_\theta \\ \ddot{\eta}_\theta' \mathbf{\Sigma}_\theta \dot{\eta}_\theta & \ddot{\eta}_\theta' \mathbf{\Sigma}_\theta \ddot{\eta}_\theta \end{pmatrix} \tag{2.2}$$

and let

$$\gamma_\theta \equiv (|M_\theta|/\nu_{20}^3(\theta))^{\frac{1}{2}} . \tag{2.3}$$

Then γ_θ is "*the curvature of $\mathscr{L}$ at θ with respect to the inner product $\mathbf{\Sigma}_\theta$*". (If we take $k = 2$, $\theta = X$, $\eta_\theta = (X, Y(X))$, and $\mathbf{\Sigma}_\theta \equiv \mathbf{I}$, then (2.3) reduces to (2.1).)

Again it can be shown that γ_θ is the rate of change of direction of η_θ with respect to arc-length along $\mathscr{L}$. The relevant quantities are illustrated in Figure 1, where the arc-length from a given point η_{θ_0} to η_θ is called "s_θ" and the angle

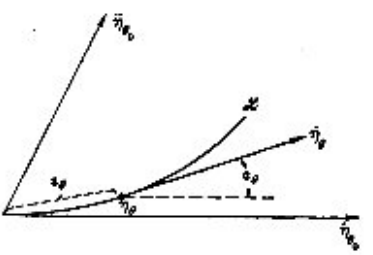

FIG. 1. The curvature of $\mathscr{L}$ at θ_0 is $da_\theta/ds_\theta|_{\theta=\theta_0}$.

between $\dot{\eta}_{\theta_0}$ and $\dot{\eta}_\theta$ called "a_θ". Then

$$\gamma_{\theta_0} = \frac{da_\theta}{ds_\theta}\Big|_{\theta_0} \tag{2.4}$$

or equivalently $\gamma_{\theta_0} = d \sin a_\theta / ds_\theta|_{\theta_0}$. Both s_θ and a_θ are defined relative to the inner product $\mathbf{\Sigma}_\theta$,

$$\frac{ds_\theta}{d\theta} \equiv (\dot{\eta}_\theta' \mathbf{\Sigma}_\theta \dot{\eta}_\theta)^{\frac{1}{2}} \tag{2.5}$$

$$\sin a_\theta \equiv \left[1 - \frac{(\dot{\eta}_{\theta_0}' \mathbf{\Sigma}_{\theta_0} \dot{\eta}_\theta)^2}{(\dot{\eta}_{\theta_0}' \mathbf{\Sigma}_{\theta_0} \dot{\eta}_{\theta_0})(\dot{\eta}_\theta' \mathbf{\Sigma}_{\theta_0} \dot{\eta}_\theta)}\right]^{\frac{1}{2}}. \tag{2.6}$$

($\mathbf{\Sigma}_{\theta_0}$ can be replaced by $\mathbf{\Sigma}_\theta$ anywhere in (2.6).) As Figure 1 indicates, for the purpose of evaluating γ_{θ_0} the k-dimensional curve $\mathscr{L}$ can be considered locally as a two-dimensional curve in the plane through η_{θ_0} spanned by $\dot{\eta}_{\theta_0}$ and $\ddot{\eta}_{\theta_0}$.

3. Curved exponential families. In this section we define statistical curvature for one parameter families $\mathscr{F}$ which are curved subsets of a larger k-parameter exponential family, "curved exponential families" for short. Denote the multiparameter family by

$$g_\eta(x) \equiv g(x) e^{\eta' x - \psi(\eta)} \tag{3.1}$$

a family of densities with respect to some given measure $m(\cdot)$, possibly discrete, on Euclidean k-space E^k. Here $\eta \in \mathscr{N}$, the subset of E^k for which $\int_{E^k} g(x) e^{\eta' x}\, dm(x) < \infty$. The convex set $\mathscr{N}$ is called *the natural parameter space* of the exponential family. If we define

$$\lambda(\eta) \equiv E_\eta x \tag{3.2}$$

the components of λ can be obtained by differentiation of ψ, $\lambda_i(\eta) = (\partial/\partial\eta_i)\psi(\eta)$. Moreover the covariance matrix $\mathbf{\Sigma}(\eta)$ of x under g_η has ijth element equal to $\partial^2\psi(\eta)/\partial\eta_i \partial\eta_j$. We denote by Λ the set of all mean vectors λ,

$$\Lambda = \{\lambda(\eta) : \eta \in \mathscr{N}\} \tag{3.3}$$

The mapping (3.2) from $\mathscr{N}$ to Λ is one-to-one, and we will often write λ instead of $\lambda(\eta)$, recognizing that λ indexes the exponential family as well as η does. $\mathbf{\Sigma}(\eta)$ has the same rank r for all η, and *we will assume rank $r \geqq 2$ to avoid trivialities.*

Now suppose that

$$\mathscr{L} \equiv \{\eta_\theta : \theta \in \Theta\} \tag{3.4}$$

is a one-parameter subset in the interior of $\mathscr{N}$, where η_θ is a continuously twice differentiable function of $\theta \in \Theta$, an interval of the real line. Define the density f_θ to be

$$f_\theta(x) \equiv g_{\eta_\theta}(x) = g(x) e^{\eta_\theta' x - \psi_\theta}, \tag{3.5}$$

where $\psi_\theta \equiv \psi(\eta_\theta)$. (Likewise $\lambda_\theta \equiv \lambda(\eta_\theta)$, $\mathbf{\Sigma}_\theta \equiv \mathbf{\Sigma}(\eta_\theta)$.) It is easy to verify that

$$\dot{\lambda}_\theta = \mathbf{\Sigma}_\theta \dot{\eta}_\theta, \qquad \dot{\psi}_\theta = \dot{\eta}_\theta' \lambda_\theta = E_\theta \dot{\eta}_\theta' x. \tag{3.6}$$

$\mathscr{F}$ will stand for the family of densities $\{f_\theta(x): \theta \in \Theta\}$, our curved exponential family.

DEFINITION. γ_θ, the statistical curvature of $\mathscr{F}$ at θ, is the geometrical curvature of $\mathscr{C} = \{\eta_\theta : \theta \in \Theta\}$ at θ with respect to the covariance inner product $\dot{\Sigma}_\theta$, as defined in (2.2) and (2.3).

EXAMPLE 1. *Bivariate normal.* x is a bivariate normal random vector with covariance matrix I and mean vector $\eta_\theta = (\theta, (\gamma_0/2)\theta^2)'$, $\theta \in \Theta = (-\infty, \infty)$,

$$x \sim \mathcal{N}_2(\eta_\theta, \mathrm{I}) . \tag{3.7}$$

Then $\dot{\eta}_\theta = (1, \gamma_0\theta)'$, $\ddot{\eta}_\theta = (0, \gamma_0)'$, and

$$M_\theta = \begin{pmatrix} 1 + \gamma_0^2\theta^2 & \gamma_0^2\theta \\ \gamma_0^2\theta & \gamma_0^2 \end{pmatrix} \tag{3.8}$$

so

$$\gamma_\theta^2 = \frac{\gamma_0^2}{(1 + \gamma_0^2\theta^2)^3} . \tag{3.9}$$

In particular $\gamma_0^2 = \gamma_0^2$, justifying the notation. This artificial but very simple curved exponential family will be used for illustrative purposes in Section 8.

EXAMPLE 2. *Poisson regression.* $x_1, x_2, \cdots, x_k$ are independent Poisson random variables, x_i having mean $a + \theta b_i$, $b_1, b_2, \cdots, b_k$ and $a > 0$ being known parameters. Θ is the interval of θ values such that $a + \theta b_i > 0$ for $i = 1, 2, \cdots, k$. Since $x = (x_1, \cdots, x_k)'$ has a k parameter exponential family of distributions if the k means are unconstrained, we apply definition (2.2) to get the elements of M_θ,

$$\nu_{20}(\theta) = \sum_{i=1}^k \frac{b_i^3}{a + \theta b_i}, \qquad \nu_{11}(\theta) = -\sum_{i=1}^k \frac{b_i^3}{(a + \theta b_i)^2}, \tag{3.10}$$

$$\nu_{02}(\theta) = \sum_{i=1}^k \frac{b_i^4}{(a + \theta b_i)^3} .$$

The formula (2.3) for γ_θ^2 simplifies at $\theta = 0$ to

$$\gamma_0^2 = \frac{1}{a}\left[\frac{\sum_{i=1}^k b_i^4}{(\sum_{i=1}^k b_i^3)^2} - \frac{(\sum_{i=1}^k b_i^3)^2}{(\sum_{i=1}^k b_i^2)^3}\right] \tag{3.11}$$

That the entries of M_θ are summations follows from the independence of $x_1, x_2, \cdots, x_k$, as mentioned in Section 6. A very similar formula holds for the analogous binomial regression model.

The *Neyman–Davies model*, $x_1, x_2, \cdots, x_k$ independent scaled χ_1^2 random variables, $x_i \overset{\text{inv}}{\sim} (1 + \theta\delta_i)\chi_1^2$, $\delta_1, \delta_2, \cdots, \delta_k$ known constants, has the same structure. (Davies (1969) uses this model, which originates in an application due to Neyman, to investigate the power of the locally most powerful test of $\theta = 0$ versus $\theta > 0$. We compare our results with his in Section 8.) By direct calculation or by the

remark at the end of Section 6 we get that M_0 has elements

$$\nu_{20}(0) = \tfrac{1}{2}\sum_{i=1}^{k}\delta_i^2\,, \qquad \nu_{11}(0) = -\sum_{i=1}^{k}\delta_i^3\,, \qquad \nu_{02}(0) = 2\sum_{i=1}^{k}\delta_i^4 \tag{3.12}$$

and so

$$\gamma_0^2 = 8\left[\frac{\sum_{i=1}^{k}\delta_i^4}{(\sum_{i=1}^{k}\delta_i^2)^2} - \frac{(\sum_{i=1}^{k}\delta_i^3)^2}{(\sum_{i=1}^{k}\delta_i^2)^3}\right]. \tag{3.13}$$

EXAMPLE 3. *Autoregressive process.* $y_0, y_1, \cdots, y_T$ are observations of the autoregressive process $y_0 = u_0$, $y_{t+1} = \theta y_t + (1-\theta^2)^{\frac{1}{2}}u_{t+1}$, $t = 1, 2, \cdots, T$. Here $u_t \overset{\text{ind}}{\sim} \mathcal{N}(0, 1)$, $t = 0, 1, \cdots, T$ and $\Theta = (-1, 1)$. Writing out the likelihood function of $(y_0, \cdots, y_T)$ shows that this is a curved exponential family with $k = 3$, the η_i vector being $\eta_\theta' = (-(1+\theta^2)/a, \theta, -\frac{1}{2})/(1-\theta^2)$, with corresponding sufficient statistics $x' = (\sum_1^{T-1} y_t^2, \sum_1^T y_t y_{t-1}, y_0^2 + y_T^2)$. For $\theta = 0$ the calculations are easy, yielding

$$M_0 = \begin{pmatrix} T & 0 \\ 0 & 8T-6 \end{pmatrix}, \qquad \gamma_0^2 = \frac{8T-6}{T^2}\,. \tag{3.14}$$

Much messier expressions are found for other values of θ. γ_θ^2 is of the form $c_\theta/T + O(1/T^2)$ as $T \to \infty$, with $c_0 = 8$, $c_{.25} = 6.25$, $c_{.5} = 3.07$, $c_{.75} = .96$. (For any T, $\gamma_{-\theta} = \gamma_\theta$, $i_{-\theta} = i_\theta$ since the mapping $(y_0, y_1, y_2, \cdots) \to (y_0, -y_1, y_2, \cdots)$ takes θ into $-\theta$ while preserving the curvature and Fisher information.) This family is least like a one-parameter exponential family at $\theta = 0$.

If $\mathscr{L}$ is a straight line through $\mathcal{N}$, $\eta_\theta = a + b\tau(\theta)$ where a and b are known vectors and $\tau(\theta)$ some real-valued twice differentiable function of θ, then $\gamma_\theta = 0$ for all θ since the curvature of a straight line is zero. In this case $f_\theta(x) = (g(x)e^{a'x})\exp[\tau(\theta)b'x - \psi_\theta]$ is a one-parameter exponential family with natural parameter $\tau(\theta)$ and sufficient statistic $b'x$. Under our definition all one-parameter exponential families $\mathscr{F}$, and only such families, have statistical curvature everywhere equal to zero. This desirable property would still hold if we defined the curvature with respect to an inner product other than $\mathbf{\Sigma}_\theta$, say $\mathbf{\Sigma}_\theta^{-1}$ or $\mathbf{I}$. The following discussion and Section 4 add support to the choice $\mathbf{\Sigma}_\theta$.

Let $l_\theta(x)$ denote the logarithm of $f_\theta(x)$,

$$l_\theta(x) \equiv \log f_\theta(x) \tag{3.15}$$

and denote the first and second partial derivations with respect to θ by

$$\dot{l}_\theta(x) \equiv \frac{\partial}{\partial\theta} l_\theta(x)\,, \qquad \ddot{l}_\theta(x) \equiv \frac{\partial^2}{\partial\theta^2} l_\theta(x)\,. \tag{3.16}$$

The moment relationships

$$E_\theta \dot{l}_\theta = 0\,, \qquad E_\theta \dot{l}_\theta^2 = -E_\theta \ddot{l}_\theta \equiv i_\theta\,, \tag{3.17}$$

where i_θ is Fisher's information, hold because the exponential family structure (3.1)—(3.5) allows us to differentiate under integral signs with impunity. (We will suppress the random element "x" in much of the subsequent notation.)

Notice that $l_\theta(x) = \eta_\theta' x - \psi_\theta + \log g(x)$ so that

$$\dot{l}_\theta(x) = \dot{\eta}_\theta'(x - \lambda_\theta), \qquad \ddot{l}_\theta(x) = \ddot{\eta}_\theta'(x - \lambda_\theta) - \dot{\eta}_\theta' \mathbf{\Sigma}_\theta \dot{\eta}_\theta, \tag{3.18}$$

where we have made use of (3.6) in taking the derivatives. Remembering that $\mathbf{\Sigma}_\theta$ is the covariance matrix of x, we see that (3.17) holds with

$$i_\theta = \dot{\eta}_\theta' \mathbf{\Sigma}_\theta \dot{\eta}_\theta . \tag{3.19}$$

As a matter of fact the covariance matrix of $(\dot{l}_\theta, \ddot{l}_\theta)$ is

$$E_\theta \begin{pmatrix} \dot{l}_\theta \\ \ddot{l}_\theta + i_\theta \end{pmatrix} (\dot{l}_\theta, \ddot{l}_\theta + i_\theta) = \begin{pmatrix} \dot{\eta}_\theta' \mathbf{\Sigma}_\theta \dot{\eta}_\theta & \dot{\eta}_\theta' \mathbf{\Sigma}_\theta \ddot{\eta}_\theta \\ \ddot{\eta}_\theta' \mathbf{\Sigma}_\theta \dot{\eta}_\theta & \ddot{\eta}_\theta' \mathbf{\Sigma}_\theta \ddot{\eta}_\theta \end{pmatrix} \tag{3.20}$$

which is just the matrix M_θ defined at (2.2). Therefore

$$\begin{aligned} &\nu_{20}(\theta) = i_\theta = E_\theta \dot{l}_\theta^2, \qquad \nu_{11}(\theta) = E_\theta \dot{l}_\theta \ddot{l}_\theta = \operatorname{Cov}_\theta (\dot{l}_\theta, \ddot{l}_\theta), \\ &\nu_{02}(\theta) = E_\theta \ddot{l}_\theta^2 - i_\theta^2 = \operatorname{Var}_\theta \ddot{l}_\theta . \end{aligned} \tag{3.21}$$

These definitions make no explicit reference to the geometrical structure of the curved exponential family. We will use them in Section 5 to provide the curvature definition for an arbitrary one-parameter family.

4. Invariance properties of the curvature. The two definitions of M_θ, the geometrical one following (3.6) and the statistical one (3.21) give two useful invariance properties of the curvature γ_θ.

i) Statistical curvature is an intrinsic property of the family $\mathcal{F}$ and does not depend on the particular parameterization used to index $\mathcal{F}$. If we let $\tilde{\theta} \equiv g(\theta)$, where g is any strictly monotone twice differentiable function, and $f_{\tilde{\theta}}(x) \equiv f_{g^{-1}\tilde{\theta}}(x)$, then $\tilde{\gamma}_{\tilde{\theta}} = \gamma_{g^{-1}\tilde{\theta}}$ for every $\tilde{\theta} \in \tilde{\Theta} \equiv g(\Theta)$. This follows from the same property of the geometrical curvature (2.3). [Note: this is not true for the Fisher information: $\tilde{\imath}_{\tilde{\theta}} = i_{g^{-1}\tilde{\theta}} (d\theta/d\tilde{\theta})^2$.]

ii) If $t = T(x)$, is sufficient for θ then $l_\theta^T(t) \equiv \partial/\partial\theta \log f_\theta^T(t) = l_\theta(x)$, where f_θ^T indicates the density of T, implying by (3.18) that $M_\theta^T = M_\theta$ and $\gamma_\theta^T = \gamma_\theta$. The statistical curvature is invariant under any mapping to a sufficient statistic, including of course all one-to-one mappings of the sample space. This property would not hold if we had chosen an inner product other than $\mathbf{\Sigma}_\theta$ in the definition of statistical curvature.

We can use property (ii) to transform an arbitrary curved exponential family into a form particularly convenient for theoretical calculations. Let θ_0 be some value of θ at which we wish to investigate the local behavior of $\mathcal{F}$. Write $\mathbf{\Sigma}_{\theta_0} \equiv \mathbf{A}'\mathbf{D}\mathbf{A}$, $\mathbf{D}$ an $r \times r$ diagonal matrix with positive diagonal elements and $\mathbf{A}$ an $r \times k$ matrix with orthonormal rows, $\mathbf{A}\mathbf{A}' = \mathbf{I}_r$ (rank $\mathbf{\Sigma}_\theta = r$, $\mathbf{I}_r$ the $r \times r$ identity matrix). Let $\tilde{x} \equiv \mathbf{\Gamma}\mathbf{D}^{-\frac{1}{2}}\mathbf{A}(x - \lambda_{\theta_0})$ where $\mathbf{\Gamma}$ is an as yet unspecified $r \times r$ orthogonal matrix. $\tilde{x}$ is an r-dimensional sufficient statistic for the family (3.1). For $\theta \in \Theta$ it has a curved exponential family of densities where we can take $\tilde{\eta}_\theta = \mathbf{\Gamma}\mathbf{D}^{\frac{1}{2}}\mathbf{A}(\eta_\theta - \eta_{\theta_0})$. (These statements are easily shown in the full rank case $r = k$ and are not difficult for $r < k$.)

Notice that $\dot{\eta}_{\theta_0} = 0$, $\dot{\lambda}_{\theta_0} = 0$, and $\dot{\tilde{\Sigma}}_{\theta_0} = I_r$. Proper choice of the rotation matrix Γ makes $\ddot{\eta}_{\theta_0}$ proportional to $\mathbf{e}_1 = (1, 0, \cdots, 0)'$ and $\dddot{\eta}_{\theta_0}$ a linear combination of $\mathbf{e}_1$ and $\mathbf{e}_2 = (0, 1, 0, \cdots, 0)'$. By (3.6), $\ddot{\lambda}_{\theta_0}$ is then also proportional to $\mathbf{e}_1$.

DEFINITION. The family $\mathscr{F}$ is in *standard form* at $\theta = \theta_0$ if $k = r$, the dimension of $\mathscr{F}$,

$$\eta_{\theta_0} = \lambda_{\theta_0} = 0\,, \qquad \Sigma_{\theta_0} = I_r \tag{4.1}$$

and

$$\dot{\eta}_{\theta_0} = \dot{\lambda}_{\theta_0} = i_{\theta_0}^{\frac{1}{2}}\mathbf{e}_1\,, \qquad \ddot{\eta}_{\theta_0} = \frac{\nu_{11}(\theta_0)}{i_{\theta_0}^{\frac{1}{2}}}\,\mathbf{e}_1 + i_{\theta_0}\gamma_{\theta_0}\mathbf{e}_2\,. \tag{4.2}$$

(The constants in (4.2) are necessary to satisfy (2.2).) We will use standard form to simplify proofs in Sections 9 and 10. If $\mathscr{F}$ is not in standard form at θ_0 the above transformation makes it so, and by property (ii) M_θ and hence all information and curvature properties remain unchanged. We could use property (i) to further standardize the situation so that $i_{\theta_0} = 1$, $\nu_{11}(\theta_0) = 0$, but that does not simplify any of the theoretical calculations which follow. Property (i) is useful for calculating curvatures, as will be shown in Section 7.

5. General definition of statistical curvature. Leaving exponential families, let

$$\mathscr{F} \equiv \{f_\theta(x),\, \theta \in \Theta\} \tag{5.1}$$

be an arbitrary family of density functions indexed by the single parameter $\theta \in \Theta$, a possibly infinite interval of the real line. The sample space $\mathscr{X}$ and carrier measure for the densities can be anything at all so we have not excluded the possibility that $\mathscr{F}$ consists of discrete distributions. Let

$$l_\theta(x) \equiv \log f_\theta(x)\,, \qquad \dot{l}_\theta(x) \equiv \frac{\partial}{\partial\theta}\, l_\theta(x)\,, \qquad \ddot{l}_\theta(x) \equiv \frac{\partial^2}{\partial\theta^2}\, l_\theta(x) \tag{5.2}$$

as in (3.15), (3.16). We assume the derivatives exist continuously and can be uniformly dominated by integrable functions in a neighborhood of the given θ, so that $E_\theta \dot{l}_\theta = 0$, $E_\theta \dot{l}_\theta^2 = -E_\theta \ddot{l}_\theta \equiv i_\theta$ as in (3.17). Finally, as in (3.20)–(3.21) we let M_θ be the covariance matrix of $(\dot{l}_\theta, \ddot{l}_\theta)$,

$$M_\theta \equiv \begin{pmatrix} \nu_{20}(\theta) & \nu_{11}(\theta) \\ \nu_{11}(\theta) & \nu_{02}(\theta) \end{pmatrix} \equiv \begin{pmatrix} E_\theta \dot{l}_\theta^2 & E_\theta \dot{l}_\theta \ddot{l}_\theta \\ E_\theta \ddot{l}_\theta \dot{l}_\theta & E_\theta \ddot{l}_\theta^2 - i_\theta^2 \end{pmatrix} \tag{5.3}$$

and define the *statistical curvature of* $\mathscr{F}$ *at* θ to be

$$\gamma_\theta \equiv (|M_\theta|/i_\theta^5)^{\frac{1}{2}} = \left[\frac{\nu_{02}(\theta)}{i_\theta^2} - \frac{\nu_{11}^2(\theta)}{i_\theta^3}\right]^{\frac{1}{2}}. \tag{5.4}$$

In making this definition we assume $0 < i_\theta < \infty$ and $\nu_{02}(\theta) < \infty$. Properties (i) and (ii) of Section 4 are verified to hold for γ_θ as defined in (5.4). Substituting $\dot{l}_\theta = \dot{f}_\theta/f_\theta$, $\ddot{l}_\theta = \ddot{f}_\theta/f_\theta - (\dot{f}_\theta/f_\theta)^2$ into (5.3), (5.4) shows that γ_θ^2 equals the bracketed term in (1.1), the crucial quantity in the Fisher–Rao theory.

What does γ_θ measure in this general situation? It is a measure of how quickly Fisher's score statistic is changing (more precisely, "turning") as θ changes. An argument along those lines is given next, further support coming in the calculations of Section 8.

Comparing (5.3) with (2.2), we can connect the two definitions by thinking of $\mathscr{S} \equiv \{l_\theta, \theta \in \Theta\}$ as a curve through the space of random variables on $\mathscr{X}$. The inner product $\langle u, v\rangle_\theta \equiv u'\Sigma_\theta v$ of (2.2) is taken to be the covariance inner product in (5.3). (Section 3 makes the analogy precise in the exponential family case.) All of the quantities in Figure 1 can now be given a statistical interpretation.

The element of arc length along $\mathscr{S}$, by analogy with (2.5), is $ds_\theta/d\theta = (E_\theta \dot{l}_\theta^2)^{\frac{1}{2}} = i_\theta^{\frac{1}{2}}$. Define

$$U_\theta(x) \equiv \frac{\dot{l}_\theta(x)}{i_\theta} + \theta . \tag{5.5}$$

U_{θ_0} is the version of Fisher's score statistic $\dot{l}_{\theta_0}$ that is the best locally unbiased estimator for θ near θ_0: $\mathrm{Var}_{\theta_0} U_{\theta_0} = 1/i_{\theta_0}$, the Cramér–Rao lower bound, and $E_{\theta_0} U_{\theta_0} = \theta_0$, $dE_\theta U_{\theta_0}/d\theta|_{\theta=\theta_0} = 1$. Therefore

$$\frac{(d/d\theta)E_\theta U_{\theta_0}}{(\mathrm{Var}_{\theta_0} U_{\theta_0})^{\frac{1}{2}}}\bigg|_{\theta=\theta_0} = \frac{ds_\theta}{d\theta}\bigg|_{\theta=\theta_0} . \tag{5.6}$$

(The quantity on the left of (5.6) is called the "efficacy" of the statistic U_{θ_0}.) We see that

$$(\theta - \theta_0) \cdot \frac{ds_\theta}{d\theta}\bigg|_{\theta=\theta_0} = \frac{E_\theta U_{\theta_0} - E_{\theta_0} U_{\theta_0}}{(\mathrm{Var}_{\theta_0} U_{\theta_0})^{\frac{1}{2}}} + O(\theta - \theta_0)^2 . \tag{5.7}$$

Therefore s_θ of Figure 1 can be interpreted locally as the number of (θ_0) standard deviations from $E_{\theta_0} U_{\theta_0}$ to $E_\theta U_{\theta_0}$.

By analogy with (2.6)

$$\sin a_\theta = \left[1 - \frac{[\mathrm{Cov}_{\theta_0}(\dot{l}_{\theta_0}, \dot{l}_\theta)]^2}{\mathrm{Var}_{\theta_0} \dot{l}_{\theta_0} \mathrm{Var}_{\theta_0} \dot{l}_\theta}\right]^{\frac{1}{2}} = [1 - \mathrm{corr}^2_{\theta_0}(\dot{l}_{\theta_0}, \dot{l}_\theta)]^{\frac{1}{2}} , \tag{5.8}$$

so $\sin^2 a_\theta$ is interpreted as the unexplained fraction of the variance in $U_\theta(x)$ after linear regression on $U_{\theta_0}(x)$, under density f_{θ_0}.

From (2.4) we get the following interpretation of the statistical curvature: *γ_{θ_0} is the derivative at $\theta = \theta_0$ of the unexplained fraction of the standard deviation of U_θ given U_{θ_0}, the derivative being taken with respect to the efficacy distance $(E_\theta U_{\theta_0} - E_{\theta_0} U_{\theta_0})/(\mathrm{Var}_{\theta_0} U_{\theta_0})^{\frac{1}{2}}$ along $\mathscr{S}$.* If this quantity is large then the locally best estimator (also the locally best test statistic) is changing quickly as θ changes and $\mathscr{S}$ is highly curved in a statistical sense. At the opposite extreme are one-parameter exponential families for which $a_\theta \equiv 0$, so U_θ is statistically equivalent to U_{θ_0} for all θ and θ_0. We pursue this interpretation of γ_θ in Section 8 to decide what constitutes a seriously large value of the curvature.

In a certain sense any smooth one-parameter family $\mathscr{S}$ can be embedded in a suitably large exponential family. Suppose at some point θ_0 in Θ, $\dot{l}_\theta$ is k times

differentiable. Consider the k-parameter exponential family

$$(5.9) \qquad g_\eta(x) \equiv \exp[l_{\theta_0}(x) + \eta_1 \dot{l}_{\theta_0}(x) + \eta_2 \ddot{l}_{\theta_0}(x) + \cdots + \eta_k l_{\theta_0}^{[k]}(x) - \psi(\eta)]\,,$$

$l_{\theta_0}^{[k]}(x) \equiv (\partial^k/\partial\theta^k) l_\theta(x)|_{\theta=\theta_0}$, $\psi(\eta)$ being chosen to make (5.9) integrate to one over $\mathscr{X}$ with respect to the carrying measure for $\mathscr{X}$. Choosing

$$\eta_\theta = \left((\theta - \theta_0), \frac{(\theta-\theta_0)^2}{2}, \cdots, \frac{(\theta-\theta_0)^k}{k!}\right)$$

gives a one-parameter family of densities $\bar{f}_\theta \equiv g_{\eta_\theta}$ approximating f_θ near $\theta = \theta_0$. If the Taylor expansion for l_θ converges at θ_0 this approximation becomes increasingly accurate as $k \to \infty$. For any value of $k \geqq 2$ definitions (5.3) and (3.21) show that $\tilde{M}_{\theta_0} = M_{\theta_0}$, so $\tilde{\imath}_{\theta_0} = i_{\theta_0}$ and $\tilde{\gamma}_{\theta_0} = \gamma_{\theta_0}$. It is reasonable to expect results proved in the context of curved exponential families to hold for sufficiently smooth nonexponential families, though no justifying theorem has been proved to this effect. This is in the same spirit as approximating an arbitrary family by a multinomial with a large number of categories, as in Fisher (1925) and Barnett (1966), but seems to make the approximation in a smoother way.

6. Repeated sampling. Suppose we sample $x_1, x_2, \cdots, x_n$ independently and identically distributed with density f_θ. We will use boldface letters to indicate quantities connected with the repeated sample, $\mathbf{x} \equiv (x_1, x_2, \cdots, x_n)'$, $\mathbf{l}_\theta(\mathbf{x}) \equiv \sum_{i=1}^n l_\theta(x_i)$, $\mathbf{U}_\theta(\mathbf{x}) = \dot{\mathbf{l}}_\theta(\mathbf{x})/\mathbf{i}_\theta + \theta$, etc. In particular

$$(6.1) \qquad \mathbf{M}_\theta = nM_\theta$$

since $\mathbf{M}_\theta$ is the covariance matrix of $(\dot{\mathbf{l}}_\theta(\mathbf{x}), \ddot{\mathbf{l}}_\theta(\mathbf{x})) = \sum_{i=1}^n (\dot{l}_\theta(x_i), \ddot{l}_\theta(x_i))$. Besides the familiar relationship $\mathbf{i}_\theta = n i_\theta$ this gives

$$(6.2) \qquad \boldsymbol{\gamma}_\theta = \frac{\gamma_\theta}{n^2}\,.$$

The curvature goes to zero at rate $1/n^2$ under repeated sampling. This makes sense since we know that linear methods work better in large samples.

In curved exponential families, (3.18)—(3.19) combine with $\mathbf{l}_\theta(\mathbf{x}) = \sum_{i=1}^n l_\theta(\mathbf{x}_i)$ to give

$$(6.3) \qquad \dot{\mathbf{l}}_\theta(\mathbf{x}) = n\dot{\gamma}_{\theta_0}'(\bar{x} - \lambda_\theta)\,, \qquad \ddot{\mathbf{l}}_\theta(\mathbf{x}) = n[\ddot{\gamma}_{\theta_0}'(\bar{x} - \lambda_\theta) - n i_\theta]\,,$$

$\bar{x} \equiv \sum_{i=1}^n x_i/n$ being the sufficient statistic for the complete family (3.1).

If the x_i are independent but not necessarily identically distributed we still have $\mathbf{l}_\theta(\mathbf{x}) = \sum_{i=1}^n l_\theta^{(i)}(x_i)$, the superscript indicating the distribution for x_i, and so $\mathbf{M}_\theta = \sum_{i=1}^n M_\theta^{(i)}$. This explains the simple form of $\mathbf{M}_\theta$ in Example 2 of Section 3.

7. Some examples. Before discussing the statistical properties of γ_θ we will expand our catalog of examples to include several nonexponential families. Those results illustrate some simple principles that make γ_θ easy to calculate in familiar statistical situations. In the first three examples we assume that the

densities given are with respect to Lebesgue measure on the real line, i.e., that we have just one observation of a continuous variable. For an independent, identically distributed (i.i.d) sample of size n the curvature is obtained from formula (6.2). This last remark applies also to Example 7, and to the examples of Section 3.

EXAMPLE 4. *Translation families.* Let $g(x)$ be a probability density function and $f_\theta(x) \equiv g(x-\theta)$. Also let $h(x) \equiv \log g(x)$. Then $\dot{l}_\theta(x) = -h^{(1)}(x-\theta)$, $\ddot{l}_\theta(x) = h^{(2)}(x-\theta)$, where $h^{(i)}(x) = d^i h(x)/dx^i$, so $E_\theta \dot{l}_\theta^i \ddot{l}_\theta^j = \int_{-\infty}^{\infty} [-h^{(1)}(x)]^i \times [h^{(2)}(x)]^j g(x)\,dx$. *Obviously M_θ and γ_θ do not depend on θ in a translation family.*

For the *t translation family*, f degrees of freedom,

$$g(x) = \frac{\Gamma\left(\frac{f+1}{2}\right)}{f^{\frac{1}{2}}\Gamma(\frac{1}{2})\Gamma\left(\frac{f}{2}\right)}\left(1 + \frac{x^2}{f}\right)^{-(f+1)/2} \tag{7.1}$$

we calculate

$$\nu_{20}(\theta) = i(\theta) = \frac{f+1}{f+3}, \tag{7.2}$$

$$\nu_{02}(\theta) = \frac{f+1}{f+3}\left[\frac{(f+2)(f^2+8f+19)}{f(f+5)(f+7)} - \frac{f+1}{f+3}\right]$$

and $\nu_{11}(\theta) = 0$ (by symmetry), giving

$$\gamma_\theta^2 = \frac{6[3f^2+18f+19]}{f(f+1)(f+5)(f+7)}, \tag{7.3}$$

a monotone decreasing function of f. Some values are as follows:

f	1	2	5	10	20	$\to \infty$
γ_θ^2	2.5	1.063	.306	.107	.0334	$\sim 18/f^2$

(7.4)

The case $f = 1$ is the *Cauchy translation family*, and the value $\frac{5}{2}$ for γ_θ^2 agrees with a closely related calculation in Fisher (1925).

For the *Gamma translation family*

$$f_\theta(x) = \frac{(x-\theta)^{a-1}e^{-(x-\theta)}}{\Gamma(a)}, \qquad x \geqq \theta \tag{7.5}$$

$a > 4$ a fixed constant, we calculate

$$M_\theta = \frac{1}{a-2}\begin{pmatrix} 1 & \frac{2}{(a-3)} \\ \frac{2}{(a-3)} & \frac{4a-10}{(a-2)(a-3)(a-4)} \end{pmatrix}, \tag{7.6}$$

$$\gamma_\theta^2 = \frac{2}{(a-3)^2}\,\frac{a-1}{a-4}.$$

(For $a \leqq 4$, ν_{02} is infinite.)

EXAMPLE 5. *Scale families.* $x \sim \theta \cdot z$ where z has a known density, $\theta \in \Theta = (0, \infty)$. If z is a positive random variable then $\log x = \log \theta + \log z$ is a translation family. By Section 4 the curvature will be the same for this family as for the original one, and by Example 4 it will not depend on θ: *For scale families γ_θ does not depend on θ.* (The argument above applied separately to the positive and negative axes gives the result in general. It can also be derived directly from (5.3).)

A particular example is the *normal with known coefficient of variation*, $x \sim \mathcal{N}(\theta, c\theta^2)$, c known. Here $x \sim \theta z$, $z \sim \mathcal{N}(1, c)$. We calculate $i_\theta = 2(c + \frac{1}{2})/(c\theta^2)$ and

$$\gamma_\theta^2 = \frac{c^2}{4(c + \frac{1}{2})^2}. \tag{7.7}$$

(Notice that $x \sim \mathcal{N}(\theta, c\theta)^2$ is a curved exponential family, $k = 2$.) The curvature is near 0 for all values of c, taking its maximum at $c = 1$:

c	$\frac{1}{4}$	$\frac{1}{2}$	1	2	4	$\to \infty$
γ_θ^2	.0370	.0625	.0740	.0640	.0439	$\sim 1/4c$

(7.8)

EXAMPLE 6. *Weibull shape parameter.* $f_\theta(x) = \theta x^{\theta-1} e^{-x^\theta}$ for $x \geqq 0$, $\theta \in \Theta = (0, \infty)$. That is $x \sim z^{1/\theta}$ where $P[z < z_0] = 1 - e^{-z_0}$ for $z_0 \geqq 0$. The transformation $\log x = 1/\theta \log z$ makes this a scale family in $1/\theta$, so once again γ_θ^2 does not depend on Θ. Taking $\theta = 1$ for convenience gives $\dot{l}_1(x) = (1 - x)\log x + 1$, $\ddot{l}_1(x) = -(x \log^2 x + 1)$, $E_1 \dot{l}_1^i \dot{l}_1^j = \int_0^\infty [\dot{l}_1(x)]^i [\ddot{l}_1(x)]^j e^{-x}\, dx$. Numerical integration gives

$$\gamma_\theta^2 = .704\,. \tag{7.9}$$

EXAMPLE 7. *Mixture problems.* $f_\theta(x) = (1 - \theta)g(x) + \theta h(x)$, g and h known densities on an arbitrary space $\mathcal{X}$. The parameter space Θ contains $[0, 1]$. We see that

$$\dot{l}_\theta = \frac{h - g}{g + \theta(h - g)}, \qquad \ddot{l}_\theta = -\dot{l}_\theta^2 \tag{7.10}$$

and for $\theta = 0$

$$\dot{l}_0 = r - 1\,, \qquad \ddot{l}_0 = -(r - 1)^2 \tag{7.11}$$

where $r(x) \equiv h(x)/g(x)$. Defining $\alpha_j \equiv E_0(r - 1)^j$ gives

$$M_0 = \begin{pmatrix} \alpha_2 & -\alpha_3 \\ -\alpha_3 & \alpha_4 - \alpha_2^2 \end{pmatrix}, \qquad \gamma_0^2 = \frac{\alpha_4 - \alpha_2^2}{\alpha_2^3} - \frac{\alpha_3^2}{\alpha_2^5}. \tag{7.12}$$

If g and h are normal densities, say $g(x) = \varphi(x) = e^{-x^2/2}/(2\pi)^{\frac{1}{2}}$, $h(x) = \varphi(x - \mu)$, we have $r(x) = \exp(\mu x - \mu^2/2)$ and

$$M_0 = \begin{pmatrix} \xi^2 - 1 & -[\xi^3 - 3\xi + 2] \\ -[\xi^3 - 3\xi + 2] & \xi^6 - 4\xi^5 - \xi^3 + 8\xi - 4 \end{pmatrix} \tag{7.13}$$

when $\xi \equiv e^{\mu^2}$. Therefore $i_0 = \xi - 1$ and

$$\gamma_0^2 = \xi^2(\xi + 1). \tag{7.14}$$

The curvature approaches 2 for μ near 0 but becomes enormous as μ increases,

(7.15)

μ	.5	.832	1	1.048	1.180	$\to \infty$
γ_0^2	4.84	24	74.68	108	320	$\sim e^{4\mu^2}$
i_0	.284	1	1.718	2	3	$\sim e^{\mu^2}$

8. Hypothesis testing. So far we have not tried to say what constitutes a "large" curvature—a value of γ_θ (or, in repeated sampling situations of γ_θ, the curvature based on all the data) of sufficient magnitude to undermine techniques based on linear approximations to the log likelihood function. We can obtain a rough idea of this value by considering the problem of testing $H_0\colon \theta = \theta_0$ versus $A_0\colon \theta > \theta_0$.

Define

$$\theta_1 \equiv \theta_0 + \frac{2}{i_{\theta_0}^{\frac{1}{2}}} \tag{8.1}$$

so that $i_{\theta_0}^{\frac{1}{2}}(\theta_1 - \theta_0) = 2$. From the discussion (5.5)—(5.7) this means that, approximately,

$$\frac{E_{\theta_1} U_{\theta_0} - E_{\theta_0} U_{\theta_0}}{(\mathrm{Var}_{\theta_0} U_{\theta_0})^{\frac{1}{2}}} = 2 \tag{8.2}$$

(where in (5.7) we have used $ds_\theta/d\theta|_{\theta=\theta_0} = i_{\theta_0}^{\frac{1}{2}}$). The locally most powerful level α test of H_0 versus A_0, LMP_α, for short, rejects for large values of U_{θ_0}. From (8.2) we would expect LMP_α to have reasonable power at θ_1 for the customary values of α. That is θ_1 should be a "statistically reasonable" alternative to θ_0.

The discussion following (5.8) shows that the unexplained fraction of the variance of U_{θ_1} after linear regression on U_{θ_0}, calculated under f_{θ_0}, is approximately $4\gamma_{\theta_0}^2$. If this quantity is large, say $4\gamma_{\theta_0}^2 \geqq \frac{1}{2}$, then U_{θ_1} differs considerably from U_{θ_0}, and the test of H_0 based on U_{θ_1} will substantially differ from that based on U_{θ_0}. Under these circumstances it is reasonable to question the use of LMP_α. *Based on those very rough calculations a value of* $\gamma_{\theta_0}^2 \geqq \frac{1}{8}$ *is "large".*

In the repeated sampling situation of Section 6 a sample of size $n > n_0$,

$$n_0 = 8\gamma_{\theta_0}^2, \tag{8.3}$$

makes $\gamma_{\theta_0}^2 = \gamma_{\theta_0}^2/n < \frac{1}{8}$, and therefore reduces the curvature below the worrisome point. For the Cauchy translation family, Example 4, $n_0 = 20$. For the Weibull shape parameter, Example 6, $n_0 = 5.6$. For the normal with known coefficient of variation, Example 5, $n_0 < 1$ for all c. At the opposite extreme we have the normal mixture problem, Example 7, with $\mu = 1$, for which $n_0 = 597.4$. We expect linear methods to work well in Example 5 for any sample size, and poorly in the last example, even for large samples.

Moving from the vague to the specific, consider Example 1, Section 3, a bivariate normal vector $x = (x_1, x_2)'$ with mean $(\theta, \gamma_0\theta^2/2)$ and covariance matrix I. Assume we wish to test $H_0\colon \theta = 0$ versus $A_0\colon \theta > 0$ on the basis of observing x. The LMP_α, which rejects for large values of x_1, has power function (probability of rejection) $1 - \beta_0(\theta) = \Phi(\theta - z_\alpha)$, where z_α and Φ are the upper α point and cdf of a standard normal variate.

The Neyman–Pearson lemma says that the most powerful level α test of $\theta = 0$ versus some specific positive alternative $\theta = \theta_1$, $\mathrm{MP}_\alpha(\theta_1)$ for short, rejects for large values of $\eta_{\theta_1}'x$. It has power function

$$1 - \beta_{\theta_1}(\theta) = \Phi(\theta(1 + \gamma_0^2\theta^2/4)^{\frac{1}{2}} \cos(A_{\theta_1} - A_\theta) - z_\alpha), \tag{8.4}$$

A_θ being the angle from the x_1 axis to η_θ, illustrated in Figure 2. As θ_1 approaches 0, $\beta_{\theta_1}(\theta)$ approaches $\beta_0(\theta)$ for all θ, justifying the notation $1 - \beta_0(\theta)$ for the power function of LMP_α.

For a given value of $\theta > 0$ the power is maximized by taking $\theta_1 = \theta$, giving "power envelope"

$$1 - \beta^*(\theta) = \Phi(\theta(1 + \gamma_0^2\theta^2/4)^{\frac{1}{2}} - z_\alpha). \tag{8.5}$$

Figure 3 compares the power envelope function, for four values of γ_0, with the power function of LMP_α, $\alpha = .05$ (which does not depend on γ_0). As predicted the difference between $1 - \beta^*(\theta)$ and $1 - \beta_0(\theta)$ increases with the curvature γ_0. In this case we can actually see that γ_{θ_0} measures how fast the locally optimum test statistic $U_{\theta_0}(x)$ becomes nonoptimal as the alternative θ increases from 0. Also according to prediction the LMP_α has reasonable power properties for $\gamma_0^2 = \frac{1}{16}$ and poor properties for $\gamma_0^2 \geqq \frac{1}{4}$.

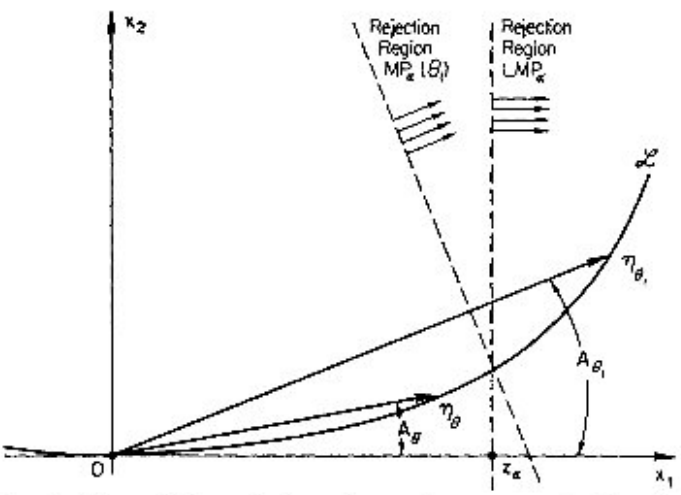

FIG. 2. Bivariate Normal, Example 1, testing $\theta = 0$ versus $\theta > 0$. The rejection region for the locally most powerful level α test, LMP_α, is compared with that for the most powerful level α test of θ versus θ_1, $\mathrm{MP}_\alpha(\theta_1)$.

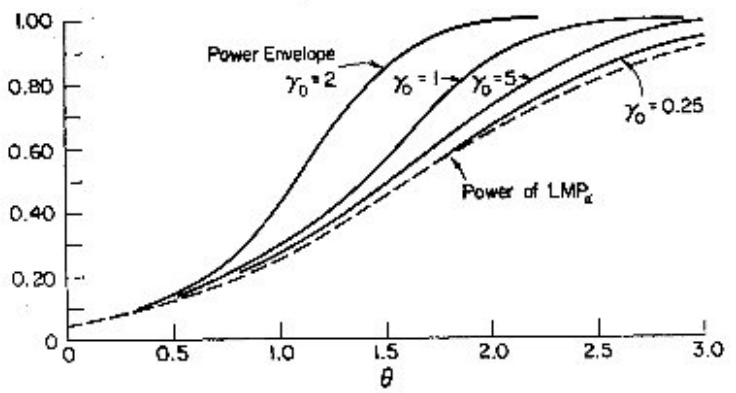

FIG. 3. Power of LMP_α, $\alpha = .05$, compared with power envelope function, Example 1.

TABLE 1

Power comparison, Example 1

a) *Power envelope* b) *Power* $MP_{.05}(2)$ c) *Power* $LMP_{.05}$

γ_0	θ							
	0	.5	1.0	1.5	2.0	2.5	3.0	3.5
.25	.05[a]	.126	.262	.453	.662	.835	.941	.985
.25	.05[b]	.125	.260	.452	.662	.834	.938	.983
.25	.05[c]	.126	.260	.442	.639	.804	.912	.968
.5	.05	.127	.269	.483	.723	.904	.982	.999
.5	.05	.121	.261	.479	.723	.901	.980	.998
.5	.05	.126	.260	.442	.639	.804	.912	.968
1.0	.05	.129	.300	.591	.882	.991	1.000	1.000
1.0	.05	.115	.280	.583	.882	.990	1.000	1.000
1.0	.05	.126	.260	.442	.639	.804	.912	.968
2.0	.05	.139	.409	.855	.998	1.000	1.000	1.000
2.0	.05	.115	.381	.850	.998	1.000	1.000	1.000
2.0	.05	.126	.260	.442	.639	.804	.912	.968

Of course no level α test can achieve the power envelope for more than one value of $\theta > 0$. $MP_\alpha(\theta_1)$ achieves it for $\theta = \theta_1$ while LMP_α optimizes for θ near 0 in the sense that $d\beta_\alpha(\theta)/d\theta|_{\theta=\theta_0} = d\beta^*(\theta)/d\theta|_{\theta=\theta_0}$. By following prescription (8.1) in choosing θ_1 we get a test which matches the power envelope at what should be a statistically interesting value of θ, one where the power is reasonably but not unreasonably high. In our example this means choosing $\theta_1 = 2$, since $i_0 = 1$. Table 1 shows that $1 - \beta_\alpha(\theta)$ stays remarkably close to $1 - \beta^*(\theta)$, and that $MP_{.05}(2)$ has better power characteristics than $LMP_{.05}$, especially for large values of γ_0.

Davies performs similar evaluations for the Neyman–Davies model of Example 2. The curvatures for the upper and lower cases graphed on page 532 of Davies (1969) are $\gamma_0^2 = .488$ and $\gamma_0^2 = .244$ respectively, while the two on page 533 are $\gamma_0^2 = .00629$ and $\gamma_0^2 = .0364$. Ignoring the "Wald's test" curve, one sees that

the magnitude of γ_0^2 is indeed a good predictor of the relative performance of LMP_α compared to $\mathrm{MP}_\alpha(\theta_1)$. His results are quite similar to those for our Example 1. (Davies chooses θ_1 so that $1 - \beta^*(\theta_1) = .8$. This is a more precise way of accomplishing what (8.1) is intended to do, but is computationally difficult in most situations.)

Section 10 shows that the Cramér–Rao lower bound for the variance of an unbiased estimator errs roughly by a factor of $1 + \gamma_\theta^2$, rejustifying the definition of $\gamma_\theta^2 \geqq \frac{1}{8}$ as a "large" curvature.

9. **The Fisher-Rao theorem.** We again assume an i.i.d. sample $x_1, x_2, \cdots, x_n$, as in Section 6. Result (1.1), originally stated by Fisher in his fundamental paper on estimation theory (1925) can be restated as

$$\lim_{n\to\infty} (\mathrm{i}_\theta - \mathrm{i}_\theta^{\hat\theta}) = i_\theta \gamma_\theta^2 \tag{9.1}$$

since γ_θ^2 equals the bracketed term in (1.1). (9.1) is derived from (1.1) and (5.4) by means of the relationships $\nu_{20}(\theta) = \mu_{20} = i_\theta$, $\nu_{11}(\theta) = \mu_{11} - \mu_{30}$, and $\nu_{02}(\theta) = \mu_{02} - 2\mu_{21} + \mu_{40} - \mu_{20}^2$, these following from (1.2), (5.3) and

$$\dot I_\theta = \dot f_\theta / f_\theta\,, \qquad \ddot I_\theta = \ddot f_\theta / f_\theta - (\dot f_\theta / f_\theta)^2 \tag{9.2}$$

To use Fisher's evocative language, asymptotically the MLE $\hat\theta(x_1, x_2, \cdots, x_n)$ extracts all but $i_\theta\gamma_\theta^2$ of the information in the sample $\mathbf{x} = (x_1, \cdots, x_n)'$. Since a single observation contains an amount i_θ of information this is equivalent to a reduction in effective sample size from n to $n - \gamma_\theta^2$, for example from n to $n - \frac{1}{2}$ in the Cauchy translation parameter problem. This is the price one pays for a one-dimensional summary of the data and, also according to Fisher, any summary statistic other than the MLE would pay a greater price. (Rao's substantial contributions to this argument are discussed toward the end of the section.)

The geometrical argument which follows shows clearly why the curvature γ_θ plays the role that it does in (9.1). It also leads quickly to a counterexample to (9.1) and shows that by working within multinomial families, Fisher and Rao chose perhaps the *least* tractable curved exponential families. We will work with a general curved exponential family in the *standard form* (4.1)—(4.2). For notational convenience we let θ_0, a particular value of θ where we wish to evaluate $\lim_{n\to\infty} (\mathrm{i}_\theta - \mathrm{i}_\theta^{\hat\theta})$, equal 0. Then we have $\eta_0 = \lambda_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_0 = \mathbf{I}_2$, $\dot\eta_0 = \dot\lambda_0 = i_0^{\frac{1}{2}}\mathbf{e}_1$, and $\ddot\eta_0 = (\nu_{11}(0)/i_0^{\frac{1}{2}})\mathbf{e}_1 + i_0\gamma_0\mathbf{e}_2$.

Fisher's argument depends on two useful results which we borrow:

1) If $T(\mathbf{x})$ is any statistic, with density say $\mathrm{f}_\theta^T(t)$ and score function (log derivative) $\dot{\mathrm{I}}_\theta^T(t) \equiv \partial/\partial\theta \log f_\theta^T(t)$, then $\dot{\mathrm{I}}_\theta^T(t) = E_\theta[\dot{\mathrm{I}}_\theta(\mathbf{x}) \mid T = t]$ (where we recall from Section 6 that $\dot{\mathrm{I}}_\theta(\mathbf{x})$ is the score based on all the data). This implies that the loss of information in going from $\mathbf{x}$ to $T(\mathbf{x})$ is

$$\mathrm{i}_\theta - \mathrm{i}_\theta^T = E_\theta \operatorname{Var}_\theta \{\dot{\mathrm{I}}_\theta(\mathbf{x}) \mid T\} \tag{9.3}$$

since $\mathrm{i}_\theta - \mathrm{i}_\theta^T = \operatorname{Var}_\theta \dot{\mathrm{I}}_\theta - \operatorname{Var}_\theta \dot{\mathrm{I}}_\theta^T$.

2) Let L_θ be the set of values of $\bar x \equiv \sum_{i=1}^n x_i/n$ for which $\dot{\mathrm{I}}_\theta(\mathbf{x}) = 0$; L_θ consists

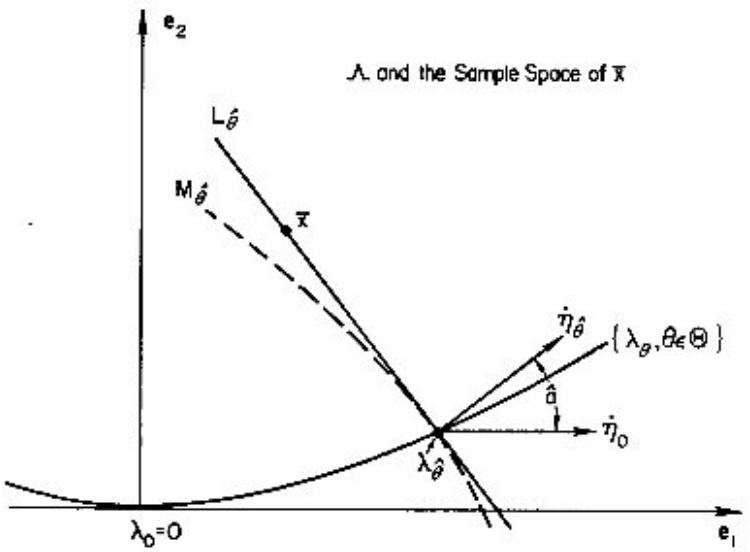

FIG. 4. A curved exponential family of dimension $r - 2$. $L_{\hat\theta}$ is the set of $\bar x$ for which $\hat\theta$ is a solution to the maximum likelihood equations. $M_{\hat\theta}$ is the level curve for another consistent efficient estimator.

of those values of the sufficient statistic $\bar x$ for which θ is a solution to the likelihood equations $\dot l_\theta(\mathbf{x}) = 0$. Then, since $\dot l_\theta = n\dot\gamma_\theta'(\bar x - \lambda_\theta)$,

$$L_\theta = \{\bar x : \dot\gamma_\theta'(\bar x - \lambda_\theta) = 0\} \tag{9.4}$$

the $r - 1$-dimensional hyperplane through λ_θ, orthogonal to $\dot\gamma_{\theta\theta}$.

Figure 4 illustrates the situation for the case $r = 2$. (Notice that the sample space, the space of possible $\bar x$ values, has been superimposed on A, the space of possible mean vectors λ.) Actually this two-dimensional picture is appropriate for any dimension since curvature is locally a two-dimensional property, as pointed out at the end of Section 2. A heuristic proof of (9.1) based on this picture now follows in five easy steps:

(i) $\dot l_0(\mathbf{x}) = n(i_0)^{\frac12}\bar x_1$ (by (6.3)).

(ii) $n^{\frac12}\bar x \to \mathcal{N}_r(\mathbf{0}, \mathbf{I})$ as $n \to \infty$ if $\theta = 0$ (since $\lambda_0 = \mathbf{0}$, $\mathbf{\Sigma}_0 = \mathbf{I}$, and central limit conditions are satisfied inside an exponential family).

(iii) Let $\hat\theta$ be the MLE and $\hat a$ the angle between $\dot\gamma_{\hat\theta\hat\theta}$ and $\dot\gamma_{\hat\theta 0} = i_0^{\frac12}\mathbf{e}_1$. Then $\hat a = i_0^{\frac12}\dot\gamma_0\hat\theta + O(\hat\theta^2)$. (Since $\eta_{\hat\theta\hat\theta} = \hat\theta\dot\gamma_{\hat\theta 0} - O(\hat\theta^2) = i_0^{\frac12}\hat\theta\mathbf{e}_1 + O(\hat\theta^2)$, the element of arclength in Figure 1 is $s_{\hat\theta} = ||\eta_{\hat\theta\hat\theta}|| + O(\hat\theta^2) = i_0^{\frac12}\hat\theta + O(\hat\theta^2)$. By (2.4) we have $a_{\hat\theta} \equiv \hat a = i_0^{\frac12}\dot\gamma_0\hat\theta + O(\hat\theta^2)$.)

(iv) $\mathrm{Var}_\theta\{\dot l_0(x)\,|\,\hat\theta\} = n^2 i_\theta \tan^2\hat a \cdot \mathrm{Var}_\theta\{\bar x_2\,|\,\hat\theta\}$. (In the case $r = 2$ this follows immediately from (i) and the geometry of the situation. For $r > 2$, $\bar x_2$ is replaced by $\nu'\bar x/||\nu||$ where $\nu = \dot\gamma_{\hat\theta\hat\theta} - ||\dot\gamma_{\hat\theta\hat\theta}||\cos\hat a \cdot \dot\gamma_{\hat\theta 0}$, the part of $\dot\gamma_{\hat\theta\hat\theta}$ orthogonal to $\dot\gamma_{\hat\theta 0}$.)

(v) $\mathrm{Var}_\theta\{\bar x_2\,|\,\hat\theta\} = 1/n + o(1/n)$. (This is plausible because of (ii) and the fact that near $\theta = \hat\theta$ the partition of the sample space generated by the "lines" L_θ looks like the partition generated by lines parallel to $L_{\hat\theta}$.)

Steps (iii) and (iv) together give $\mathrm{Var}_\theta\{\dot{l}_0 \mid \hat{\theta}\} = n^2 i_0^2 \gamma_0^2 \hat{\theta}^2 (1 + O(\hat{\theta}))\, \mathrm{Var}_0\{\bar{x}_2 \mid \hat{\theta}\}$, which, combined with (v), gives

$$\mathrm{Var}_\theta\{\dot{l}_0(x) \mid \hat{\theta}\} = n i_0^2 \gamma_0^2 \hat{\theta}^2 (1 + O(\hat{\theta}))(1 + o_n(1)) \tag{9.5}$$

$o_n(1) \to 0$ as $n \to \infty$, $O(\hat{\theta}) \to 0$ as $\hat{\theta} \to 0$. The heuristic proof of (9.1) is completed by (9.3), giving

$$\lim_{n\to\infty} l_0 - l_0^{\hat{\theta}} = \lim_{n\to\infty} E_0 \,\mathrm{Var}_0\{\dot{l}_0 \mid \hat{\theta}\} = i_0 \gamma_0^2 . \tag{9.6}$$

Here we have used

$$\lim_{n\to\infty} n F_0\{|\hat{\theta}|^3\} = 0 , \qquad \lim_{n\to\infty} n E_0 \hat{\theta}^2 = i_0^{-1} , \tag{9.7}$$

which one might hope for in view of $n^{\frac{1}{2}}\hat{\theta} \to \mathcal{N}(0, i_0^{-1})$.

All of the weak links in this chain of reasoning can be made solid except for (v). Its fatal flaw is shown by a counterexample to (9.1) based on the trinomial distribution

$$P\{\text{observed object is in category } j\} = \lambda_j , \qquad j = 1, 2, 3 \tag{9.8}$$
$$(\text{so } \lambda_j \geqq 0, \lambda_1 + \lambda_2 + \lambda_3 = 1) .$$

The trinomial can be considered as an exponential family of form (3.1) with $k = r = 2$; $\lambda = (\lambda_1, \lambda_2)'$, $\eta = (\eta_1, \eta_2)'$, $\eta_j = \log[\lambda_j/(1 - \lambda_1 - \lambda_2)]$, $j = 1, 2$, and $\phi(\eta) = \log(1 + e^{\eta_1} + e^{\eta_2})$. The x vector takes on three possible values: (1, 0), (0, 1), (0, 0), corresponding to the observed object being in the first, second, or third categories respectively. The carrier measure $m(\cdot)$ puts mass one at each of these three x values.

The counterexample is a one parameter family $\mathcal{F}$ with L_θ passing through the fixed point $c = (-2^{\frac{1}{2}}, -1)$ as illustrated in Figure 5, the parameter θ being

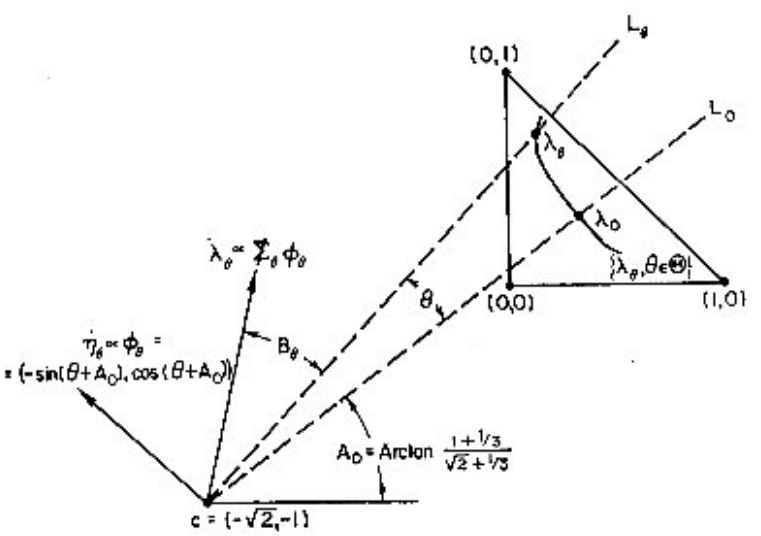

FIG. 5. Counterexample to (9.1) based on trinomial. Each line L_θ contains at most one possible sample point $\bar{x}$.

the angle between L_0 and L_θ. Such a family does exist, as the following construction shows: let $\lambda_0 \equiv (\frac{1}{3}, \frac{1}{3})$ and

$$\lambda_{\theta_0} \equiv \lambda_0 + \int_0^{\theta_0} \mu_\theta(\dot{\Sigma}_\theta \phi_\theta)\, d\theta \tag{9.9}$$

where $\mu_\theta \equiv ||\lambda_\theta - c||/(||\dot{\Sigma}_\theta \phi_\theta|| \sin B_\theta)$, $\dot{\Sigma}_\theta$ is the covariance matrix of x under f_θ, the vector ϕ_θ and the angle B_θ being defined as in Figure 5. Definition (9.9) gives $\lambda_\theta \in L_\theta$ and also that, by (3.6), $\dot{\eta}_\theta \propto \phi_\theta$, the normal vector to L_θ, as necessitated by (9.4).

$\mathcal{F}$ is a curved exponential family having the following property: if $\bar{x}_{(1)}$ and $\bar{x}_{(2)}$ are two values of $\bar{x} \equiv \sum_1^n x_i/n$ giving the same MLE $\hat{\theta}$, then both $\bar{x}_{(1)}$ and $\bar{x}_{(2)}$ lie on $L_{\hat{\theta}}$. But $\bar{x}_{(i)} = (n_{1(i)}/n, n_{2(i)}/n)$, $i = 1, 2$ the $n_{j(i)}$ being nonnegative integers. This implies either $\bar{x}_{(1)} = \bar{x}_{(2)}$ or

$$\frac{n_{2(2)} - n_{2(1)}}{n_{1(2)} - n_{1(1)}} = \frac{n_{2(1)} + 1 \cdot n}{n_{1(1)} + 2^1 \cdot n}. \tag{9.10}$$

Since (9.10) would make 2^1 a rational number, $\bar{x}_{(1)}$ must equal $\bar{x}_{(2)}$. In short there is at most one possible $\bar{x}$ value corresponding to any $\hat{\theta}$, and so the MLE is a sufficient statistic in $\mathcal{F}$, implying $\mathbf{i} - \mathbf{i}^{\hat{\theta}} = 0$ for all n. But γ_θ^2 must be positive for all θ values since $\dot{\eta}_\theta$ is always changing direction. This completes the counterexample.

REMARK 1. Let $\varphi(t) \equiv E_0 e^{it'x}$ be the characteristic function of f_0. If $|\varphi(t)|^p$ is integrable for some $p \geqq 1$ then $n^{\frac{1}{2}}\bar{x}$ has a density function converging uniformly to $(2\pi)^{-k/2} \exp(-\|x\|^2/2)$. See Efron and Truax (1968), Gnedenko and Kolmogorov (1954). Under those conditions (9.1) can be verified. The technical details, which depend on an exponential bound to the density of $\bar{x}$, are indicated in the Appendix.

REMARK 2. Instead of working with the MLE $\hat{\theta}$ itself we can consider the coarser statistic which only records which interval $\hat{\theta}$ lies in, among intervals of the form $(i\varepsilon_n, (i+1)\varepsilon_n)$, $i = 0, \pm 1, \pm 2, \cdots$. The line $L_{\hat{\theta}}$ in Figure 4 is now replaced by a pair of lines $L_{i\varepsilon_n}$, $L_{(i+1)\varepsilon_n}$, and step (v) can be weakened to say only that the conditional distribution of $\bar{x}_2$, given that $\bar{x}$ is between the two lines, has variance $1/n + o(1/n)$. However in order for statement (iv) to still have meaning we need to take $\varepsilon_n = o(1/n)$ (so that the conditional variance of $\dot{\mathbf{l}}_0$ will still be due mainly to the slope of the lines $L_{i\varepsilon_n}$, $L_{(i+1)\varepsilon_n}$, and not to the distance between them). It turns out (Efron and Truax (1968)) to be possible to choose ε_n in this way and to get the proper convergence of the conditional variance if f_0 is nonlattice, $|\varphi(t)| < 1$ for all $t \neq 0$. (This excludes the multinomial.) In this case it is possible to show that $\limsup_{n\to\infty}(\mathbf{i}_\theta - \mathbf{i}_\theta^{\hat{\theta}}) \leqq i_\theta \gamma_\theta^2$.

REMARK 3. If $\bar{\theta}(\bar{x})$ is any other consistent efficient estimator of θ, and M_θ is the set of $\bar{x}$ values having $\bar{\theta}(\bar{x}) = \theta$, then as in Figure 4, $M_{\hat{\theta}}$ passes through $\lambda_{\hat{\theta}}$ and is tangent to $L_{\hat{\theta}}$ at that point. See Section 10. The increment of $[\lim_{n\to\infty}(\mathbf{i}_\theta - \mathbf{i}_\theta^{\bar{\theta}}) - i_\theta \gamma_\theta^2]$ above zero is due to the quadratic term in the expansion

of $M_{\hat\theta}$ near $\lambda_{\hat\theta}$. The details are almost identical to those of Section 10 and will not be given here. (See (10.25).)

REMARK 4. It is possible for two of the surfaces (9.4), say L_{θ_0} and $L_{\hat\theta}$, to intersect. If $\bar x \in L_{\theta_0} \cap L_{\hat\theta}$ then both 0 and $\hat\theta$ are solutions to the likelihood equation. As $\hat\theta$ decreases to zero in Figure 4, $I_{\theta_0} \cap L_{\hat\theta}$ converges to a point (in general an $r-2$ dimensional flat) on $L_{\theta_0} = \{ce_2\}$ a distance $\rho_{\theta_0} \equiv 1/\gamma_{\theta_0}$ above $\mathbf{0}$. Values of $\bar x$ on L_{θ_0} which lie above this point are local *maxima* of the likelihood function, while those lying below are local *minima*.

REMARK 5. Rao (1961, 1962, 1963) uses a different definition of the information which avoids the difficulty illustrated by the counterexample. (9.3) can be written as $\mathbf{i}_\theta - \mathbf{i}_\theta^T = \inf E_\theta\{\dot l_\theta(x) - h(T(x))\}^2$, the infimum being over all choices of the function $h(\cdot)$. Rao redefines $\mathbf{i}_\theta^T$ by restricting the function h to be quadratic. Rao states that he believes the two definitions to be equivalent, but the counterexample can be used to show that they are not.

REMARK 6. Is (9.1) a useful fact, assuming it is true? Fisher seemed to think of Fisher information as a perfect measure of the amount of information available to the statistician. For ordinary "first order efficiency" calculations in large samples this is true enough, in the following sense: let $T(x)$ be a statistic having Fisher information $\mathbf{i}_\theta^T$. Then in a neighborhood of any given value θ_0 of θ we can construct, under suitable regularity conditions, a function $\tilde T(T)$, that is approximately $\mathcal{N}(\theta, 1/\mathbf{i}_\theta^T)$, as compared with $\mathcal{N}(\theta, 1/\mathbf{i}_\theta^{\hat\theta})$ for the MLE. If $\mathbf{i}_\theta^T/\mathbf{i}_\theta^{\hat\theta} = .8$ for example, then any statistic $h(\tilde T)$ will have almost the same distribution as $h(\hat\theta)$ with $\hat\theta$ based on a sample 80% as large.

This argument breaks down for information discrepancies as small as those contemplated in (9.1), since the central limit theorem is in general not capable of supporting such fine distinctions. To give substance to Fisher and Rao's theorem we must demonstrate that in specific statistical problems the Fisher information determines relative performance at the level of accuracy suggested by (9.1). Rao (1963) showed that this indeed was the case for the problem of estimating θ with squared error loss. We review his results from the point of view of this paper in Section 10.

10. Estimation with squared error loss. Suppose we wish to estimate the parameter θ in a curved exponential family on the basis of an i.i.d. sample $x_1, x_2, \cdots, x_n$, using a squared error loss function to evaluate possible estimators. We will only consider estimators that are smooth functions of the sufficient statistic $\bar x$ and are consistent and efficient in the usual sense (see (10.5)—(10.7) below). The following result will be discussed: let $\bar\theta(\bar x)$ be such an estimator, the form of $\bar\theta$ not depending on n, and let $\phi(\theta) \equiv E_\theta \mathbf{U}_{\theta_0}(\bar x)$ where as before $\mathbf{U}_{\theta_0}(\bar x) \equiv \dot l_{\theta_0}/\mathbf{i}_{\theta_0} + \theta_0$ is the best locally unbiased estimator of θ near θ_0. Also let $b_\theta \equiv E_\theta\bar\theta(\bar x) - \theta$ be the bias of $\bar\theta$, a quantity which will turn out to be of order

$O(1/n)$ in the theory below. Then

$$\text{(10.1)}\qquad \operatorname{Var}_{\theta_0}\bar{\theta} = \frac{1}{ni_{\theta_0}} + \frac{1}{n^2 i_{\theta_0}}\left\{\gamma^2_{\theta_0} + 4\frac{\Gamma^2_{\theta_0}}{i_{\theta_0}} + \Delta^{\bar{\theta}}_{\theta_0}\right\} + 2\frac{b_{\theta_0}}{ni_{\theta_0}} + o\left(\frac{1}{n^2}\right)$$

where $\Delta^{\bar{\theta}}_{\theta_0} \geqq 0$ and for the MLE $\hat{\theta}$, $\Delta^{\hat{\theta}}_{\theta_0} \equiv 0$. The quantity Γ_{θ_0} is the ordinary curvature at $\theta = \theta_0$ of the two-dimensional curve $(\theta, \phi(\theta))$ as defined at (2.1).

Before verifying (10.1) several remarks are pertinent.

1) The term $1/ni_{\theta_0}$ is the Cramér–Rao lower bound for the variance of an unbiased estimator. The bracketed quantity in (10.1) expresses the coefficient of the $1/n^2 i_{\theta_0}$ term as the sum of three nonnegative quantities: $\gamma^2_{\theta_0}$, the statistical curvature, which is invariant under transformations of θ; $4\Gamma^2_{\theta_0}/i_{\theta_0}$, the "naming curvature", which depends on how $\mathscr{F}$ is parametrized (however, notice that $4\Gamma^2_{\theta_0}/i_{\theta_0}$ is invariant under *linear* reparametrizations $\theta \to \alpha + \beta\theta$); and $\Delta^{\hat{\theta}}_{\theta_0}$, which can be made zero by using the MLE. Taken literally (10.1) says that the MLE is superior to other efficient estimations with the same bias structure.

2) The estimators $\bar{\theta}$ will generally be biased by an amount of order $1/n$. This affects mean square error to order $1/n^2$. A simple adjustment, noted below at Remark 11, produces estimators biased only to order $1/n^2$; (10.1), with the bias term $2b_{\theta_0}/ni_{\theta_0}$ removed, is valid for such estimators. Among such bias corrected estimators, (10.1) says that the MLE has asymptotically smallest variance.

3) The Fisher information is essentially invariant under reparametrizations of $\mathscr{F}$, in the sense that if $\mu = \mu(\theta)$ is a differentiable monotonic function then $\mathbf{i}_\mu{}^T = \mathbf{i}_\theta{}^T(d\theta/d\mu)^2$ for every statistic $T(x)$. The squared error estimation problem is *not* invariant under reparametrization and this accounts for the presence of the $4\Gamma^2_{\theta_0}$ term in (10.1). For a given θ_0 the "best" parametrization is in terms of $\phi(\theta)$, the expectation of the best locally unbiased estimator of θ. (Notice that ϕ will be the same, except for scale and translation constants, no matter what "θ" we begin with.) It will turn out that if the MLE $\hat{\theta}$ is unbiased for θ then $\phi \equiv \theta$ for all choices of θ_0, so we are automatically using the best parametrization.

4) (10.1) is not a special case of the Bhattacharyya lower bounds. The second Bhattacharyya bound, applying to estimators biased by amount $O(1/n^2)$ or less, is of the form

$$\text{(10.2)}\qquad \operatorname{Var}_{\theta_0}\bar{\theta} \geqq \frac{1}{ni_{\theta_0}} + \frac{1}{n^2 i_{\theta_0}}\left\{\frac{4\Gamma^2_{\theta_0}}{i_{\theta_0}}\right\} + O\left(\frac{1}{n^3}\right),$$

and the higher Bhattacharyya bounds are identical until order $O(1/n^3)$, so these bounds relate only to the naming part of the estimation problem. It is possible for an estimator to achieve equality in (10.2), but then it cannot be efficient in a neighborhood of θ_0, so (10.1) is not contradicted.

5) Even if $\mathscr{F}$ is not a curved exponential family we can use (10.1) to get an improved approximation to $\operatorname{Var}_{\theta_0}\bar{\theta}$, compared with the Cramér–Rao lower bound $1/ni_{\theta_0}$. The Cauchy translation family discussed at (7.4) has $i_{\theta_0} = \frac{1}{2}$, $\gamma^2_{\theta_0} = \frac{5}{2}$. The MLE $\hat{\theta}$ is unbiased in this case, so $\Gamma^2_{\theta_0} = 0$ and (10.1) is of the form $\operatorname{Var}_{\theta_0}\hat{\theta} = 1/ni_0 + \gamma^2_{\theta_0}/n^2 i_{\theta_0} + O(1/n^3)$. Numerical comparisons of this formula with the

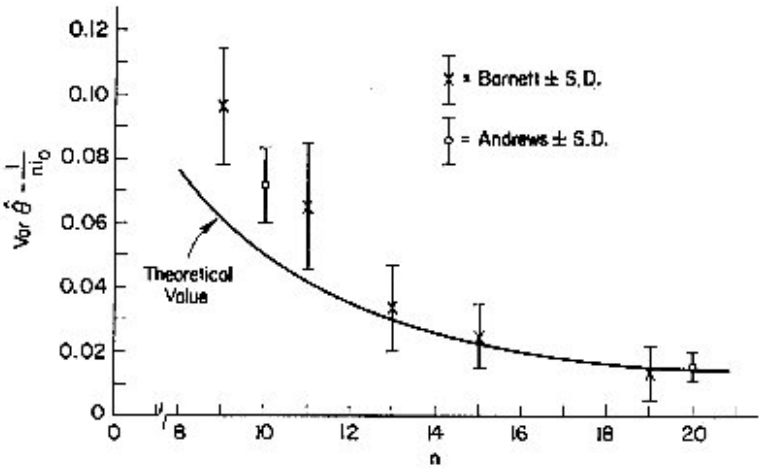

FIG. 6. Variance of MLE minus Cramér–Rao lower bound, for estimating the Cauchy translation parameter. Theoretical value from (10.1) compared with Monte Carlo results.

Monte Carlo studies of Barnett (1966) and also of Andrews et al. (1972) are shown in Figure 6. The theoretical values are obviously too small for $n \leqq 11$, but seem to be more accurate than the Monte Carlo results for $n \geqq 13$. For $n = 40$ Andrews et al. estimate $\mathrm{Var}_{\theta_0}\hat{\theta} - 1/ni_{\theta_0} = .0025 \pm .0017$ while (10.1) gives .0031.

6) For estimating a translation parameter Pitman's estimator is known to have smaller variance than the MLE. However, (10.1) suggests that this effect must be of magnitude at most $O(1/n^3)$.

7) Nothing in (10.1), except the application to general curved exponential families, is new. Rao (1963) states the result for curved multinomial families, and notes that for the MLE it was previously derived by Haldane and Smith (1956). The identification of the bracketed terms with curvatures is new, as well as the line of proof which leads to a rigorous verification.

8) The similarity of (9.1) and (10.1) can be viewed as a vindication of the belief that Fisher information is an accurate measure of the information contained in a given statistic. This conclusion is premature; the squared error estimation problem is very closely related to the information calculation, a fact which would be more obvious if we had presented a geometric argument below, as in Section 9, instead of using analytic methods. It is more reasonable to say that the curvature γ_θ is the leading term defining the nonlinearity of a family $\mathscr{F}$, and must play a central role in all calculations like (9.1) and (10.1). On the other hand in the absence of evidence to the contrary it seems difficult to dispute Fisher and Rao's assertion that the MLE provides the most informative one-dimensional summary statistic even when there is no one-dimensional sufficient statistic.

Our derivation of (10.1) will be done with the curved exponential family $\mathscr{F}$

in standard form, and assuming $\theta_0 = 0$. Neither of these assumptions affects the generality of the result. (The transformation to standard form maps any estimator into an estimator having the same variance, and leaves the quantities i_{θ_0}, γ_{θ_0}, and Γ_{θ_0} unchanged.) We assume that the estimator $\tilde{\theta}(\bar{x})$ has continuous third partial derivatives with respect to the components of $\bar{x}$, so that around $\bar{x} = 0$ it has the Taylor's series expansion

$$\tilde{\theta}(\bar{x}) = a_0 + \mathbf{a}'\bar{x} + (\bar{x}'\mathbf{A}\bar{x})/2 + O(\bar{x}^3)\,, \tag{10.3}$$

where a_0 is a scalar, $\mathbf{a}$ is a $r \times 1$ vector, and $\mathbf{A}$ an $r \times r$ matrix, r being the dimension of the full exponential family containing $\mathcal{F}$.

Here $O(\bar{x}^3)$ indicates a term that near the origin is bounded in absolute value by some polynomial in the components of $\bar{x}$ containing only terms of order 3.

Differentiating (10.3) with respect to the components of $\bar{x}$ gives the gradient vector

$$\nabla\tilde{\theta}(\bar{x}) = \mathbf{a} + \mathbf{A}\bar{x} + O(\bar{x}^2)\,. \tag{10.4}$$

In order for $\tilde{\theta}$ to be consistent and efficient, (10.3) must have the special form shown in the lemma:

LEMMA. *A consistent, efficient estimator $\tilde{\theta}(\bar{x})$, having continuous third partial derivatives near $\bar{x} = \mathbf{0}$, has the Taylor series expansion*

$$\tilde{\theta}(\bar{x}) = \frac{\bar{x}_1}{i_0^{\frac{1}{2}}} - \frac{\mu_{111}}{i_0^3}\frac{\bar{x}_1^2}{2} + \frac{\gamma_{0\cdot 1}}{i_0^{\frac{1}{2}}}\bar{x}_1\bar{x}_2 + \frac{\bar{x}'_{(1)}\mathbf{A}_{(1)}\bar{x}_{(1)}}{2} + O(\bar{x}^3) \tag{10.5}$$

assuming $\mathcal{F}$ is in standard form at $\theta = 0$. Here $\bar{x}_i$ indicates the ith component of $\bar{x}$, $\bar{x}_{(1)} \equiv (\bar{x}_2, \bar{x}_3, \cdots, \bar{x}_r)$, and $\mathbf{A}_{(1)}$ is the matrix $\mathbf{A}$ with its first row and column removed. For the MLE $\hat{\theta}(\bar{x})$, $\mathbf{A}_{(1)} = \mathbf{0}$. *As in* (1.1), $\mu_{111} = E_0\dot{f}_0\ddot{f}_0/f_0^2$.

The proof of the lemma is based on two simple facts: in order for a continuous estimator $\tilde{\theta}(\bar{x})$ to be consistent it must have "Fisher consistency",

$$\tilde{\theta}(\lambda_\theta) = \theta\,, \tag{10.6}$$

since $\bar{x} \to_p \lambda_\theta$ under repeated independent sampling from f_θ. Moreover, letting

$$\nabla_\theta \equiv \nabla\tilde{\theta}(\bar{x})|_{\bar{x}=\lambda_\theta}\,,$$

$$\lim_{n\to\infty}\frac{1}{\mathbf{i}_\theta \operatorname{Var}_\theta \tilde{\theta}} = \frac{(\dot{\eta}_\theta'\boldsymbol{\Sigma}_\theta\nabla_\theta)^2}{(\dot{\eta}_\theta'\boldsymbol{\Sigma}_\theta\dot{\eta}_\theta)(\nabla_\theta'\boldsymbol{\Sigma}_\theta\nabla_\theta)} \tag{10.7}$$

so $\tilde{\theta}$ will be first order efficient at θ, ($\lim_{n\to\infty} \mathbf{i}_\theta \operatorname{Var}_\theta \tilde{\theta} = 1$), if and only if

$$\nabla_\theta \equiv \nabla\tilde{\theta}(\bar{x})|_{\bar{x}=\lambda_\theta} = c_\theta\dot{\eta}_\theta \tag{10.8}$$

for some scalar c_θ. Taken together (10.6) and (10.8) say that the level surface $M_\theta \equiv \{\bar{x}: \tilde{\theta}(\bar{x}) = \theta\}$ of an efficient consistent estimator $\tilde{\theta}$ must cross $\{\lambda_\theta, \theta \in \Theta\}$ at λ_θ, and at that point must be parallel to the level surface (9.4) of the MLE, as shown in Figure 4. (10.7) merely says that the linear term in the expansion of $\tilde{\theta}(\bar{x})$ about λ_θ, $\tilde{\theta} = \theta + \nabla_\theta'(\bar{x} - \lambda_\theta) + O((\bar{x} - \lambda_\theta)^2)$, must be proportional to

the score statistic $\dot{\mathbf{l}}_\theta = n\dot{\gamma}_\theta'(\bar{x} - \lambda_\theta)$ in order to get first order efficiency. A proof follows from a greatly simplified version of the argument below, but the result is well known and will not be derived here.

The proof of (10.5) is obtained by seeing what form of (10.3) is necessary in order that (10.6) and (10.8) hold for λ_θ near 0. We will need the Taylor series expansions

$$\dot{\gamma}_\theta = i_0^{\frac{1}{2}}\mathbf{e}_1 + \left[\frac{\nu_{11}}{i_0^{\frac{1}{2}}}\mathbf{e}_1 + i_0\gamma_0\mathbf{e}_2\right]\theta + o(\theta)\,, \tag{10.9}$$

$$\lambda_\theta = i_0^{\frac{1}{2}}\mathbf{c}_1\theta + O(\theta^2)$$

and a more accurate expansion for the first component of λ_θ,

$$\mathbf{e}_1'\lambda_\theta = i_0^{\frac{1}{2}}\theta + \frac{\mu_{11}}{i_0^{\frac{1}{2}}}\frac{\theta^2}{2} + o(\theta^2)\,. \tag{10.10}$$

(10.9) follows from the standard form relationships (4.1)—(4.2). To prove (10.10) notice that $\mathbf{e}_1'\lambda_\theta = E_\theta x_1 = (1/i_0^{\frac{1}{2}})E_\theta \dot{l}_0(x)$ (see (3.18)). Formally

$$\begin{aligned} E_\theta \dot{l}_0 &= \int_{\mathcal{X}} \frac{\dot{f}_0}{f_0}(x)\left[f_0(x) + \theta\dot{f}_0(x) + \frac{\theta^2}{2}\ddot{f}_0(x) + o(\theta^2)\right]dm(x) \\ &= i_0\theta + \frac{\mu_{11}\theta^2}{2} + o(\theta^2)\,, \end{aligned} \tag{10.11}$$

a result which is easy to verify rigorously in an exponential family.

(10.4), (10.9), and (10.8) combine to give (writing $c_\theta = c_0 + \dot{c}_0\theta + o(\theta)$)

$$\begin{aligned} \mathbf{a} + \mathbf{A}(i_0^{\frac{1}{2}}\mathbf{e}_1\theta) + O(\theta^2) \\ = c_0 i_0^{\frac{1}{2}}\mathbf{e}_1 + \left[\dot{c}_0 i_0^{\frac{1}{2}}\mathbf{e}_1 + \frac{c_0\nu_{11}(0)}{i_0^{\frac{1}{2}}}\mathbf{e}_1 + c_0 i_0\gamma_0\mathbf{e}_2\right]\theta + o(\theta) \end{aligned} \tag{10.12}$$

implying

$$\mathbf{a} = c_0 i_0^{\frac{1}{2}}\mathbf{e}_1 \tag{10.13}$$

and

$$i_0^{\frac{1}{2}}\mathbf{A}\mathbf{c}_1 = \left(\dot{c}_0 i_0^{\frac{1}{2}} + \frac{c_0\nu_{11}(0)}{i_0^{\frac{1}{2}}}\right)\mathbf{e}_1 + c_0 i_0\gamma_0\mathbf{e}_2\,. \tag{10.14}$$

Notice that (10.14) shows that

$$A_{31} = A_{41} = \cdots = A_{r1} = 0\,. \tag{10.15}$$

(10.9), (10.10), (10.13), (10.6), and (10.3) combine to give

$$\theta = a_0 + c_0 i_0^{\frac{1}{2}}\left[i_0^{\frac{1}{2}}\theta + \frac{\mu_{11}}{i_0^{\frac{1}{2}}}\frac{\theta^2}{2}\right] + \frac{i_0 A_{11}}{2}\theta^2 + o(\theta^2)\,, \tag{10.16}$$

implying

$$a_0 = 0\,, \tag{10.17}$$

$c_0 = 1/i_0$, and $c_0\mu_{11} + i_0 A_{11} = 0$. Therefore

$$\mathbf{a} = \frac{1}{i_0^{\frac{1}{2}}}\mathbf{e}_1\,, \qquad A_{11} = -\frac{\mu_{11}}{i_0^2}\,, \qquad A_{21} = \frac{\gamma_0}{i_0^{\frac{1}{2}}}\,, \tag{10.18}$$

the first of these following from (10.13), the last from (10.14). Taken together, (10.15), (10.17) and (10.18) are equivalent to (10.5). Finally, for the MLE, $\hat{\theta}((0, \bar{x}_{(1)})') = 0$, implying $\mathbf{A}_{(1)} = 0$. This completes the proof of (10.5).

Two more simple results give (10.1) from (10.5). First of all, the Cramér–Rao lower bound for the variance of a possibly biased estimator $T(\bar{x})$ can be rewritten as an equality in the following useful form:

$$E_0 T^2 = \frac{1}{n i_0} + E_0\left(T - \frac{\bar{x}_1}{i_0^2}\right)^2 + 2\frac{b_0}{n i_0}. \tag{10.19}$$

((10.19) follows from $\mathrm{Cov}_0(T, \dot{\mathrm{l}}_0) = 1 + b_0'$.) Notice that $\dot{\mathrm{l}}_0/\mathrm{i}_0 = \bar{x}_1/i_0^2$ by (6.3) so this statistic is just the best locally unbiased estimator of θ, $\mathbf{U}_0$, introduced at (5.5). For an unbiased estimator, (10.19) says that $\mathrm{Var}_0 T$ exceeds the Cramér–Rao lower bound by the expected squared error of T in predicting $\mathbf{U}_0$. In a curved exponential family the regularity conditions necessary for (10.19) are satisfied if $E_\theta T^2 < \infty$ for θ in a neighborhood of 0. The second fact needed is that if z is standard multivariate normal, $z \sim \mathcal{N}_r(\mathbf{0}, \mathbf{I})$, and $\mathbf{A}$ is an $r \times r$ symmetric matrix, then $E(z'\mathbf{A}z)/2 = \mathrm{tr}\,\mathbf{A}/2$ and

$$\mathrm{Var}\frac{z'\mathbf{A}z}{2} = \tfrac{1}{2}\,\mathrm{tr}\,\mathbf{A}^2. \tag{10.20}$$

As $n \to \infty$, $z_n \equiv n^{\frac{1}{2}}\bar{x} \to \mathcal{N}_r(\mathbf{0}, \mathbf{I})$, and because f_0 is inside an exponential family the moments of z_n converge to the moments of $z \sim \mathcal{N}_r(\mathbf{0}, \mathbf{I})$. Ignoring the $O(\bar{x}^3)$ term, an omission justified (under an additional restriction on $\tilde{\theta}$) in Remark 12 below, (10.3) and (10.5) give

$$E_0\tilde{\theta} = E_0\frac{\bar{x}'\mathbf{A}\bar{x}}{2} = \frac{1}{n}\frac{\mathrm{tr}\,\mathbf{A}}{2} = \frac{1}{n}\left(-\frac{\mu_{21}}{2i_0^2} + \frac{\mathrm{tr}\,\mathbf{A}_{(1)}}{2}\right). \tag{10.21}$$

Moreover (10.5) combines with (10.19) and (10.20) to give

$$E_0\tilde{\theta}^2 = \frac{1}{n i_0} + \frac{1}{n^2}\left(\frac{\gamma_0^2}{i_0} + \frac{\mu_{21}^2}{2i_0^4} + \frac{\mathrm{tr}\,\mathbf{A}_{(1)}^2}{2} + \frac{\mathrm{tr}^2\,\mathbf{A}}{4}\right) + \frac{2b_0}{n i_0} + o\left(\frac{1}{n^2}\right). \tag{10.22}$$

Therefore,

$$\mathrm{Var}_0\tilde{\theta} = \frac{1}{n i_0} + \frac{1}{n^2}\left(\frac{\gamma_0^2}{i_0} + \frac{\mu_{21}^2}{2i_0^4} + \frac{\mathrm{tr}\,\mathbf{A}_{(1)}^2}{2}\right) + \frac{2b_0}{n i_0} + o\left(\frac{1}{n^2}\right). \tag{10.23}$$

Finally, (10.11) gives $\phi(\theta) = \theta + (\mu_{11}/2i_0)\theta^2 + o(\theta^2)$, where $\phi(\theta) = E_\theta\dot{\mathrm{l}}_0/\mathrm{i}_0 = E_\theta\dot{l}_0(x)/i_0$, and then (2.1) gives the curvature squared of $(\theta, \phi(\theta))$ equal to $\mu_{11}^2/8i_0^2$ at $\theta = 0$. This completes the proof of (10.1). We see that the term $\Delta_{\theta_0}^{\theta}$ is

$$\Delta_0^{\theta} = i_0\,\mathrm{tr}\,\mathbf{A}_{(1)}^2/2 \tag{10.24}$$

and so equals 0 for the MLE.

Several more remarks can now be made about (10.1).

9) The bias of the MLE up to $O(1/n)$ is, by (10.21), equal to $-\mu_{11}/(2i_0 n)$. If $\hat{\theta}$ is unbiased to $O(1/n)$, as it is for example in any translation parameter estimation problem involving a symmetric density, then we must have $\mu_{11} = 0$. By

(10.23) we then have $\mathrm{Var}_\theta \hat{\theta} = 1/ni_\theta + \gamma_\theta^2/n^2 i_\theta + o(1/n^2)$. The naming curvature term disappears from (10.1) in this case, so θ must be equivalent to the best name, ϕ, at every point in $\mathscr{F}$.

10) The expression (10.24) for the excess variance of $\tilde{\theta}$ over the MLE also occurs in the theory of Section 9,

$$\lim_{n\to\infty} \mathbf{i}_\theta - \mathbf{i}_\theta^{\tilde{\theta}} = i_\theta \gamma_\theta^2 + \Delta_\theta^{\tilde{\theta}}, \tag{10.25}$$

see Rao (1963).

11) Let $\mathbf{A}(\theta_0)$ be the matrix $\mathbf{A}$ in the Taylor expansion (10.3) when we have put $\mathscr{F}$ into standard form at $\theta = \theta_0$, and define $B_\theta^{\tilde{\theta}} \equiv \mathrm{tr}\, \mathbf{A}(\theta)/2$. Then up to $O(1/n)$, $B_{\theta_0}^{\tilde{\theta}}/n$ is the bias of $\tilde{\theta}$ when $\theta = \theta_0$. It is easy to show, by calculations similar to those in Remark 12 below, that $\tilde{\theta}_n \equiv \tilde{\theta} - B_{\tilde{\theta}(\bar{x})}^{\tilde{\theta}}/n$ has bias of order $O(1/n^2)$ and variance as given in (10.23) but with the term $2b_\theta/ni_\theta$ removed. See Rao (1963). For the MLE $\hat{\theta}$, $B_\theta^{\hat{\theta}} = -(\mu_{111}(\theta)/2i_\theta^2)$. The estimator $\hat{\theta} - B_{\hat{\theta}(\bar{x})}^{\hat{\theta}}/n + B_{\hat{\theta}(\bar{x})}^{\tilde{\theta}}/n$ has variance as given in (10.1) but with the term $\Delta_{\theta_0}^{\tilde{\theta}}$ removed. The point is that by modifying the MLE we can obtain an estimator with the same bias structure and smaller variance than any other consistent, efficient estimator $\tilde{\theta}$.

12) We have ignored the $O(\bar{x}^3)$ term in (10.3) in the derivation of (10.23) and (10.1). To justify this requires the following result: let $\mathscr{C}_n$ be the cube $\{z : |z_i| \leqq n^\alpha,\ i = 1, 2, \cdots, r\}$, $0 < \alpha < \frac{1}{6}$, and $I_n(z)$ the indicator function of $\mathscr{C}_n$. Define $z_n \equiv n^{\frac{1}{2}}\bar{x}$ (so $z_n \to_{\mathscr{L}} \mathscr{N}_r(\mathbf{0}, \mathbf{I})$) and let $p(z_n)$ be a polynomial of degree l in the coordinates of z_n. Then

$$E_\theta p(z_n)[1 - I_n(z_n)] = O(n^{l\alpha} \exp\{-\tfrac{1}{2}n^{2\alpha}\}) \tag{10.26}$$

as discussed in the Appendix.

Now write (10.5) as $\tilde{\theta} - \bar{x}_1/i_\theta^{\frac{1}{2}} = Q + R$ where Q is the quadratic term $\bar{x}'\mathbf{A}\bar{x}$, $\mathbf{A}$ having the special form indicated in the lemma, and R is the remainder term $O(\bar{x}^3)$. Also define $S(\bar{x}) \equiv Q(\bar{x})I_n(n^{\frac{1}{2}}\bar{x})$, $T(\bar{x}) \equiv Q(\bar{x})[1 - I_n(n^{\frac{1}{2}}\bar{x})]$, and $V = T + R$ (so $Q = S + T$, $\tilde{\theta} - \bar{x}_1/i_\theta^{\frac{1}{2}} = S + V$). Notice that

$$|V| = |O(\bar{x}^3)| < Kn^{-3(\frac{1}{2}-\alpha)} \qquad \text{for} \quad n^{\frac{1}{2}}\bar{x} \in \mathscr{C}_n \tag{10.27}$$

for some positive constant K. (We use below the same symbol K to represent any bounding constant.) To the assumptions of the lemma we now add *that* $|\tilde{\theta} - \bar{x}_1/i_\theta^{\frac{1}{2}}|$ *is uniformly bounded*, giving

$$|V| < K, \qquad n^{\frac{1}{2}}\bar{x} \notin \mathscr{C}_n. \tag{10.28}$$

(With only slightly greater effort below, the boundedness condition can be relaxed to $|\tilde{\theta}| \leqq K(n^{\frac{1}{2}}\|\bar{x}\|)^k$ for $n^{\frac{1}{2}}\bar{x} \notin \mathscr{C}_n$ for some positive constants K, k.) By (10.26) and (10.27),

$$E_\theta|V|^l = O(n^{-3l(\frac{1}{2}-\alpha)}) \tag{10.29}$$

for any $l \geqq 0$, while

$$E_\theta|T|^l = O(n^{3\alpha l}e^{-n^{2\alpha}/2}). \tag{10.30}$$

Formulas (10.21) and (10.23) were derived assuming $\tilde{\theta} - \bar{x}_1/i_0^{\frac{1}{2}} = Q$. But

$$|E_0 Q - E_0 S| \leqq E_0|T| = O(n^{\alpha} e^{-n^{2\alpha}/2})$$

and

$$|E_0(\tilde{\theta} - \bar{x}_1/i_0^{\frac{1}{2}}) - E_0 S| \leq E_0|V| = O(n^{-3/4-\alpha}) .$$

Since $\alpha < \frac{1}{6}$ this shows that $E_0\tilde{\theta} = E_0 Q + o(1/n)$, so (10.21) is valid. Likewise

$$\begin{aligned} |E_0(\tilde{\theta} - \bar{x}_1/i_0^{\frac{1}{2}})^2 - E_0 Q^2| &= |E_0(S+V)^2 - E_0(S+T)^2| \\ &= |E_0[2SV + V^2 - T^2]| \end{aligned}$$

(since $ST \equiv 0$), which is $\leqq 2E_0|SV| + E_0|V|^2 + E_0|T|^2$. The last two terms are $o(n^{-2})$ by (10.29) and (10.30). Notice that $SV = O(\bar{x}^6)$ and $SV = 0$ for $n^{\frac{1}{2}}\bar{x} \notin \mathscr{C}_n$, so

$$|SV| < Kn^{-3/4-\alpha} . \tag{10.31}$$

Taking $\alpha < \frac{1}{16}$ makes $E_0|SV| = o(n^{-2})$, completing the proof that (10.28) is valid. We remark that a more careful proof, assuming $\tilde{\theta}$ four times continuously differentiable, allows one to replace $o(1/n^2)$ bs $O(1/n^5)$ in (10.1).

Acknowledgment. Much of this work was done while I was visiting Imperial College, London, Department of Mathematics. I appreciate the assistance of Margaret Ansell in carrying out the more difficult numerical computations. The Associate Editor provided extensive help, especially with the Appendix.

APPENDIX

Complete proofs of the statements made in Sections 9 and 10 require large deviation results of the type discussed in Chernoff (1952) and the references therein. Suppose $x_1, x_2, \cdots, x_n, \cdots$ are independent, identically distributed real valued random variables such that $Ex_i = 0$, Var $x_i = 1$, and $\psi(s) \equiv Ee^{sx}$ exists for $|s| < s_0$, s_0 some positive constant. Then $\psi(s) = 1 + s^2/2 + O(s^3)$ for s near 0, so

$$\log \psi(s) = s^2/2 + O(s^3) . \tag{A1}$$

Define $I_{[y,\infty)}(z) = 1$ or 0 for $z \geqq y$ or $z < y$, respectively. Because $e^{ns(\bar{x}_n - y)} \geqq I_{[y,\infty)}(\bar{x}_n)$ for all values of $\bar{x}_n \equiv \sum_{i=1}^n x_i/n$ we have, for $|s| < s_0$,

$$P\{\bar{x}_n \geqq y\} \leqq Ee^{ns(\bar{x}_n - y)} = [\psi(s)e^{-sy}]^n . \tag{A2}$$

LEMMA. *For c_n a sequence of numbers going to infinity, $c_n = o(n^{\frac{1}{2}})$, and l a non-negative integer,*

$$E\{(n^{\frac{1}{2}}\bar{x}_n)^l I_{[c_n,\infty)}(n^{\frac{1}{2}}\bar{x}_n)\} \leqq c_n{}^l e^{-c_n{}^2/2 + o_n(1)} . \tag{A3}$$

PROOF. Let $F_n(y) \equiv P\{\bar{x}_n \geqq y\}$, so $\bar{F}_n(y) \leqq [\psi(s)e^{-sy}]^n$ for $|s| < s_0$ by (A2). We have

$$E\{(n^{\frac{1}{2}}\bar{x}_n)^l I_{[c_n,\infty)}(n^{\frac{1}{2}}\bar{x}_n)\} = -n^{l/2}\int_{c_n/n^{\frac{1}{2}}}^{\infty} x^l \, d\bar{F}_n(x)$$

and integration by parts gives

$$-\int_{c_n/n^{\frac{1}{2}}}^{\infty} x^l\, d\bar{F}_n(x) = \left(\frac{c_n}{n^{\frac{1}{2}}}\right)^l \bar{F}_n\left(\frac{c_n}{n^{\frac{1}{2}}}\right) + l\int_{c_n/n^{\frac{1}{2}}}^{\infty} x^{l-1}\bar{F}_n(x)\,dx$$

$$\leqq \left(\frac{c_n}{n^{\frac{1}{2}}}\right)^l \left[\psi\left(\frac{c_n}{n^{\frac{1}{2}}}\right)e^{-sc_n/n^{\frac{1}{2}}}\right]^n + l\int_{c_n/n^{\frac{1}{2}}}^{\infty} x^{l-1}[\psi(s)e^{-sx}]^n\,dx\,.$$

Taking $s = c_n/n^{\frac{1}{2}}$ gives

$$\text{(A4)}\qquad E\{n^{\frac{1}{2}}\bar{x}_n)^l I_{[c_n,\infty)}(n^{\frac{1}{2}}\bar{x}_n)\} \leqq \psi^n\left(\frac{c_n}{n^{\frac{1}{2}}}\right)e^{-c_n^2}\left[c_n^l + \frac{l}{c_n}E\left(c_n + \frac{G}{c_n}\right)^{l-1}\right],$$

where G has density e^{-g} for $g \geqq 0$, 0 otherwise. Finally

$$\text{(A5)}\qquad \psi^n\left(\frac{c_n}{n^{\frac{1}{2}}}\right) = e^{n\log\psi(c_n/n^{\frac{1}{2}})} = e^{c_n^3/2 + O(c_n^3/n^{\frac{1}{2}})}$$

by (A1). Combining (A4) and (A5) gives (A3) with

$$\text{(A6)}\qquad o_n(1) = O([c_n/n^{\frac{1}{2}}]^3) + \log\{1 + lc_n^{-l}E(1 + G/c_n^3)^{l-1}\}\,,$$

where we now use $c_n = o(n^{\frac{1}{2}})$, $c_n \to \infty$.

Now let $x_1, x_2, \cdots, x_n, \cdots$ be independent identically distributed random vectors, dimension k, $Ex_i = \mathbf{0}$, $\operatorname{Cov} x_i = \mathbf{I}$, such that $\psi(t) = Ee^{t'x_i}$ exists for $\|t\| < t_0$, some positive constant. For any unit vector v define $x_i^v \equiv v'x_i$. Then (A3) holds with $\bar{x}_n$ replaced by $\bar{x}_n^v$. The term $o_n(1)$ is defined as in (A6), with the big O term being the one in the expression $\log\psi(t) = \|t\|^2/2 + O(t^3)$. (Notice that $o_n(1)$ does not depend on v.) (10.26) now follows easily.

LEMMA. *If $|E_0 e^{t'x_i}|^\rho$ is integrable as a function of t for some $\rho \geqq 1$ then $g_n(z)$, the density of $z = n^{\frac{1}{2}}\bar{x}_n$, exists and satisfies*

$$\text{(A7)}\qquad g_n(z) < \frac{2^{\frac{1}{2}}}{(2\pi)^{k/2}}\, e^{-(\|z\|/4)\min\{c_n,\|z\|+o_n(1)\}}\,,$$

$c_n = o(n^{\frac{1}{2}})$, $c_n \to \infty$.

PROOF. Consider the univariate case, with n even. Define

$$\text{(A8)}\qquad h(z) \equiv \int_{-\infty}^{\infty} g_{n/2}(w)g_{n/2}(z-w)\,dw$$

$$= \int_{-\infty}^{z/2} g_{n/2}(w)g_{n/2}(z-w)\,dw + \int_{z/2}^{\infty} g_{n/2}(w)g_{n/2}(z-w)\,dw\,.$$

Here $g_{n/2}(z)$, the density of $(n/2)^{\frac{1}{2}}\bar{x}_{n/2}$, is known to exist and to converge uniformly to $(2\pi)^{-\frac{1}{2}}\exp(-z^2/2)$, see page 244 of Gnedenko and Kolmogorov (1954). Then $M_n = \sup_z |g_n(z)| = (2\pi)^{-\frac{1}{2}} + o_n(1)$, so for $0 \leqq z \leqq c_n$

$$h(z) \leqq M_{n/2}\left\{\int_{-\infty}^{z/2} g_{n/2}(z-w)\,dw + \int_{z/2}^{\infty} g_{n/2}(w)\,dw\right\}$$

$$\leqq 2M_{n/2}e^{-(z/8)\min\{2c_n,z\}+o_n(1)}$$

where we have used the bound $P\{n^{\frac{1}{2}}\bar{x}_n \geqq z\} \leqq \exp[-z/2\min\{c_n, z\} + o_n(1)]$ obtained by setting $y = z/n^{\frac{1}{2}}$ and $s = \min\{z_n/n^{\frac{1}{2}}, c_n/n^{\frac{1}{2}}\}$ in (A2). But $g_n(z) = 2^{\frac{1}{2}}h(2^{\frac{1}{2}}z)$, giving (A7). The same proof with trivial modifications works for n odd. For

the multivariate case the integrals in (A8) are over the regions $R_1 = \{w : z'w < \|z\|^2/2\}$ and $R_2 = \{w : z'w > \|z\|^2/2\}$.

Remark 1 of Section 9 follows because (A7) makes step (v) of the heuristic proof valid. All the other approximations involved in the proof are handled by power series expansions and the bounding arguments of Remark 12, Section 10.

REFERENCES

[1] ANDREWS, F., BICKEL, P., HAMPEL, P., HUBER, P., ROGERS, W., and TUKEY, J. (1972). *Robust Estimates of Location*. Princeton Univ. Press.

[2] BARNETT, V. D. (1966). Evaluation of the maximum likelihood estimator when the likelihood equation has multiple roots. *Biometrika* **53** 151–165.

[3] CHERNOFF, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23** 493–507.

[4] DAVIES, R. B. (1969). Beta optimal tests and an application to the summary evaluation of experiments. *J. Roy. Statist. Soc. Ser. B* **31** 524–538.

[5] DAVIES, R. B. (1971). Rank tests for Lehmann's alternative. *J. Amer. Statist. Assoc.* **66** 879–883.

[6] EFRON, B. and TRUAX, D. (1968). Large deviations theory in exponential families. *Ann. Math. Statist.* **39** 1402–1424.

[7] FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **122** 700–725.

[8] GNEDENKO, B. V., and KOLMOGOROV, A. N. (1954). (Translated by K. Chung.) *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, Cambridge, Mass.

[9] HALDANE, J. B. S. and SMITH, S. M. (1956). The sampling distribution of a maximum likelihood estimate. *Biometrika* **43** 96–103.

[10] RAO, C. R. (1961). Asymptotic efficiency and limiting information. (J. Neyman, Ed.). *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 531–545. Univ. of California Press.

[11] RAO, C. R. (1962). Efficient estimates and optimum inference procedures in large samples. *J. Roy. Statist. Soc. Ser. B.* **24** 46–72.

[12] RAO, C. R. (1963). Criteria of estimation in large samples. *Sankhyā* **25** 189–206.

[13] STRUIK, D. J. (1950). *Differential Geometry*. Addison-Wesley, Reading, Mass.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

DISCUSSION ON PROFESSOR EFRON'S PAPER

Professor Efron's paper was presented at the 1974 Annual Meeting of the Institute of Mathematical Statistics at Edmonton, Alberta. Professors D. R. Cox, A. P. Dawid, J. K. Ghosh, N. Keiding, L. M. Le Cam, D. V. Lindley, J. Pfanzagl, D. A. Pierce, C. R. Rao and J. Reeds were invited discussants. The Editor greatly appreciates the willing assistance of Professor Efron as well as the discussants in arranging this discussion paper. Professor Rao's remarks arrived after the author's reply to the discussion was received and are not referred to for that reason.

C. R. RAO

Indian Statistical Institute, New Delhi

I am delighted to see the paper by Bradley Efron and also the paper by J. K. Ghosh and K. Subrahmaniam (*Sankhyā A*, 1975 **36** 325–358) on the subject of second order efficiency. Having worked for some time on second order efficiency of estimators, I was aware of the importance of measures of how closely a given model can be approximated by an exponential family $\{f_\theta = C(\theta)\exp[K(\theta)T(X)]\}$. Measures of this sort are of course closely related to what Professor Efron calls the curvature of a statistical problem. What is quite new about Professor Efron's measure is its invariance under smooth 1 — 1 transformations and the elegant geometric interpretation which makes the term so apt and illuminating and provides new tools and insights into the subject.

My endeavour in this area was motivated by two results in the literature on estimation which seemed to contradict Fisher's claims about MLE's. (maximum likelihood estimators). One is the concept of super efficiency, according to which MLE is not efficient in the sense defined by Fisher. Another is the concept of BANE (best asymptotically normal estimator), according to which ML is only one out of a very wide class of estimation procedures.

The first task was to redefine the concept of efficiency of an estimator since its asymptotic variance is a poor indicator of its performance in statistical inference. To do this it is necessary to see how well an optimum inference procedure based on a given estimator T_n alone compares with that based on all the observations. Following Fisher's ideas, I thought it is relevant, at least in large samples, to consider the score function $\dot{l}(\theta)$ (see Efron's paper for notations) as basic to all inference problems. Then the problem reduces to examining how closely $\dot{l}(\theta)$ and T_n are related. Under the additional condition that T_n is consistent for θ, T_n was defined to be *first order efficient* if

$$\text{(1)} \qquad \text{plim}_{n\to\infty} |n^{-\frac{1}{2}}\dot{l}(\theta) - \alpha - \beta n^{\frac{1}{2}}(T_n - \theta)| \to 0 .$$

There are a large number of estimators which are first order efficient. To distinguish among them, it is natural to examine the rapidity of convergence in (1), which led to the consideration of the random variable (rv)

$$\text{(2)} \qquad |\dot{l}(\theta) - n^{\frac{1}{2}}\alpha - n\beta(T_n - \theta)|$$

which is $n^{\frac{1}{2}}$ times the rv in (1). The asymptotic variance of (2) was defined as the *second order efficiency*. Instead of (2) we may as well consider the rv

$$\text{(3)} \qquad |\dot{l}(\theta) - n^{\frac{1}{2}}\alpha - n\beta(T_n - \theta) - \lambda n(T_n - \theta)^2| .$$

and define its minimum asymptotic variance for a proper choice of λ as the second order efficiency. Fisher suggested the use of

$$\text{(4)} \qquad \lim_{n\to\infty} n(i - i_{T_n})$$

to distinguish between alternate estimators, but the computation of (4) is extremely difficult.

The definition arising out of (3) was criticised as not being directly related to an inference problem, although it attempts to examine how close T_n is to $\dot{l}(\theta)$. This led to another definition of second order efficiency based on the expansion (under some conditions) of the variance of T_n after correcting for bias

$$V(T_n) = \frac{1}{in} + \frac{\phi(\theta)}{n^2} + o\left(\frac{1}{n^2}\right). \tag{5}$$

The quantity $\phi(\theta)$ was considered as a measure of second order efficiency. A major component of $\phi(\theta)$ was the measure based on (3).

With this background, the work of Efron is valuable in many ways.

(i) The results due to Fisher and me were confined to multinomial distributions. Efron, and also Ghosh and Subrahmaniam extend the results to a wider class of distributions.

(ii) Efron relates second order efficiency to what he calls curvature of a statistical problem, which appears to be natural and throws further light on problems of inference (providing, for instance, an intimate connection between curvature and properties of test criteria).

(iii) Efron provides a decomposition of $\phi(\theta)$ in (5), which is extremely interesting.

(iv) Efron suggests the use of a most powerful test at a suitably chosen alternative in preference to a locally most powerful test, which seems to be an attractive idea worth pursuing.

No doubt Efron's work has led to considerable clarification of second order efficiency and its relevance in problems of inference. However, there are many problems which require deeper investigation.

(i) Efron shows by an example that measures of second order efficiency based on (3) and (4) can be different. In fact, as he observes, it may be shown (from definition) that the measure based on (4) is smaller than that on (3). But the question remains: under what conditions are the two measures the same, and is the MLE efficient under the measure (4)?

(ii) I have considered Fisher's score function $l(\theta)$ as a basic in problems of inference. Perhaps, following Barnard and Sprott, one should consider $1(\theta)$ itself. How should efficiency of T_n be defined in such a case?

(iii) How can the result based on quadratic loss function as in (5) be extended to more general loss functions?

DON A. PIERCE

Oregon State University

I think that I am not alone in having had great difficulty with the reasoning of Fisher's 1925 paper. Professor Efron's elegant contribution to clarifying these ideas is very helpful.

The part of Fisher's paper which has intrigued and puzzled me most is the final section in which he suggests the use of $\ddot{\mathbf{l}}_{\hat\theta}(\mathbf{x})$, in Efron's notation, as an ancillary statistic. I would like to indicate here how the geometry of this paper helps clarify this, although there are many details yet unclear to me.

It is characteristic of "curvature" that $-\ddot{\mathbf{l}}_{\hat\theta}(\mathbf{x}) \neq \mathbf{i}_{\hat\theta}$. In fact, one can always parameterize so that $\mathrm{Cov}_{\theta_0}(\dot{l}_{\theta_0}, \ddot{l}_{\theta_0}) = 0$, and then $\gamma^2_{\theta_0} = \mathrm{Var}_{\theta_0}(\ddot{l}_{\theta_0})/i^2_{\theta_0}$. Fisher seems to suggest using $-\ddot{\mathbf{l}}_{\hat\theta}(\mathbf{x})$, rather than $\mathbf{i}_{\hat\theta}$, as a post-data measure of precision of $\hat\theta$. This is also suggested by standard asymptotic Bayesian arguments, but the sampling theory justification has never been clear to me. Such use of $\ddot{\mathbf{l}}_{\hat\theta}$ would be significant relative to the order of n^{-1} of approximation to $\mathrm{Var}\,(\hat\theta)$ considered in this paper, for $-\ddot{\mathbf{l}}_{\hat\theta} = \mathbf{i}_{\hat\theta} + O_p(n^{\frac{1}{2}})$ and thus $-1/\ddot{\mathbf{l}}_{\hat\theta} = 1/\mathbf{i}_{\hat\theta} + O_p(n^{-\frac{1}{2}})$.

The geometrical structure exposed in this paper is indeed very helpful in understanding the role of $\ddot{\mathbf{l}}_{\hat\theta}$ as an ancillary statistic. For a curved exponential family of dimension k think of the projection from the sample point $\mathbf{x} \in E^n$ to the MLE $\lambda_{\hat\theta}$, where $\lambda_\theta = E(\mathbf{x})$ as an orthogonal projection (relative to $\sum_{\hat\theta}^{-1}$) first to $\hat\lambda$ in the local osculating plane of the curve λ_θ and then a projection from $\hat\lambda$ to $\lambda_{\hat\theta}$. The argument below suggests that $(-\ddot{\mathbf{l}}_{\hat\theta}(\mathbf{x}) - \mathbf{i}_{\hat\theta})/\mathbf{i}_{\hat\theta}$ is a useful measure of the signed distance from $\hat\lambda$ to the curve λ_θ, positive when $\hat\lambda$ is on the outside of the curve. This is useful ancillary information because the projection from $\hat\lambda$ to $\lambda_{\hat\theta}$ is a contraction (resp. expansion) mapping when $\hat\lambda$ is on the outside (resp. inside) of the curve λ_θ. The extent of this contraction is a function of the distance from $\hat\lambda$ to the curve λ_θ, as measured by the above statistic. Thus the conditional precision of $\lambda_{\hat\theta}$ given $\ddot{\mathbf{l}}_{\hat\theta}(\mathbf{x})$ is either greater or less than the unconditional precision. Furthermore, it appears plausible that the component of $\hat\lambda$ orthogonal to the curve λ_θ at $\lambda_{\hat\theta}$ is itself uninformative regarding the value of λ_θ.

More precisely, consider the situation of Figure 4 with the additional assumption that θ is a choice of parameter such that $\mathrm{Cov}_\theta(\dot{l}_\theta, \ddot{l}_\theta) = 0$. The point (x_1, x_2) corresponds to the λ of the above discussion. It follows directly from (6.3) and the relations given in the second paragraph after (9.2) that

$$\bar{x}_1 = \dot{\mathbf{l}}_0/n(i_0)^{\frac{1}{2}}, \qquad \bar{x}_2 = -[-\ddot{\mathbf{l}}_0 - ni_0]/ni_0\gamma_0 .$$

Near the origin the curve λ_θ is approximately a segment of a circle with center at e_2/γ_0, and the arc distance of $\lambda_{\hat\theta}$ from the origin is to first order $i_0^{\frac{1}{2}}\hat\theta$. Proportionality of arc lengths to radii gives

$$\begin{aligned} i_0^{\frac{1}{2}}\hat\theta/\bar{x}_1 &\doteq (1/\gamma_0)/(1/\gamma_0 - \bar{x}_2) \\ &= (1 - \gamma_0\bar{x}_2)^{-1}, \end{aligned}$$

so

$$\begin{aligned} \hat\theta &\doteq (\bar{x}_1/i_0^{\frac{1}{2}})(1 - \gamma_0\bar{x}_2)^{-1} \\ &= (\bar{x}_1/i_0^{\frac{1}{2}})[1 + (-\ddot{\mathbf{l}}_0 - ni_0)/ni_0]^{-1} \\ &= (\bar{x}_1/i_0^{\frac{1}{2}})[ni_0/(-\ddot{\mathbf{l}}_0)] . \end{aligned} \tag{1}$$

Equation (1) can be seen to agree with the rigorously established (10.5) of the paper, where $\mu_{11} = 0$ since $\nu_{11} = 0$.

Thus we have

$$\text{(2)} \qquad \begin{aligned} \operatorname{Var}(\hat{\theta} \mid \ddot{\imath}_0) &\doteq (1/ni_0)[ni_0/(-\ddot{\imath}_0)]^2 \\ &= [ni_0/(-\ddot{\imath}_0)][1/(-\ddot{\imath}_0)] . \end{aligned}$$

Since $-\ddot{\imath}_0 = ni_0 + O_p(n^{\frac{1}{2}})$ this expression can be *either greater or less* than $1/ni_0$ by an amount $O_p(n^{-\frac{3}{2}})$.

I do not know the effect of conditioning on $\ddot{\imath}_{\hat{\theta}}$ rather than $\ddot{\imath}_0$ nor can I see yet whether $1/(-\ddot{\imath}_{\hat{\theta}})$ as suggested by Fisher is a good approximation to $\operatorname{Var}(\hat{\theta} \mid \ddot{\imath}_{\hat{\theta}})$. Note that the expression in (2) differs by $O_p(n^{-\frac{3}{2}})$ from $1/(-\ddot{\imath}_0)$. I also do not know the effect of relaxing the assumption that one has parameterized so that $\operatorname{Cov}_0(\dot{l}, \ddot{l}_0) = 0$.

It appears, then, that the curvature γ_θ is essentially the standard deviation of an approximately ancillary statistic. This interpretation might have a number of advantages over that furnished by relations such as (1.1) and (10.1). Loosely put, the degree of curvature relates to the amount of information in the sample which is not captured by the MLE; information in a sense regarding not θ but rather the precision of $\hat{\theta}$. Moreover, this information can be largely recovered through appropriate use of $\ddot{\imath}_{\hat{\theta}}$.

D. R. Cox

Imperial College, London

Dr. Efron's impressive paper throws much light on a longstanding problem. I will confine my comments to one aspect that he has not treated. For an approach to statistical inference in which evidence in unique sets of data is interpreted via frequencies in hypothetical repetitions, appropriate conditioning is important, at least theoretically, in making the hypothetical repetitions relevant to the data under study. Thus for the translation family, Example 4, Fisher (1934) provided a simple definitive solution to inference about θ by conditioning on the ancillary statistic, the set of differences among order statistics. This leads to the use of normalized likelihood as giving confidence limits. Curvature here measures the variation among the different kinds of likelihood functions that can arise. It would be useful to make this more specific and to draw any implications about the comparison of conditional and unconditional inference.

More importantly, what are the implications of conditional inference for some of the other problems, for instance Example 1? Here, if $x = (x_1, x_2)$, $x_2 - \frac{1}{2}\gamma_0(x_1^2 - 1)$ is approximately ancillary in some sense, at least for small $\gamma_0\theta$. Existence of an approximate ancillary must be connected with the approximate constancy of γ_θ as a function of θ; it would be good to have the connexions explored.

REFERENCES

FISHER, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. Ser. A.* **144** 285–307.

D. V. LINDLEY

The University of Iowa

My first comment is to repeat the point made in discussing C. R. Rao's (1962) paper, namely that it is doubtful whether any general measure of second-order efficiency is possible. The reason for suggesting this is that an admissible estimate is typically, to order n^{-1}, equivalent to the maximum likelihood estimate, for a wide class of loss functions: but to order n^{-2} its asymptotic form depends on some features of the loss structure. Consequently the second-order "correction" to the maximum likelihood estimate typically depends on the loss structure, as does its efficiency. The point is discussed more fully in Lindley (1961).

Efron's thought-provoking paper does not introduce curvature solely for second-order efficiency properties; nevertheless the definition of curvature he proposes suffers from a defect in some statistical problems. The defect arises from the fact that it involves an integration over sample space and thereby violates the likelihood principle. Put it this way: suppose we have some data x and its associated likelihood function, $l_\theta(x)$, then, according to Efron, we have to consider what other data we might have had, but did not, before any inference can be made. These data are needed before the integrations, symbolized by E_θ in the paper, can be performed. That such data are needed is puzzling and any reasonable axiomatization of inference seems to deny their relevance. The author tacitly assumes that the other data are samples of the same size, but many practical problems do not naturally fit into this framework. Even the notation helps to reinforce this view. Likelihood is a function of θ for fixed x and yet Efron lowers the status of the variable to that of a subscript and the constant appears in the place customarily reserved for the argument. The notation $l(\theta \mid x)$ is surely to be preferred.

An example of the misuse of the integration is provided by the discussion of the t-translation family [Example 4 of Section 7: see also the remark after (8.3)]. If samples are taken from a t-distribution with low degrees of freedom, then it will be found that a substantial majority of them look very like samples from a normal distribution—the comparison being made through the t- and normal likelihoods. It is only rarely (how rarely depends on f and n) that a sample arises which is clearly nonnormal and its log-likelihood is markedly not quadratic. But because of the integration, or averaging, over all samples, these "peculiar" samples get put in with the "normal" ones and nonstandard estimates proposed. Looked at without prejudice, I think you will find this is a surprising thing to do. The argument can be extended to query whether it is reasonable to look for a point estimate in the "peculiar" cases: for example, when the likelihood is bimodal. I would go further and suggest that point estimation is not a good model for *any* inference procedure, though it does occasionally occur in a decision context. Estimation is solved by describing the likelihood function or the posterior distribution.

These criticisms have less force *before* the data, x, are to hand. If it is a question of experimental design, or choice of a survey sample, then naturally one has to consider what data *might* be obtained, and integration becomes natural and necessary. Hence curvature could have a place in these fields and it would be interesting to see whether, in some sense, linear designs were better than "curved" ones. However, the argument of my first paragraph would show that if a terminal (as distinct from design) decision problem is contemplated after the experimentation, then the choice of design would again involve a loss function, so that no general measure seems possible. Some experiments are not associated with terminal decisions and are genuinely inferential in character. In these one is collecting information about parameters and Shannon's measure is essentially the only one to use. I have tried to see whether some second-order expansion of it might lead to anything analogous to Efron's curvature, but without success.

REFERENCES

LINDLEY, D. V. (1961). The use of prior probability distributions in statistical inference and decisions. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 453-468.

LUCIEN LE CAM

University of California, Berkeley

Professor Bradley Efron is to be congratulated for a clear and informative discussion of the differential properties of families of measures. The paper is certainly a step in the right direction. However, as I shall try to explain below, much remains to be done.

The paper tends to give the impression that the curvature measures the loss of information sustained by using a one dimensional summary of the data. This is perhaps so if "information" is measured by Fisher's number. However, one can define other measures of loss of information more directly in terms of performance in testing or other decision problems. See for instance E. N. Torgersen (1970). These definitions are usable for arbitrary families, whether or not they are smoothly differentiable.

It can probably be shown that these other measures of loss of information are related to Fisher's numbers in certain special situations, but not in general. One could roughly say that Torgersen's formula for testing deficiencies relies on finite differences instead of relying on the first and second derivatives used to compute curvatures. Efron's curvature has the merit of being easily computable, but one should not take it for granted that computations with differences, which may be difficult, should not be attempted.

The part of the paper which relates to the presumed excellency of maximum likelihood estimates should be taken with a great deal of caution. It is easy to modify Bahadur's example (1958) to construct one parameter families of densities which are infinitely differentiable, satisfy all kinds of reasonable conditions locally but are such that, when the number of observations tends to infinity,

the maximum likelihood estimate always converges to infinity, no matter what the true value of θ is.

It is also easy to find exponential families where, for reasonable numbers of observations, maximum likelihood estimates are difficult to compute and definitely worse (in the sense of expected square deviations) than some readily available alternatives. An example occurs in bioassay using the logit method (see Berkson (1951)). Another example with an interesting discussion is given by T. S. Ferguson (1958).

Finally, it seems that the entire asymptotic argument relies essentially on a replacement of the actual logarithm of likelihood ratio by a suitable approximation which is quadratic in θ.

If this is indeed the case, the technique of using a preliminary estimate, fitting a quadratic around the estimated value and then maximizing the quadratic should give the same asymptotic results. Preliminary considerations suggest that this technique may well work better than straight maximum likelihood estimation in the finite sample situation.

REFERENCES

[1] BAHADUR, R. R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhyā* **20** 207–210.

[2] BERKSON, J. (1951). Relative precision of minimum chi-square and maximum likelihood estimates of regression coefficients. *Proc. Second Berkeley Symp. Math. Statist. Prob.* 471–479. Univ. of California Press.

[3] FERGUSON, T. S. (1958). A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities. *Ann. Math. Statist.* **29** 1046–1062.

[4] TORGERSEN, E. N. (1970). Comparison of experiments when the parameter space is finite. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **16** 219–249.

J. K. GHOSH

Indian Statistical Institute, Calcutta

Thanks to my work on second order efficiency, I was aware of the significance of the quantity which Professor Efron calls the curvature of a statistical problem. What enhances the importance of it is the elegant geometric interpretation of it, which affords new techniques and deeper insight into the problem.

It is natural to expect that this quantity also plays an important role in asymptotic problems of testing hypotheses. By considering a number of examples of curved exponential families, Professor Efron has shown that this is indeed the case and unless curvature is small such commonly used methods as maximising the local power perform rather poorly for moderate sample sizes. Pfanzagl (1974) has arrived at the same conclusion. (Pfanzagl's D = (curvature)^2/4.)

Probably even more interesting than this is the suggestion by both Pfanzagl and Efron to use a suitable most powerful test instead of a locally most powerful test when the curvature is appreciable. Following Davies, Professor Efron suggests the use of a most powerful test against an alternative θ_1 such that its

power at θ_1 is about .8 and recommends the thumb rule of taking $\theta_1 = \theta_0 + 2/I_{\theta_0}^{\frac{1}{2}}$. These suggestions must be tried out in lots of problems involving nonexponential families to see if one does get reasonable tests this way even for moderate samples. (Pfanzagl (1974) provides some criteria for comparing two tests.) I report below some calculations for a curved nonexponential family, namely, the Cauchy with unknown location parameter. To make matters worse, I take sample size $N = 1$.

Suppose then that I have a random variable X with density $f_\theta(x) = 1/\pi \cdot 1/(1 + (x - \theta)^2)$ and want to test $H_0(\theta = 0)$ vs. $H_1(\theta > 0)$. Let ϕ_0 be the most powerful test of Davies and ϕ_1 the test: reject H_0 iff $X > C$. The second test has its greatest power against $\theta = 2C$ and seems to me a reasonable one. For $\alpha = .05$, ϕ_0 is most powerful against $\theta = 5$ (approximately) and ϕ_1 is most powerful against $\theta = 13$ (approximately). The following table compares ϕ_0 and ϕ_1.

	$\theta = 5$	$\theta = 13$
ϕ_0	.8	.06
ϕ_1	.2	.95

If $\alpha = .2$, ϕ_0 and ϕ_1 are nearly the same and are most powerful against $\theta = 2(2)^{\frac{1}{2}}$ which is the alternative obtained by Efron's thumb rule. I refrain from drawing any conclusion.

It is not difficult to come up with analogues of curvature when one has more parameters than one. Extension of the results due to Rao and Fisher to multiparameter families is provided in Ghosh and Subramanyam (1974). But it is now necessary to study testing problems of composite hypotheses along the lines of investigation carried out by Efron and Pfanzagl for simple hypotheses.

How relevant is curvature for a Bayesian? Ghosh and Subramanyam (1974) have shown how one can construct a Bayesian proof of the second order efficiency of the MLE. What is lacking and would be useful to have is a study of relevance of curvature in Bayesian analysis. The difficulty here is that one cannot think of any simple and convincing reason why a Bayesian would prefer the linear exponential families to nonexponential ones. All is grist that comes to the mill of the lucky man who not only has a prior but knows what it is.

It is a little disappointing, though not really surprising in retrospect, that curvature has nothing to do with the geometrical curvature of the likelihood curves. Curvature is, however, useful in the problems that Sprott (1973) discusses. For it is easy to show that his two approaches of minimizing $F_A(\phi)$ or $F(\phi)$ (in his notations) coincide iff one has a linear exponential family. (This statement is true provided the MLE satisfies the likelihood equation with probability one for all θ.) For example (2.3) of Sprott (1973), the curvature is fairly large for x near .5 and so Sprott's transformation which minimizes $F_A(\phi)$ may not be efficient in normalising the likelihood for x near .5. Incidentally, I suspect that for small curvature one can reparametrize in such a way that the approach of a posterior to normality, guaranteed by the Bernstein–von Mises theorem, would

be faster with the new parameter than with the original. (This may be an answer to the question of relevance of curvature for a Bayesian.)

It may be worth pointing out here that the results of Pfanzagl (1973) and those of Fisher and Rao (i.e. results like (10.1) of Efron) are not really comparable. In fact for all the efficient estimators considered by Efron or Ghosh and Subramanyam (1974), inequality (6.4) of Pfanzagl (1973, page 1005) reduces to an equality. This result, which is not very hard to show, will appear in Ghosh and Srinivasan (1975).

Finally, a question suggested by the beautiful counter example of Professor Efron. Is there any example such that among the Fisher consistent efficient estimators the MLE does not minimize the loss in Fisher's information for all values of θ? It seems reasonable to expect that such examples do exist.

REFERENCES

[1] GHOSH, J. K. and SUBRAMANYAM, K. (1974). Second order efficiency of maximum likelihood estimators. *Sankhya Ser. A.* (To appear).

[2] GHOSH, J. K. and SRINIVASAN, C. (1975). Asymptotic sufficiency and second order efficiency. Unpublished.

[3] PFANZAGL, J. (1973). Asymptotic expansions related to minimum contrast estimators. *Ann. Statist.* **1** 993–1026.

[4] PFANZAGL, J. (1974). Nonexistence of tests with deficiency zero. University of Cologne, preprint in Statistics #8.

[5] SPROTT, D. A. (1973). Normal likelihoods and their relation to large sample theory of estimation. *Biometrika* **60** 457–465.

J. PFANZAGL

University of Cologne

In hypothesis testing, one-parameter exponential families are distinguished by the fact that for one-sided alternatives uniformly most powerful tests exist for arbitrary sample sizes. For other families, the test has to be chosen with particular alternatives in mind. It is intuitively clear that the dependence of the test on these particular alternatives will be weak if the family is close to an exponential one. Is it possible to measure "nonexponentiality" (for this and other purposes) by a single quantity? Mr. Efron's suggestion to use the "curvature" γ_θ for this purpose is based on a geometric analogy. Therefore, its usefulness for statistical theory is not obvious in advance. It is the purpose of this note to draw attention to some results of asymptotic theory where the function γ_θ has been in use already for some time. Whether curvature admits an easy statistical interpretation in nonasymptotic theory seems doubtful.

"Nonexponentiality" implies in particular that a LMP (locally most powerful) test will not be MP (most powerful) against the statistically reasonable alternatives. The author uses a particular example to support his claim (see end of Section 8) that $\gamma^2_{\theta_0}$ is a good predictor for the relative performance of the LMP test compared to the test which is MP against a specific alternative. In this

connection he suggests that the difference in power can be neglected if $\gamma_{\theta_0}^3 \leqq \frac{1}{8}$. For the case of a sample of n i.i.d. variables this entails that the difference in power can be neglected if the sample size exceeds $8\gamma_{\theta_0}^3$ (see 8.3).

Since this rule is rather arbitrary, the reader should be aware of other results which make the role of γ_{θ_0} more clear. These results concern the case of n i.i.d. variables, the distribution of which is nonatomic and sufficiently regular (as a function of θ). To define for a given level α-test "deficiency at rejection level β" we determine first the alternative closest to the hypothesis which can be rejected with probability β by some level α-test. (The test for which this is achieved is called β-optimal.) In order to reach rejection probability β for this alternative with the given test, the sample size has to be increased. The additional number of observations needed for this purpose is the "deficiency at rejection level β."

For the LMP α-test the deficiency at rejection level β is asymptotically equal to

$$\frac{1}{4}\gamma_{\theta_0}^2(N_\beta - N_\alpha)^2 + o(n^0), \tag{1}$$

where N_ε is the ε-quantile of the standard normal distribution. (See Chibisov (1973, Corollary 2) and Pfanzagl (1973, Section 8, formula 24) or Pfanzagl (1975, Proposition 1, formula 6.2).)

This result enables one to check whether the rule suggested by (8.3) is reasonable. For $\alpha = .01$ and $\beta = .99$ the deficiency is $5.4\gamma_{\theta_0}^2 + o(n^0)$. Mr. Efron suggests in (8.3) not to worry about curvature if $n \geqq 8\gamma_{\theta_0}^3$. To follow this suggestion and to use a LMP test instead of a β-optimal test could mean to waste more than half of the sample.

The following is another asymptotic result (for nonatomic families) illustrating the statistical relevance of curvature. If a sequence of tests is β_0-optimal, then its deficiency at rejection level β is at least

$$\frac{1}{4}\gamma_{\theta_0}^2(N_\beta - N_{\beta_0})^2 + o(n^0) \tag{2}$$

(see Pfanzagl 1975, Corollary 2, formula 6.5). Hence a sequence of tests having asymptotic deficiency zero for more than one alternative cannot exist unless the curvature is zero.

In another attempt to demonstrate the statistical relevance of "curvature," Mr. Efron refers to a result of Fisher (see (9.1)). Mr. Efron is careful enough not to follow Fisher's abuse of language using a suggestive word for a mathematical construct (such as "information" or "likelihood") without paying any attention to the question whether the interpretation thus suggested is meaningful from the operational point of view.

A statement like "Since a single observation contains an amount i_θ of information this [namely the use of a MLE instead of the whole sample] is equivalent to a reduction in effective sample size from n to $n - \gamma_\theta^2 \cdots$" (see beginning of Section 9) is misleading, at least, since for nonatomic families the level α-test based on the MLE has asymptotic deficiency zero at the rejection level $(1 - \alpha)$, and not asymptotic deficiency γ_θ^2, as the statement quoted above might suggest.

(See Chibisov 1973, Corollary 3 or Pfanzagl 1973, formula 23 for $t = -2N_n L^{-1}_{(01)(0)}$ or Pfanzagl 1975, end of Section 6.) Probably the statement quoted above is meant as the interpretation Fisher himself would give to (9.1). Since this interpretation is unjustified, how can (9.1) convince the reader that "curvature" is statistically significant?

REFERENCES

[1] Chibisov, D. M. (1973). Asymptotic expansions for some asymptotically optimal tests. *Proc. Prague Symp. Asymptotic Statist.* **2** 37–68.

[2] Pfanzagl, J. (1973). Asymptotically optimum estimation and test procedures. *Proc. Prague Symp. Asymptotic Statist.* **1** 201–272.

[3] Pfanzagl, J. (1975). On asymptotically complete classes. *Statistical Inference*. **2** 1 43 M. Puri, ed. Academic Press.

NIELS KEIDING

University of Copenhagen

1. An important feature of Efron's paper is the study of the loss of information resulting from summarizing the data in n replications $X_1, \cdots, X_n$ of a multivariate random variable into a one-dimensional statistic $T(\mathbf{X}) = T(X_1, \cdots, X_n)$. In most of the paper it is assumed that the X_i's are observable and that their distribution belongs to an exponential family of which the statistical model forms a "curved subset", in the sense of the mean value parametrization. The basic result in this connection is formula (9.3), stating that the information loss from n replications is

$$\mathbf{i}_\theta - \mathbf{i}_\theta^T = E_\theta \operatorname{Var}_\theta \{\dot{\mathbf{l}}_\theta(\mathbf{X}) \mid T\},$$

where for $T = \hat{\theta}$, the right hand side is $i_\theta \gamma_\theta^2$, independent of n. (Notice that it is an implicit consequence of this that $\hat{\theta}$ cannot itself have the form $\Sigma t(X_i)$).

A somewhat related problem is that of incomplete observation of an exponential family, where the statistician is "forced" to work with nonsufficient reduction of data. It is here assumed that the statistical problem is specified in terms of an exponential family where only a function $Y = Y(X)$ of each component may be observed. If Y is a linear function of the canonical statistic X, there seems to be a canonical way of decomposing the parameter vector into an efficiently estimable part and a nonidentifiable part, using the concepts of "mixed parametrization" and "cut" introduced and further studied by Barndorff–Nielsen (1973, 1974) and Barndorff–Nielsen and Blaesild (1975), and in the case of continuously distributed random variables this seems to hold as soon as the level curves of Y are hyperplanes. Asymptotic results for arbitrary "curved" functions Y were given by Sundberg (1974) who points out that the same formula as above applies for the information loss, which here in general will be of order n.

It is clear that the two situations might be combined: a "curved" model with incomplete observation. An example of this was discussed by Fisher (1958, Section 57.1).

2. The relation (10.1) for the asymptotic variance of any consistent and efficient estimator $\hat{\theta}$ contains the term $\Delta^{\gamma}_{\theta_0}$, being always nonnegative and zero for the MLE. This quantity was computed by Rao (1963) for several estimation methods in the multinomial distribution, as noted by Efron. It would be interesting if some geometrical interpretation, or at least a bit more transparent expression than (10.24) could be given for this quantity, which must be related to the intuitive discussion by Fisher (1958, Section 57) of "the contribution to χ^2 of errors of estimation".

3. Curved exponential families occur frequently in population process and life testing models leading to occurrence/exposure estimates of birth or death intensities. One familiar example is that of estimating the mean μ^{-1} of an exponential distribution from a sample of n, censored at a fixed point t. If D is the number of variables less than t, and S the sum of these $+ (n - D)t$, then the likelihood function is $\mu^D e^{-\mu S}$, yielding $\hat{\mu} = D/S$.

We shall here comment a little upon the similar example of estimating the birth intensity λ in a pure (linear) birth process (X_u) from continuous observation of the process in $[0, t]$. See Keiding (1974) for details of the problem.

Assuming $X_0 = x_0$, degenerate, the likelihood is

$$\lambda^{X_t - x_0} e^{-\lambda S_t}$$

with $S_t = \int_0^t X_u \, du$. Setting $B_t = X_t - x_0$, the maximum likelihood estimator is $\hat{\lambda} = B_t/S_t$. It is readily seen that the Fisher information

$$i_\lambda = x_0(e^{\lambda t} - 1)/\lambda^2$$

and the statistical curvature γ_λ is given by

$$\gamma_\lambda^2 = \frac{1}{x_0}\left[\frac{1}{1 - e^{-\lambda t}} - \frac{(\lambda t)^2 e^{2\lambda t}}{(e^{\lambda t} - 1)^3}\right].$$

In the spirit of the paper, we quote some values of γ_λ^2 $(x_0 = 1)$ in Table 1.

Two asymptotic schemes are inviting: large initial population size $(x_0 \to \infty)$ for fixed t and large observation period $(t \to \infty)$ for fixed x_0. Being a branching process, a birth process with $X_0 = x_0$ may be interpreted as a sum of x_0 birth processes with $X_0 = 1$ and the same λ. Therefore the first scheme is still within the realm of independent identical replications, and may be treated with the methods of Efron's paper. This was done by Beyer, Keiding and Simonsen (1975) for this case as well as for the life-testing situation outlined above.

The second scheme, however, is a "real" stochastic process situation, and we encounter here the trouble that the minimal sufficient statistic is not consistent.

TABLE 1

Statistical curvature for the birth process with $x_0 = 1$

λt	0	0.1	0.5	1	2	5	∞
γ_λ^2	0	0.009	0.052	0.125	0.319	0.835	1

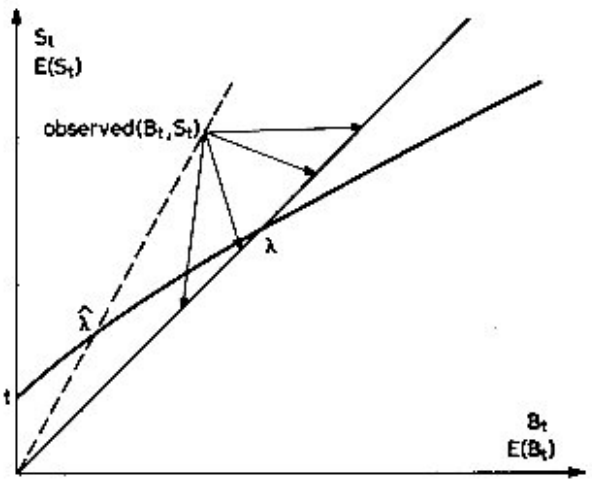

FIG. 1. The canonical sample space of the birth process estimation problem. The curve is the statistical model corresponding to $0 < \lambda < \infty$ (mean value parametrisation). The full-drawn line is the set of points for which $B_t = \lambda S_t$ where λ is the "true" value, and the broken line is the set where $B_t = \hat{\lambda} S_t$.

in fact, as $t \to \infty$

$$e^{-\lambda t}(B_t, S_t) \to (1, \lambda^{-1})W$$

almost surely, where the random variable W is gamma distributed with form parameter x_0 and expectation x_0. Nevertheless $\hat{\lambda} \to \lambda$ a.s., as illustrated in Figure 1. Here λ^{-1} is the slope of the full-drawn line, $\hat{\lambda}^{-1}$ is the slope of the broken line (connecting the observed (B_t, S_t) and the origin.) Normalising with $e^{-\lambda t}$, the minimal sufficient statistic will converge towards some $(1, \lambda^{-1})W$ (shown by arrows), but the empirical line will always converge towards the correct line.

In the standard situation the asymptotic normality of $\hat{\theta}$ is based upon the asymptotic normality of the minimal sufficient statistic combined with pure differential geometry, as noted by Efron in Section 9. It is therefore no surprise that asymptotic normality breaks down here. Notice also that $\gamma_\lambda \to 1$ (not 0) as $t \to \infty$. However, for given "nuisance statistic" W, the minimal sufficient statistic *is* asymptotically normal with asymptotic variance proportional to W^{-1}, and hence also $\hat{\lambda}$ is asymptotically normal. (Marginally, the distribution of $e^{\lambda t/2}(\hat{\lambda} - \lambda)$ converges towards a Student distribution with $2x_0$ d.f., which may be interpreted as the mixture of the normal distributions over the gamma distributed inverse variances.)

It is thus tempting to investigate the problem obtained by conditioning on $W = w$, replacing the "nuisance statistic" W by a nuisance parameter w, see Keiding (1974). The resulting "conditional" maximum likelihood estimator λ^* has the same first-order efficiency properties as $\hat{\lambda}$. A comparison of second-order efficiencies is not yet completed.

4. A more general aspect of the last example is: can curved exponential families be "avoided"? In the birth process situation a stopping rule like "sample until $X_t = n$" will make the minimal sufficient statistic one-dimensional, in fact equal to S_τ, $\tau = \inf\{t \mid X_t = n\}$. Also it should be mentioned that conditioning on statistics which are in some sense ancillary (see Barndorff–Nielsen (1973) for a survey of ancillarity) may completely change the curvature properties of the problem.

REFERENCES

BARNDORFF-NIELSEN, O. (1973). *Exponential Families and Conditioning*. Univ. of Copenhagen.

BARNDORFF-NIELSEN, O. (1974). Factorization of likelihood functions for exponential families. *J. Roy. Statist. Soc. Ser. B.* (Submitted).

BARNDORFF-NIELSEN, O. and BLAESILD, P. (1975). *S*-ancillarity in exponential families. To appear in *Sankhyā A* 37.

BEYER, J. E., KEIDING, N. and SIMONSEN, W. (1975). The exact behaviour of the maximum likelihood estimator in the pure birth process and the pure death process. *Scand. J. Statist.* **2**. To appear.

FISHER, R. A. (1958). *Statistical Methods for Research Workers*. 13th ed. Oliver & Boyd, Edinburgh.

KEIDING, N. (1974). Estimation in the birth process. *Biometrika* **61** 71–80 and 647.

RAO, C. R. (1963). Criteria of estimation in large samples. *Sankhyā A* **25** 189–206.

SUNDBERG, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.* **1** 49–58.

A. P. DAWID

University College London

With his introduction of the concept of statistical curvature, Professor Efron has provided, not merely a valuable theoretical tool, but a new way of looking at statistical problems which at once unifies what has gone before and opens up new territory.

The general study of curvature belongs to Differential Geometry, a subject which has proved an invaluable tool in Physics, both Newtonian and Einsteinian. It may have much to offer Statistics. A good introduction is Laugwitz (1965) while Hicks (1965) emphasises a coordinate-free approach more suitable for Statistics.

In general differentiable spaces, we cannot talk about curvature until we have chosen, somewhat arbitrarily, a *linear connexion*: this defines what we mean by "displacement of a vector *parallel to itself* along a curve." For example, consider an observer who lives and measures on a plane inverted in its unit circle. To him, a circle through the origin looks like a straight line, and he would consider its tangents as parallel; to us they are not. The need for the parallelism concept may be seen from Efron's Figure 1: a_θ is the angle between (i) $\dot{\gamma}_{\theta}$ and (ii) $\dot{\gamma}_{\theta_0}$ *displaced parallel to itself* along $\mathscr{S}$ to γ_{θ}. This depends on our connexion.

Let us try to frame Statistics within Differential Geometry as follows (ignoring obvious technical difficulties): Let $\mathscr{P}$ be the family of all distributions over $\mathscr{X}$

equivalent to a carrier measure μ. A *curve* $\mathscr{C}$ in $\mathscr{P}$ is a 1-parameter family in $\mathscr{P}$, say $\mathscr{C} = \{P_\theta\}$ with densities $\{f_\theta\}$, having suitable regularity properties.

If $\mathscr{M}$ is the vector space of signed measures m on $\mathscr{X}$, with $m \ll \mu$ and $m(\mathscr{X}) = 0$, we may define the *tangent* to $\mathscr{C}$ at $P = P_\theta$ as $m_\theta \in \mathscr{M}$, with $m_\theta = $ "$\lim_{\delta \to 0}$" $(P_{\theta-\delta} - P_\theta)/\delta$. (Equivalently, $dm_\theta/d\mu = \dot{f}_\theta$). Conversely $m \in \mathscr{M}$ is tangent to some curve at P.

Let $\mathscr{V}_P$ be the vector space of random variables $T(x)$ having $E_P|T(X)| = 0$. For given P, there is a natural isomorphism between $\mathscr{M}$ and $\mathscr{V}_P$: $dm = T(x)\,dP$. Then m_θ maps into $\dot{l}_\theta(x)$, which may again be identified with the *tangent* to $\mathscr{C}$ at P_θ.

Now let $P_{\theta_0}, P_{\theta_1} \subset \mathscr{C}$, with tangent spaces $\mathscr{V}_0$, $\mathscr{V}_1$, and let $T_0 \subset \mathscr{V}_0$, $T_1 \in \mathscr{V}_1$. To be able to talk about the angle between T_0 and T_1 we must put them into the same space. We may do this by a *parallel displacement* of T_0 along $\mathscr{C}$ to θ_1, where it becomes $T_0' \in \mathscr{V}_1$.

The parallel displacement used implicitly by Efron—what I propose to call the "Efron connexion"—has

$$T_0' = T_0 - E_{\theta_1}(T_0) \,. \tag{1}$$

This happens to be independent of the curve $\mathscr{C}$, which is not always so. Noting $(d/d\theta)E_\theta[T] = E_\theta[T\dot{l}_\theta]$ for fixed T, we can generate (1) by the infinitesimal displacement rule (having $\theta_1 = \theta_0 + d\theta$):

$$T_0' = T_0 - E_{\theta_0}(T_0 \dot{l}_{\theta_0}) \cdot d\theta \,. \tag{2}$$

For curvature, we look at the angle between $\dot{l}'_{\theta_0} = \dot{l}_{\theta_0} - \ddot{l}_{\theta_0} \cdot d\theta$ and $\dot{l}_{\theta_1} = \dot{l}_{\theta_0} + \ddot{l}_{\theta_0} d\theta$. We may measure this by any convenient inner product, but in our statistical set-up there appears to be only one natural inner product in $\mathscr{V}_P$, namely $\langle T, U\rangle = E_P(TU)$. (For any parametric family $\{P_\phi\}$, this yields the information inner product, with matrix $(E_\phi[(\partial l/\partial\phi_i)(\partial l/\partial\phi_j)]$.) Hence we may call this the *information metric*). This leads to Efron's measurement of angle and of curvature.

The "straight lines" have a characterisation independent of the metric: $\ddot{l}_{\theta_0}$ must displace to become a scalar multiple of $\dot{l}_{\theta_1}$. By reparametrisation, the multiple may be taken as unity. This leads to the differential equation

$$\dddot{l}_\theta + \dot{l}_\theta = 0 \tag{3}$$

characterising exponential families.

The Efron connexion is not, however, the only available one (although it probably is the only one that fits in neatly with repeated sampling, as in Efron's Section 6). An alternative obvious definition of parallel displacement considers $\mathscr{M}$ as the tangent space and uses the identity transformation (again, independent of path). This is equivalent to transforming $\mathscr{V}_0$ into $\mathscr{V}_1$ with

$$T_0' = T_0 \cdot \left(\frac{dP_{\theta_0}}{dP_{\theta_1}}\right), \tag{4}$$

yielding the infinitesimal displacement

$$T_\theta' = T_\theta - T_\theta \dot{l}_{\theta_0} \cdot d\theta \,. \tag{5}$$

To measure curvature with this connexion, using the information metric, M_θ in Efron's (2.3) must be replaced by the covariance matrix of $\dot{l}_\theta$ and $(\ddot{l}_\theta + \dot{l}_\theta^2)$. The "straight lines" now have $\ddot{l}_\theta + \dot{l}_\theta^2 = 0$, which yields mixture families: $P_\theta = (1-\theta)P_0 + \theta P_1$. Thus the above connexion may be termed the "mixture connexion".

Now the information metric makes $\mathscr{S}$ into a *Riemannian space*, and from this point of view there is a serious deficiency in both connexions above: they are *not compatible* with the metric. That is, the length of T_θ at $P_{\theta_0}(\text{viz } |E_{\theta_0}(T_\theta^2)|^{\frac{1}{2}})$ is not the same as that of its parallel translate T_θ' at P_{θ_1}. It may be checked that the infinitesimal displacement

$$T_\theta' = T_\theta - \tfrac{1}{2}[T_\theta \dot{l}_{\theta_0} + E_{\theta_0}(T_\theta \dot{l}_{\theta_0})] \cdot d\theta \tag{6}$$

yields a connexion—the "information connexion"—that *is* compatible with the information metric. Curvature for this connexion (which is the *geodesic curvature* associated with the information metric) uses the covariance matrix of $\dot{l}_\theta$ and $\ddot{l}_\theta + \frac{1}{2}\dot{l}_\theta^2$.

We can calculate the *torsion* and *curvature tensors* (Hicks, page 59) for the above connexions. We find that all have zero torsion (equivalently: are *symmetric*, or *affine*). There is a unique affine connexion compatible with a given metric, hence (6) supplies it for the information metric.

We find zero curvature for the Efron and mixture connexions, while the curvature tensor R associated with the information connexion has

$$R(T, U)V = \tfrac{1}{4}[T \cdot E(UV) - U \cdot E(TV)] \,. \tag{7}$$

The Riemann–Christoffel curvature tensor K of type 0, 4 (Hicks, page 72) is then given by:

$$K(T, U, V, W) = \tfrac{1}{4}[E(TV)E(UW) - E(TW)E(UV)] \,. \tag{8}$$

From this we find that the space $\mathscr{S}$, with the information metric, has constant, positive, Riemannian curvature $\frac{1}{4}$.

The *geodesics* (shortest paths) for the information metric are the "straight lines" of the information connexion, satisfying

$$\ddot{l}_\theta + \tfrac{1}{2}\dot{l}_\theta^2 + \tfrac{1}{2}i_\theta = 0 \,. \tag{9}$$

Solutions of (9) are *closed* curves, parametrized by an angle θ, having an angle-valued sufficient statistic t, with density of the form

$$f(t \mid \theta) = 1 + \cos(t - \theta) \tag{10}$$

with respect to a probability measure ν over the unit circle for which $\int_0^{2\pi} e^{it}\, d\nu(t) = 0$. Such curves have $i_\theta \equiv 1$, and total length 2π. Thus $\mathscr{S}$ looks rather like the surface of a sphere of radius 2, opposite points being identified.

The nonvanishing of (7) means that the information parallel displacement depends on path, which makes it less immediately intelligible than the Efron and mixture displacements. Can we give any interesting statistical interpretation to the information connexion, and its associated families (10)?

REFERENCES

[1] HICKS, N. J. (1965). *Notes on Differential Geometry*. Van Nostrand, Princeton.
[2] LAUGWITZ, D. (1965). *Differential and Riemannian Geometry*. Academic Press, New York.

JIM REEDS
Harvard University

1. Ideas of geometrical curvature are not completely new to statistics. Efron's paper is the logical successor to papers applying the differential geometric point of view to statistical estimation. Rao (1945) and Bhattacharyya (1943) viewed the multiparameter Fisher information as defining a local (Riemannian) metric (Eisenhart (1926 and 1960), Spivak (1970)) on the parameter space; the integrated arc length of a geodesic connecting two parameter values then defines a global metric or distance function on parameter space. Holland (1973), Huzurbazar (1950 and 1956) and Mitchell (1962) exploited transformation properties of the Fisher information viewed as a Riemannian metric. Holland, for instance, sought covariance stabilizing transformations (like the square root transformation of univariate Poissons). Such a transformation makes the Fisher information matrix, expressed in transformed coordinates, a constant matrix. "When can it be found?" is the question "When is a given Riemannian manifold locally isometric to a Euclidean space?" Riemann gave the answer: "When the Riemannian curvature (or, in two dimensions, the Gaussian curvature) vanishes identically." This always happens only in dimension one. In all higher dimensions non-Euclidean manifolds—and noncovariance stabilizable parameter spaces—occur.

Recent unpublished work of Tadashi Yoshizawa (1971) makes explicit use of the inherent Riemannian structure in parameter estimation problems. He shows how one can isometrically embed the parameter space into a Euclidean space of sufficiently high dimension, and then read off the (first order) asymptotic properties of the estimation problem by inspecting the parameter space as a curved submanifold of a Euclidean space.

Thus curvature of one sort is not new to statistics. But Efron's curvature is of a different sort—not the Riemannian or "intrinsic" curvature but instead the curvature of embedding, associated with the particular way a parameter space is placed inside a higher-dimensional "natural parameter" space. Riemann curvature—measured by the curvature tensor—is determined solely by the "first fundamental form" or metric tensor, the physicists' metric ground form, the statisticians' Fisher information matrix. Efron's curvature, curvature of embedding, is measured by the "second fundamental form" and depends on more than

Fisher information. The distinction is illustrated by a cylinder embedded in Euclidean 3-space. This surface has curvature of embedding but no Riemann curvature, for any piece of it can be unrolled without distorting lengths. A sphere in 3-space has both sorts of curvature; a parabola in the plane has only curvature of embedding. No submanifolds of Euclidean space have Riemann curvature without curvature of embedding.

Efron takes the *natural* parameter space as Euclidean, with *constant* metric given by the Fisher information evaluated at the true value of the parameter, θ_0. The actual parameter space is a submanifold of natural parameter space; its curvature of embedding is calculated with respect to this constant Euclidean structure on the natural parameter space. Efron's discussion in the second paragraph of Section 2 is unclear; one might falsely assume that the natural parameter space was endowed with the (nonconstant) metric provided by the Fisher information as a function of θ.

The point of Efron's paper is that the curvature of embedding, calculated in this way, has an effect on statistical procedures, an effect amenable to quantitative study.

2. The main result of Section 10 may be generalized to a multivariate curved exponential family. Both this result and Efron's suffer from a defect which might be overcome in future work. The defect is that both make statements about the coefficients of the asymptotic expansions of the variance, *not about the variance itself*. Thus, the conclusions are of the form

$$\text{“Var}\,(T_n) = \frac{a}{n} + \frac{b}{n^2} + o(n^{-2}) \quad (\text{or} \quad O(n^{-3}))\,,$$

$$\text{and} \qquad a \geqq \alpha\,, \quad \text{and if} \quad a = \alpha\,, \quad b \geqq \beta\,,\text{”}$$

where α and β are certain theoretical lower bounds. This should be contrasted with a stronger type of conclusion:

$$\text{“Var}\,(T_n) \geqq \frac{\alpha}{n} + \frac{\beta}{n^2}\,,\text{”}$$

where α and β have the same meaning as above. (If T_n is such that $\text{Var}\,(T_n)$ has an asymptotic expansion at all, the second conclusion implies the first.) Both the Cramér–Rao and the Bhattacharyya inequalities provide conclusions of the second type. In a sense, we can trace the difference to the different methods used to prove the various inequalities. The classical proof of the Cramér–Rao bound proceeds by constructing a certain variance-covariance matrix, and using its positive semidefiniteness to get the desired results. This is to be contrasted with the method used in the present theorems: Taylor expansions of the functional form of the estimate, coupled with systematic discarding of negligible terms.

It is conceivable that a proof of the theorem of Section 10 could be constructed by the classical method, by considering the joint covariance of the estimate, the

first derivative of the log likelihood, the square of the first derivative of the log likelihood, and the product of the first and second derivatives of the log likelihood. This is conjectured on the grounds of the simple form the covariance matrix takes, when only terms of order up through $1/n^2$ are considered.

We may define a curved q-parameter exponential family by means of a smooth map $\eta: \Theta \to H$, where Θ is some open subset of $\mathbb{R}^q$, and H is the natural parameter space of a k-variate exponential family. To simplify the discussion that follows, we will assume that η is an embedding in the sense of differential geometry: η is a C^∞ injection, with differential of full rank at each point, and that "smooth"—whenever it appears in this discussion—means C^∞. Note that according to this set-up, Θ is not a submanifold of H; but $\eta(\Theta)$ is. An estimate is a function $T: \mathcal{X}^n \to \Theta$, mapping the space of the sufficient statistic to the parameter space.

If we restrict ourselves to estimates T that depend only on the sufficient statistic $\bar{x}_n = n^{-1}(x_1 + \cdots + x_n)$ (and not on n), and which satisfy certain regularity conditions, we may prove:

THEOREM. *Let T depend only on $\bar{x}_n$, the sufficient statistic for a curved q-parameter exponential family. Suppose T is smooth in some neighborhood of $E(\bar{x}_n)$, and suppose T grows (as a function of $\bar{x}_n$) no faster than exponentially.*

If T is a consistent and first order efficient estimate of θ, the variance of T possesses an asymptotic expansion

$$\operatorname{Var}(T(\bar{x}_n)) = \text{CRLB} + \frac{A}{n^2} + \frac{B}{n^3} + \frac{C}{n^3} + O(n^{-3}).$$

(Here CRLB denotes the Cramér–Rao Lower Bound,

- A denotes the "naming" or "Bhattacharyya" curvature, which can be made zero by an appropriate reparameterization of parameter space. It is independent of T.
- B is the "Efron excess", or statistical curvature term and is independent of T.
- C depends only on the function T, and vanishes for the particular choice $T =$ the maximum likelihood estimate.

All these quantities are q by q positive semidefinite symmetric matrices.)

The proof of this multivariate theorem parallels Efron's univariate arguments. It shares the use of affine transformations to bring the problem into "standard form," calculations with Taylor expansions to exhibit the consequences of consistency and first order efficiency, and finally, replacement of T by a Taylor approximation, and the calculation of expectations and variances of the Taylor approximant.

The key quantity of interest in the conclusion of this theorem is the term B, the "Efron" or "statistical curvature" excess. It is the multivariate generalization of γ^2/i, and (like γ^2/i) may be defined in several ways.

(1) Let $\gamma_i(\theta)$, in the vicinity of θ_0, have an expansion

$$\gamma_i(\theta) = a^i + \sum_j b_j^i(\theta^j - \theta_0^j) + \tfrac{1}{2}\sum_{jk} c_{jk}^i(\theta^j - \theta_0^j)(\theta^k - \theta_0^k) + \cdots$$

where θ has coordinates $(\theta^1, \theta^2, \cdots, \theta^r)$. Let (g^{ij}) denote the inverse of the Fisher information matrix for θ, and let (G_{rs}) denote the Fisher information matrix for the natural parameter γ_i. Let

$$D_{ik} = \sum_{rs} b_i^r G_{rs} b_k^s ,$$

$$E_{i,mn} = \sum_{rs} b_i^r G_{rs} c_{mn}^s$$

and

$$F_{jk,mn} = \sum_{rs} c_{jk}^r G_{rs} c_{mn}^s .$$

Let the inverse of $D = (D_{ik})$ be $D^{-1} = (D^{ij})$. Let

$$\bar{F}_{jk,mn} = F_{jk,mn} - \sum_{ih} E_{i,jk} F_{h,mn} D^{ih} .$$

Then

$$B^{ij} = \sum_{kl}\sum_{mn} g^{im} g^{jn} g^{kl} \bar{F}_{mk,nl} .$$

If, at θ_0, the Fisher matrices of both θ and γ_i are equal to identity matrices, this simplifies to

$$B^{ij} = \sum_{k,r} c_{ik}^r c_{kj}^r ,$$

where the summation extends over $r \geqq q + 1$.

(2) Let l be the log likelihood function. If

$$l_i = \frac{\partial}{\partial \theta_i} l$$

and

$$\ddot{l}_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} l ,$$

we may form the linear regression of $\ddot{l}$ on $\dot{l}$ as follows:

$$\hat{\ddot{l}}_{jk} = \sum_i \beta_{jk}^i \dot{l}_i ,$$

and we may calculate the regression-residual variance:

$$\operatorname{Cov}(\ddot{l}_{ij} - \hat{\ddot{l}}_{ij}, \ddot{l}_{mn} - \hat{\ddot{l}}_{mn}) = \alpha_{ij,mn} .$$

Then

$$B^{ij} = \sum_{mn}\sum_{kl} g^{im} g^{jn} \alpha_{mn,ln} g^{kl} .$$

(3) Let Ω_{rij} be the components of the second fundamental form of the imbedding $\gamma_i : \Theta \to H$ (see Eisenhart (1926 and 1960)) where H has the Euclidean structure induced by the Fisher information evaluated at $\gamma_i(\theta_0)$. Then

$$B^{ij} = \sum_{mn}\sum_{kl}\sum_r g^{im}\Omega_{r|mn} g^{kl}\Omega_{r\,kn} g^{nj} .$$

Similar formulas hold for the naming curvature term A. In the special case where both the Fisher information matrices are equal to identity matrices (at θ_0) and where

$$b_j^i = \left.\frac{\partial \gamma_i^i}{\partial \theta^j}\right|_{\theta_0} = \delta_{ij} ,$$

the ijth term of the naming curvature is given by

$$A^{ij} = \sum_{a,b} \left(c^i_{ab} + \frac{\partial G_{ab}}{\partial \theta_i}\Big|_{\theta_0} \right) \left(c^j_{ab} + \frac{\partial G_{ab}}{\partial \theta_j}\Big|_{\theta_0} \right),$$

where the summation extends over $1 \leqq a, b \leqq q$.

Notice that in the univariate case the naming curvature term A can always be made to vanish identically by a suitable reparameterization, but in the multivariate case this cannot in general be done. It *can* always be made to vanish at isolated points, but there need not in general exist reparameterizations which make the naming curvature vanish globally. This is related to the general nonexistence of multivariate covariance stabilizing transformations. In the univariate case, the naming curvature vanishes identically exactly when we parameterize the curve by arc length: that is, it vanishes when the variance is stabilized. In the multivariate setting, however, we cannot in general covariance stabilize, and we cannot in general make the naming curvature identically zero. Perhaps the easiest example is provided by the trivariate normal distribution, with unit covariance matrix, with the mean vector constrained to have unit length (and, to avoid global topological problems, with first coordinate positive). Thus, in the multivariate case the naming curvature term takes on added significance, and must be viewed as serious an object of study as the statistical curvature term itself.

REFERENCES

[1] BHATTACHARYYA, A. (1943). On a measure of divergence between two statistical populations. *Bull. Calcutta Math. Soc.* **35** 99–109.

[2] EISENHART, L. (1926 and 1960). *Riemannian Geometry*. Princeton Univ. Press.

[3] HOLLAND, P. (1973). Covariance stabilizing transformations. *Ann. Statist.* **1** 84–92.

[4] HURZURBAZAR, V. (1950). Probability distributions and orthogonal parameters. *Proc. Cambridge Philos. Soc.* **46** 281.

[5] HURZURBAZAR, V. (1956). Sufficient statistics and orthogonal parameters. *Sankhyā* **17** 217–220.

[6] MITCHELL, A. (1962). Sufficient statistics and orthogonal parameters. *Proc. Cambridge Philos. Soc.* **58** 326–337.

[7] RAO, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37** 81–91.

[8] SPIVAK, M. (1970). *Differential Geometry*. Publish or Perish, Boston.

[9] YOSHIZAWA, T. (1971). Memorandum TYH-2, *A Geometrical Interpretation of Location and Scale Parameters*. Statist. Dept., Harvard Univ. Cambridge.

REPLY TO DISCUSSION

The discussants are (almost) uniformly constructive and informative in their comments. They point out many important facts, and even whole areas, that the paper misses. Only two of them consider me basically deranged in my thought processes. In what follows I have tried to answer a few specific points, without exploring much further the bigger questions raised.

Professors Cox and Pierce suggest that the distance from $(\bar{x}_1, \bar{x}_2)$, to λ_0 is a

useful approximate ancillary statistic. (See Figure 4. It is simplest to assume that the family $\mathscr{F}$ is in standard form at $\theta = 0$, and that we are considering θ values near zero.) I particularly like Pierce's suggestion that the ancillary information has to do with the *precision* of $\hat{\theta}$ and not its location. To make things really easy, consider repeated sampling in Example 1, and suppose that we happen to get $\hat{\theta} = 0$, that is $\bar{x}_1 = 0$. (See Figure 2.) The likelihood function for θ is proportional to $\exp\{-(n/2)[1 - \gamma_0\bar{x}_2 - \gamma_0^2\theta^2/4]\theta^2\}$ which for θ in the interval $\hat{\theta} \pm c/n^{\frac{1}{2}}$ behaves like $\exp\{-(n/2)[1 - \gamma_0\bar{x}_2]\theta^2\}$. That is, the likelihood function for θ is approximately $\mathscr{N}(\hat{\theta}, [1 - \gamma_0\bar{x}_2]/(ni_{\hat{\theta}}))$. The distance from $(\bar{x}_1, \bar{x}_2)$ to $\hat{\theta}, \bar{x}_2$ in this case, modifies the unconditional variance $(ni_{\hat{\theta}})^{-1}$ by the factor $[1 - \gamma_0\bar{x}_2]$. It is probably possible to extend this likelihood analysis to a genuine conditional variance statement, as Pierce suggests.

Bayesians and other nonfrequentist statisticians do not like averages taken over the sample space $\mathscr{X}$ with θ fixed. Professor Lindley raises this objection to the curvature γ_θ^2, as it has been raised to the Fisher information i_θ itself. Those who believe in direct interpretation of likelihood functions prefer $-\ddot{l}_{\hat{\theta}}(x)$, the actual curvature of the log likelihood function at its maximum, to the average value i_θ. (Incidentally, I use θ as a subscript rather than an argument to save writing parentheses!) I find some force in these kinds of considerations but, perhaps because of my training, can never be convinced without the support of some relevant averaging property, be it frequentist, conditional frequentist, Bayesian, or otherwise. (See my discussion following Blyth (1970).)

If a Cauchy translation sample of size 10 yields a very normal looking likelihood function, say $\mathscr{N}(0, .3)$, should we behave as if the MLE has variance .3? Professor Lindley answers "yes" on Bayesian grounds, in the absence of prior information. Professor Pierce's remarks indicate that the curvature may have something helpful to say to frequentists about such problems.

Returning to less slippery ground, here is a calculation of asymptotic Bayes risk that makes use of the curvature. In a curved exponential with an i.i.d. sample of size n, let θ have prior distribution $\mathscr{N}(\theta_0, c_n/n)$, where c_n is going sufficiently slowly to infinity. Then the Bayes risk is asymptotically

$$\frac{1}{n_{i_{\theta_0}}} + \frac{1}{n^2 i_{\theta_0}}\left\{\gamma_{\theta_0}^2 + \frac{4\Gamma_{\theta_0}^2}{i_{\theta_0}}\right\} + o\left(\frac{1}{n^2}\right)$$

which equals to order $1/n^2$ the squared error risk of the biased corrected MLE at $\theta = \theta_0$. (This result follows, with some effort, from (10.19).)

Professor Le Cam's warning about over-reliance on local methods is well taken. As a matter of fact, my paper is most concerned with curvature as a check on the appropriateness of first order local properties such as Fisher's information and the Cramér–Rao lower bound. In the situation of Figure 6, curvature can be used quantitatively to improve the first order approximation. I hope, but of course am not certain, that other situations will be similarly obliging.

Le Cam's criticism of the MLE as a point estimator should not be confused

with Fisher's preference for it is an information gatherer. A function of the MLE may be better than the MLE itself for any specific estimation problem. This is the case in the Berkson example quoted. Berkson finds a "better" estimator than the MLE, which eventually is improved by Rao-Blackwellizing it on the sufficient statistics. This gives a function of the MLE! (It has to because the situation involves a genuine uncurved exponential family.) Figure 4 becomes more convincing the more you study it. Locally the straight level line $L_{\hat{\theta}}$ seems intuitively preferable to any curved competitor $M_{\hat{\theta}}$. (See Dr. Keiding's remarks and my reply.)

Quadratic approximations to the log likelihood function have been used successfully by many authors, notably Professor Le Cam himself. They are the basis of Rao's work in second order efficiency. They can be used to produce estimators other than the MLE which are second order efficient. Whether there is a corresponding theory of third order efficiency, and whether the MLE is still the champion, is an interesting open question.

After a long fallow period there seems to be a revival of interest in second order efficiency and related topics. I am eager to see Professor Ghosh's work with Subrahmaniam and Srinivasan. (Also, I must apologize for not having been aware of Pfanzagl and Chibisov's papers, which demonstrate rigorously the relevance of what I have called curvature to hypothesis testing problems, even outside an exponential family framework.) As Ghosh suggests and as I mentioned in discussing Pierce's comments, there is some connection between γ_θ^2 and the geometrical curvature of the likelihood function, but not one I understand clearly yet. Professor Ghosh's last question can be partially answered in the affirmative: in the counter-example of Figure 5, change c to $(-2^{\frac{1}{2}}, \frac{1}{3})$. Then the MLE of any $\bar{x}$ vector with $\bar{x}_1 = \frac{1}{3}$ is zero, but if $\bar{x}_1 \neq \frac{1}{3}$ each $\bar{x}$ corresponds to a unique $\hat{\theta}$. For n any multiple of 3, $\hat{\theta}$ will lose information because of the grouping of those $\bar{x}$ vectors with $\bar{x}_1 = \frac{1}{3}$. It is easy to curve the level lines of another consistent efficient estimator $\tilde{\theta}$, à la Figure 4, so that the vectors with $\bar{x}_1 = \frac{1}{3}$ are separated, and $\tilde{\theta}(\bar{x})$ is different for all different $\bar{x}$ vectors, so no information is lost. This works for any fixed n divisible by 3, but I am less certain about finding a $\tilde{\theta}$ that works for all values of n.

There is less difference between Professor Pfanzagl and me than the tone of his comments indicates. His results (1) and (2) follow from (8.4). I should have said earlier that a rescaled version of this equation holds as an approximation when testing $\theta = 0$ versus $\theta > 0$ under i.i.d. sampling in any curved exponential family,

$$1 - \beta_{\tilde{\theta}_1}(\tilde{\theta}) \approx \Phi(\tilde{\theta}(1 + \tilde{\gamma}_0^2\tilde{\theta}^2/4)^{\frac{1}{2}} \cos(A_{\tilde{\theta}_1} - A_{\tilde{\theta}}) - z_\alpha),$$

where $\tilde{\theta} \equiv (ni_0)^{\frac{1}{2}}\theta$, $\tilde{\gamma}_0 \equiv \gamma_0/n^{\frac{1}{2}}$, and $A_{\tilde{\theta}} \equiv \tan^{-1}(\tilde{\gamma}_0\tilde{\theta}/2)$. In order for this approximation to be sufficiently accurate to yield Pfanzagl's asymptotic results, the family must be nonatomic. However, the type of power comparisons presented in Table 3 are less sensitive as well as more familiar. For $\alpha = .01$, power $= .99$,

$\gamma_0^2/n = \frac{1}{8}$, the case Pfanzagl discusses, the locally most powerful test has approximate power .94 compared with the envelope value .99. I consider this borderline acceptable, and will stick to my suggestion of $\gamma_0^2/n > \frac{1}{8}$ as a rough indicator of nonnegligible curvature effects.

Fisher defined γ_θ^2 as the loss of information in using $\hat{\theta}$ instead of the whole sample. Rao's results on estimation with squared error loss partially vindicate this definition. Pfanzagl's own work shows that γ_θ^2 plays a key role in the loss of effective sample size in hypothesis testing problems. Then why does he seem to say that γ_θ^2 has no statistical significance? The fact that the level α test based on the MLE is asymptotically equivalent to the β optimal test with power $1 - \alpha$ has nothing to do with the existence of curvature effects. There still is no uniformly most powerful test. The global deviations of any attainable power curve from the power envelope are still ruled by the magnitude of γ_θ^2.

I was happy to see that Dr. Keiding had found a definite use for curved exponential families in his work on birth processes. Time series problems offer many other examples, of which my Example 3 is close to the simplest. (With Dr. Reeds' multiparameter theory available we are now in a position to analyze the second order asymptotics of higher autoregressive schemes.) The geometric interpretation of the penalty $\Delta_{\theta_0}^{\bar{\theta}}$ for not using the MLE is simple in the case $r = 2$. Comparing (10.24) with (10.5) shows that it equals one-half of the squared curvature of the level curve $M_{\theta_0} = \{\bar{x} : \bar{\theta}(\bar{x}) = \theta_0\}$. See Figure 4.

Dr. Dawid raises a deep question: why have I chosen to represent families of probability distributions by their log densities rather than, say, the density functions themselves? This latter representation would make mixture families rather than exponential families straight lines, as he points out. What I have called the matrix M_θ then has elements μ_{kj} as at (1.2) rather than ν_{kj} as at (3.21). Dawid makes the interesting observation that still another definition is needed to make straight lines into geodesics in the information metric. (Rao 1945a and 1945b, has proposed using this type of geodesic distance to measure the separation of probability distributions. Atkinson and Mitchell have calculated Rao distances for many familiar distribution families.) I can't answer Dr. Dawid's deep question except to say that my definition was motivated by what seemed to be the most pressing statistical considerations. He makes a good case for other definitions also yielding useful results for the statistician.

My paper considers only one parameter families. Dr. Reeds gives a convincing extension to the multiparameter case. Having been frustrated myself by the intricacies of the higher order differential geometry, I am impressed! Hopefully, his "B", the analogue of γ_θ^2, will also play the correct corresponding role vis-à-vis Fisher information and hypothesis testing.

Two technical comments: (i) a version of the usual super-efficiency examples prevents Reeds' formula (2) from holding generally. In my Example 1, Figure 2, let $\bar{\theta}(\bar{x}) = \bar{x}_1$ except in a band of width $\pm x_1^3$ on either side of $\mathscr{L}$. Within this band modify $\bar{\theta}$ so that it is consistent and first order efficient. Then (10.19) can

be used to show that $\tilde{\theta}$ satisfies (10.1) with the $1/n^2 i_{\theta_0}\{\ \}$ term set equal to zero at $\theta_0 = 0$. (ii) It is not true in general, even in the one parameter case, that the "arc-length parameter" has naming curvature equal to zero. Let $\sigma(\theta)$ be this parameter measured from $\theta = \theta_0 = 0$, where we assume for convenience that $i_0 = 1$. By definition $\sigma(\theta) = \int_0^\theta i_{\theta'}^{\frac{1}{2}} d\theta'$ so that $d\sigma(\theta)/d\theta = i_\theta^{\frac{1}{2}}$, $d^2\sigma(\theta)/d\theta^2 = (di_\theta/d\theta)/2(i_\theta)^{\frac{1}{2}}$. It is easy to show by an expansion similar to (10.10) that in terms of the quantities $\mu_{k,l}$ defined at (1.2),

$$di_\theta/d\theta = 2\mu_{11} - \mu_{30}\,.$$

This gives the Taylor expansion about zero

$$\sigma(\theta) = \theta + \left(\mu_{11} - \frac{\mu_{30}}{2}\right)\frac{\theta^2}{2} + o(\theta^2)\,,$$

μ_{11} and μ_{30} being evaluated at $\theta = 0$.

The parameter $\phi(\theta)$ which figures in the definition of Γ_{θ_0} in (10.1) has Taylor expansion

$$\phi(\theta) = \theta + \mu_{11}\frac{\theta^2}{2} + o(\theta^2)$$

as given in (10.11). Therefore the naming curvature $\Gamma^2_{\theta_0}$ will not be zero for the arc-length parameter unless $\mu_{30} = 0$. (That is, Fisher's score function has third moment zero.)

It is not clear to me whether or not one can always choose a reparameterization for $\mathscr{F}$ which has naming curvature identically zero, even in the one-parameter case. We probably wouldn't want to estimate such a parameter anyway unless it had something more to recommend it than $\Gamma^2_\theta = 0$. I didn't mean to imply that naming curvature is less important than statistical curvature, onyl that it depends on the name.

Finally, I would like to thank the Editor for arranging this discussion which involved a large amount of extra work on his part. I hope the Annals of Statistics will continue the entertaining and enlightening policy of providing occasional discussion papers.

REFERENCES

[1] ATKINSON, C., and MITCHELL, A. F. *Rao's Distance Measure*. Unpublished.

[2] BLYTH, C. R. (1970). On the inference and decision models of statistics. *Ann. Math. Statist.* 3 1034–1058.

[3] RAO, C. R. (1945a). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* 37 81–91.

[4] RAO, C. R. (1945b). On the distance between two populations. *Sankhyā* 9 246–248.