The Annals of Statistics
1990, Vol. 18, No. 1, 38–62

FISHER'S INFORMATION IN TERMS OF THE HAZARD RATE[1]

BY BRADLEY EFRON AND IAIN M. JOHNSTONE

Stanford University

If $\{g_\theta(t)\}$ is a regular family of probability densities on the real line, with corresponding hazard rates $\{h_\theta(t)\}$, then the Fisher information for θ can be expressed in terms of the hazard rate as follows:

$$\mathscr{I}_\theta \equiv \int \left(\frac{\dot{g}_\theta}{g_\theta}\right)^2 g_\theta = \int \left(\frac{\dot{h}_\theta}{h_\theta}\right)^2 g_\theta, \qquad \theta \in \mathbb{R},$$

where the dot denotes $\partial/\partial\theta$. This identity shows that the hazard rate transform of a probability density has an unexpected length-preserving property. We explore this property in continuous and discrete settings, some geometric consequences and curvature formulas, its connection with martingale theory and its relation to statistical issues in the theory of life-time distributions and censored data.

1. Introduction. Fisher's information for the parameter θ in a family of density functions $g_\theta(t)$ on the real line is defined to be

$$\mathscr{I}_\theta \equiv \int_{-\infty}^{\infty} \left[\frac{\dot{g}_\theta(t)}{g_\theta(t)}\right]^2 g_\theta(t)\, dt, \tag{1.1}$$

where the dot indicates differentiation with respect to θ,

$$\dot{g}_\theta(t) \equiv \frac{\partial}{\partial\theta} g_\theta(t). \tag{1.2}$$

This definition assumes that we are dealing with a family of continuous distributions on the real line for which the partial derivative (1.2) exists, see Section 5a.4 of Rao (1973), Section 2.6 of Lehmann (1983) or Section I.7 of Ibragimov and Has'minskii (1981). The parameter θ may be a p-dimensional vector, in which case

$$\dot{g}_\theta(t) \equiv \left(\frac{\partial g_\theta(t)}{\partial\theta_1} \cdots \frac{\partial g_\theta(t)}{\partial\theta_p}\right)'$$

and $\mathscr{I}_\theta$ is the $p \times p$ Fisher information matrix. The results which follow hold for the matrix case, after the obvious notational changes, but for the sake of simple exposition we will take θ real-valued.

Received January 1987; revised April 1989.

[1]Research supported in part by NSF Grants DMS-86-00235 and 84-51750 (Johnstone) and NIH Grant PHS GM21215-12 and the Sloan Foundation (Johnstone).

AMS 1980 *subject classifications.* Primary 62F12, 62G05; secondary 62P10.

Key words and phrases. Score function, length-preserving transformation, information metric, proportional hazards model, counting process, martingale, Greenwood's formula, binary trees, tangent space, α-expectation, statistical curvature.

The survival function, or right-sided cumulative distribution function (right cdf) corresponding to density $g_\theta(t)$ is

$$G_\theta(t) \equiv \int_t^\infty g_\theta(s)\,ds = \text{Prob}_\theta\{T \ge t\}, \tag{1.3}$$

T indicating a generic random variable with density $g_\theta(t)$. The *hazard rate* for T is then defined to be

$$h_\theta(t) \equiv \frac{g_\theta(t)}{G_\theta(t)}. \tag{1.4}$$

Hazard rates are useful in discussing life-time distributions. They have the interpretation

$$h_\theta(t)\Delta = \frac{g_\theta(t)\Delta}{G_\theta(t)} \doteq \text{Prob}\{T \in [t, t+\Delta) | T \ge t\}, \tag{1.5}$$

so $h_\theta(t)\Delta$ equals the probability of dying in an infinitesimal interval $(t, t+\Delta)$, conditional upon survival until time t. See Chapter 2 of Cox and Oakes (1984) or Chapter 1 of Kalbfleisch and Prentice (1980) for nice discussions of the hazard rate.

Our central result is an expression for the Fisher information in terms of the hazard rate:

$$\mathscr{I}_\theta = \int_{-\infty}^\infty \left[\frac{\dot{h}_\theta(t)}{h_\theta(t)}\right]^2 g_\theta(t)\,dt \qquad \left[\dot{h}_\theta(t) \equiv \frac{\partial}{\partial\theta} h_\theta(t)\right]. \tag{1.6}$$

In other words one can replace the usual *score function* $(\partial/\partial\theta)\log g_\theta(t) = \dot{g}_\theta(t)/g_\theta(t)$ in (1.1) by its hazard rate analog $(\partial/\partial\theta)\log h_\theta(t)$. [However, $\int_{-\infty}^\infty g_\theta(t)\dot{h}_\theta(t)/h_\theta(t)\,dt$ does not usually equal 0, while

$$\int_{-\infty}^\infty g_\theta(t)\frac{\dot{g}_\theta(t)}{g_\theta(t)}\,dt = 0$$

under mild regularative conditions.] Formula (1.6) is simple and easy to derive, but it has interesting statistical and probabilistic implications, which are the main topic of this paper.

EXAMPLE 1. (a) The negative exponential density: $g_\theta(t) = (1/\theta)e^{-t/\theta}$ for $t \ge 0$, so $\dot{g}_\theta(t)/g_\theta(t) = (t-\theta)/\theta^2$. In this case $h_\theta(t) = 1/\theta$, $\dot{h}_\theta(t)/h_\theta(t) = -1/\theta$, and (1.6) gives $\mathscr{I}_\theta = \int_0^\infty g_\theta(t)(1/\theta^2)\,dt = 1/\theta^2$.

(b) The proportional hazards model of survival analysis [e.g., Cox and Oakes (1984), Chapter 7, and Efron (1977)], sets $h_\theta(t; z) = \psi(z; \theta)h_0(t)$, where z is a covariate vector, $h_0(t)$ an arbitrary baseline hazard rate and θ may be a vector. In this case $\partial \log h_\theta(t; z)/\partial\theta = \partial \log \psi(z; \theta)/\partial\theta$ does not depend on t. If no censoring is present, then $\mathscr{I}_\theta = (\partial\psi/\partial\theta)(\partial\psi/\partial\theta)'/\psi^2$, where prime denotes transpose. For the popular log-linear form $\psi(z; \theta) = e^{z\theta}$, $\mathscr{I}_\theta = zz'$. If the covariate z is time varying, $z = z(t)$, then $\mathscr{I}_\theta = Ez(T)z(T)'$.

EXAMPLE 2. Standard Gaussian distribution: $g_\theta(t) = \phi(t - \theta)$, where $\phi(t) = (2\pi)^{-1/2}e^{-t^2/2}$. Then (1.2) gives $\mathscr{I}_\theta = 1 = \int_{-\infty}^{\infty}(t-\theta)^2\phi(t-\theta)\,dt$. In this case (1.6) produces an identity involving Mills' ratio,

$$1 = \int_{-\infty}^{\infty}\left[(t-\theta) - \frac{\phi(t-\theta)}{1-\Phi(t-\theta)}\right]^2\phi(t-\theta)\,dt, \tag{1.7}$$

where $\Phi(t)$ is the standard normal cumulative $\int_{-\infty}^{t}\phi(s)\,ds$. This example and its generalization to translation families appears in Gill (1980), page 128.

Equality (1.6) says that the functional transformation $\dot{g}_\theta(t)/g_\theta(t) \to \dot{h}_\theta(t)/h_\theta(t)$ preserves length in the L_2 norm $\|b\|^2 \equiv \int_{-\infty}^{\infty}b(t)^2 g_\theta(t)\,dt$. Section 2, which concerns the case where T is continuous, discusses a pair of length-preserving linear transformations on the L_2 space defined by g_θ. Equality (1.6) is then seen as a special case of a more general result [which appears in James (1986)], holding for L_2 functions $b(T)$ of a continuous variate T:

$$\operatorname{var}\{b(T)\} = E\left[b(T) - \bar{b}(T)\right]^2, \quad \text{where } \bar{b}(t) \equiv E\{b(T)|T \geq t\}. \tag{1.8}$$

EXAMPLE 3. Suppose $g(t) = t^{\nu-1}e^{-t}/\Gamma_\nu$ for $t > 0$, so T has a gamma distribution. By considering $-T$ instead of T, we can let $\bar{b}(t) = E\{b(T)|T < t\}$ in (1.8). Then (1.8) gives for $b(t) = t$ the (apparently new) identity

$$E\left[\frac{T - \Gamma_{\nu+1}(T)}{\Gamma_\nu}\right]^2 = \nu, \tag{1.9}$$

where $\Gamma_{\nu+1}(t)$ is the incomplete gamma function $\int_0^t s^\nu e^{-s}\,ds$.

To derive (1.6) from (1.8), choose $b(t) = \dot{g}_\theta(t)/g_\theta(t)$. Then

$$\bar{b}(t) = \frac{1}{G_\theta(t)}\int_t^\infty \frac{\dot{g}_\theta(s)}{g_\theta(s)}g_\theta(s)\,ds = \frac{\dot{G}_\theta(t)}{G_\theta(t)}, \tag{1.10}$$

so

$$b(t) - \bar{b}(t) = \frac{\dot{g}_\theta(t)}{g_\theta(t)} - \frac{\dot{G}_\theta(t)}{G_\theta(t)} = \frac{\dot{h}_\theta(t)}{h_\theta(t)}, \tag{1.11}$$

and (1.8) gives (1.6). (Regularity conditions for this argument appear in Remark A, Section 4.)

There is a more statistical way to look at (1.6) and (1.8). These results are closely related to the left-to-right conditioning calculations which arise in survival analysis and censored data, for example in the Kaplan–Meier estimate, the log-rank (or Mantel–Haenszel) test and Cox's proportional hazard model. [See Miller (1981), Cox and Oakes (1984) and Kalbfleisch and Prentice (1980).] Our results for the case where T is discrete, Section 3, will be proved using the left-to-right conditioning argument. Because of the connection with survival analysis, the generic random variables T will be referred to as "life-times," though of course this has no bearing on the results.

Unified formulas covering discrete and continuous cases are given in Section 4, using the language of counting processes and martingale theory. For example, a martingale argument generalizing that of Section 3 gives a probabilistic proof of (1.6). Among various remarks we present an application of the deviations lemma of Section 3 to Greenwood's formula and an extension to binary trees that includes decomposable models for contingency tables.

Geometrical aspects of (1.6) and (1.8) dominate in the concluding Sections 5 and 6. Although the identity (1.6) is stated for one-parameter families $\{g_\theta\}$, the functional transformation from density $g(t)$ to hazard rate $h(t) = g(t)/G(t)$ depends only on the density g, and not on the family to which it belongs. Choosing different one-parameter families ("curves") through g, one sees that the linear transformation A_g: $\dot{g}_\theta(t)/g_\theta(t) \to \dot{h}_\theta(t)/h_\theta(t)$ depends only on g. This suggests that we think of A_g as a length-preserving linear transformation from a tangent space at g to a tangent space at h. Since in the identity g is essentially arbitrary, these tangent spaces are infinite dimensional. Section 5 tries to make this precise.

Formula (1.6) has other geometric consequences. Section 6 applies it along with extensions derived in Lemma 2.1 to compute the statistical curvature [cf. Efron (1975)] of one-parameter families of distributions specified in terms of hazard functions $h_\theta(t)$.

The survival analysis literature contains much related work. Identity (1.6) is at least implicitly known, since the inverse of the right side of (1.6) appears as the asymptotic variance of the MLE of θ [see, e.g., Borgan (1984)]. James (1986, 1987) emphasizes (1.8) in the study of estimating equations with censored data. Ritov and Wellner (1988) have independently studied the operators A and B of our Section 2 (their R and L), derived (1.8) and given applications to information calculations for regression models for survival data.

In summary, this paper is a set of analytic, statistical and geometric variations on the theme laid out by (1.6) and (1.8). Beyond verifying these our purposes are (i) to understand more fully the mapping $\dot{g}(t)/g(t) \to \dot{h}(t)/h(t)$, its inverse and its length preserving properties, (ii) to derive the discrete analog of (1.8) in a way that suggests the relation with martingale theory, (iii) to understand the statistical basis of (1.6), especially its connection with familiar arguments from the theory of life-time distributions and censored data and (iv) to explore geometric implications—nonparametrically via infinite-dimensional isometries and parametrically through curvature calculations for hazard models. Whether (1.6) or (1.8) dominates depends on whether the viewpoint is parametric or nonparametric. Although martingales and counting processes appear, they are not given their full due, in view of the elegant and sophisticated treatment of Ritov and Wellner (1988).

2. Continuous case. We begin with a direct analytic proof for the general version of the information identity $\mathscr{I}_\theta = E[\dot{h}_\theta(T)/h_\theta(T)]^2$, namely that $\mathrm{var}\{b(T)\} = E[b(T) - \bar{b}(T)]^2$, (1.8), in the continuous case where T has density function $g_\theta(t)$ on the real line. Section 3, which concerns the discrete case, gives another proof of the main result based on more intuitive probability calculations.

The subscript θ plays no role in the identity (1.8): We now assume simply that $g(t)$ is a probability density function with respect to Lebesgue measure and that $G(t) = \int_t^\infty g(s)\,ds$, $h(t) = g(t)/G(t)$. Thus, let $L_2(g)$ be the space of functions which are square integrable with respect to the density g. For a and $b \in L_2(g)$ let $\langle a, b\rangle$ denote the inner product $\int_{-\infty}^{\infty} a(t)b(t)g(t)\,dt$. For any a and b in $L_2(g)$ define transformed functions

$$\bar{b}(s) = \frac{1}{G(s)}\int_s^\infty b(t)g(t)\,dt, \qquad (2.1)$$
$$\tilde{a}(t) = \int_{-\infty}^t a(s)h(s)\,ds.$$

We will be interested in two related linear transformations on $L_2(g)$,

$$(Ab)(s) = b(s) - \bar{b}(s), \qquad (2.2)$$
$$(Ba)(t) = a(t) - \tilde{a}(t).$$

(Our mnemonic: A adjusts for "advance" times, B for "backward" times.)

The proof of (1.8) consists of showing that A and B are essentially adjoint and inverse transformations, from which (1.8) follows as a length-preserving identity for functions in $L_2(g)$.

LEMMA 1. *For functions a and b in $L_2(g)$,*

$$\langle Ba, b\rangle = \langle a - \tilde{a}, b\rangle = \langle a, b - \bar{b}\rangle = \langle a, Ab\rangle, \qquad (2.3a)$$

$$\bar{\tilde{a}} = \tilde{a} + \bar{a} \qquad (2.3b)$$

and

$$\tilde{\bar{b}} = \tilde{b} + \bar{b} - Eb, \qquad (2.3c)$$

where $Eb = \int_{-\infty}^{\infty} b(t)g(t)\,dt$. (The proof of Lemma 1 appears below.)

Thus, A and B are adjoint transformations. Results (2.3b) and (2.3c) show that A and B are also inverses in the following sense:

$$ABa = A(a - \tilde{a}) = (a - \bar{a}) - (\tilde{a} - \bar{\tilde{a}}) = a \qquad (2.4a)$$

and

$$BAb = B(b - \bar{b}) = (b - \tilde{b}) - (\bar{b} - \tilde{\bar{b}}) = b - Eb. \qquad (2.4b)$$

Our main result (1.8) follows immediately from (2.3a) and (2.4c),

$$\operatorname{var}\{b\} = \langle b, BAb\rangle = \langle Ab, Ab\rangle = E[b - \bar{b}]^2. \qquad (2.5)$$

Let $L_2^0(g)$ be those functions b in $L_2(g)$ having $Eb = \int_{-\infty}^{\infty} b(t)g(t)\,dt = 0$. If $a \in L_2(g)$, then $b = Ba = a - \tilde{a} \in L_2^0(g)$. [This follows from (2.3a), taking $b(t) \equiv 1 \equiv \bar{b}(t)$.] Then $b - \bar{b} = Ab = ABa = a$ according to (2.4a). Substituting $b = a - \tilde{a}$ and $b - \bar{b} = a$ in (2.5) gives another form of identity (1.8),

$$E\{a^2\} = E\{a - \tilde{a}\}^2. \qquad (2.6)$$

In summary, the transformations A and B have norm 1 on $L_2(g)$ and the restriction of A to $L_2^0(g)$ equals B^{-1} and provides an isometry of $L_2^0(g)$ onto $L_2(g)$. This development makes it look like A ranges over a space of dimension one greater than B. For the discrete situation, discussed in Section 3, we will see that A and B actually range over spaces of the same dimension.

PROOF OF LEMMA 1. For a and b (and hence the conditional expectation $\bar{b}$) in $L_2(g)$, we apply Fubini's theorem:

$$\begin{aligned}\langle \tilde{a}, b\rangle &= \int_t b(t)\left[\int_{s<t} a(s)h(s)\,ds\right]g(t)\,dt\\ &= \int_s\int_{t\geq s} a(s)h(s)b(t)g(t)\,dt\,ds\\ &= \int_s a(s)g(s)\frac{1}{G(s)}\left[\int_{t\geq s} b(t)g(t)\,dt\right]ds\\ &= \int_s a(s)\bar{b}(s)g(s)\,ds = \langle a, \bar{b}\rangle,\end{aligned} \tag{2.7}$$

verifying (2.3a). [Formula (2.7) shows that $\tilde{a} \in L^2(g)$: $\|\tilde{a}\| = \sup\{\langle a, \bar{b}\rangle/\|b\|: b \in L^2(g)\} \leq \|a\| < \infty$.] Formulas (2.3b) and (2.3c) are also based on an interchange of integrations; we shall consider only (2.3c) here:

$$\begin{aligned}\tilde{\bar{b}}(s) &= \int_{t<s} h(t)\frac{1}{G(t)}\left[\int_{u\geq t} b(u)g(u)\,du\right]dt\\ &= \int_{u<s}\int_{t\leq u}\frac{h(t)}{G(t)}b(u)g(u)\,dt\,du\\ &\quad+\int_{u\geq s}\int_{t<s}\frac{h(t)}{G(t)}b(u)g(u)\,dt\,du.\end{aligned}$$

Since

$$\frac{h(t)}{G(t)} = \frac{g(t)}{G(t)^2} = \frac{\partial}{\partial t}\left(\frac{1}{G(t)}\right),$$

we get

$$\begin{aligned}\tilde{\bar{b}}(s) &= \int_{u<s} b(u)\left[\frac{1}{G(u)} - 1\right]g(u)\,du + \int_{u\geq s} b(u)g(u)\,du\left[\frac{1}{G(s)} - 1\right]\\ &= \tilde{b}(s) + \bar{b}(s) - Eb,\end{aligned}$$

which is (2.3c). □

3. Discrete case. Results (1.6) and (1.8) require minor modifications when T is discrete. This section discusses the discrete case. A probabilistic proof of the main result is given, emphasizing the connection of (1.6) and (1.8) [now (3.8) and (3.16)] with familiar martingale arguments from survival analysis. Some aspects of the structure of the linear transformations A and B are seen more clearly in the discrete case.

Suppose then that T can take on N possible values,

$$\text{Prob}_\theta\{T = i\} = g_{\theta, i}, \qquad i = 1, 2, \ldots, N, \tag{3.1}$$

$\sum_{i=1}^N g_{\theta, i} = 1$. The parameter θ will usually be omitted from the notation, $g_i \equiv g_{\theta, i}$. The right cdf

$$G_i \equiv G_{\theta, i} = \sum_{j \geq i} g_{\theta, i} \tag{3.2}$$

is the probability $T \geq i$. The ith discrete hazard is

$$h_i \equiv h_{\theta, i} = \frac{g_{\theta, i}}{G_{\theta, i}}. \tag{3.3}$$

Let $T_1, T_2, \ldots, T_n$ represent an independent and identically distributed (iid) sample from the discrete distribution (3.1), and let

$$s_i = \#\{T_l = i\}, \qquad i = 1, 2, \ldots, N, \tag{3.4}$$

be the number of counts in category i. Also define

$$n_i \equiv \sum_{j \geq i} s_j, \qquad i = 1, 2, \ldots, N, \tag{3.5}$$

so $n_i = \#\{T_l \geq i\}$.

We then have the following elementary result.

LEMMA 2 (Left-to-right identity). *The probability of observing counts* $\mathbf{s} = (s_1, s_2, \ldots, s_N)$ *is*

$$\left(\frac{n!}{s_1! s_2! \ldots s_N!}\right) \prod_{i=1}^N g_i^{s_i} = \prod_{i=1}^N \binom{n_i}{s_i} h_i^{s_i} (1 - h_i)^{n_i - s_i}. \tag{3.6}$$

PROOF. The left side is the usual multinomial expression. The right side is obtained by successive conditioning beginning at the left end of the time scale: If n_i of the life-times $T_1, T_2, \ldots, T_n$ are known to exceed $i - 1$, then the number s_i dying at time i is conditionally binomial,

$$s_i | \mathbf{s}_{i-1} \sim s_i | n_i \sim \text{Bi}(n_i, h_i). \tag{3.7}$$

Multiplying the successive binomial densities (3.7) gives the right side of (3.6). □

Barlow, Bartholomew, Bremner and Brunk (1972) give a likelihood-based derivation of Lemma 2 on pages 104 and 105.

The left-to-right identity leads directly to a discrete version of (1.6),

$$\mathscr{I}_\theta \equiv \sum_{i=1}^N \left(\frac{\dot{g}_i}{g_i}\right)^2 g_i = \sum_{i=1}^N \left(\frac{\dot{h}_i}{h_i}\right)^2 \frac{g_i}{1 - h_i}. \tag{3.8}$$

The continuous case can be thought of as the discrete case with $h_i \to 0$, in which case (3.8) $\to$ (1.6). Here are the steps from (3.6) to (3.8):

If $\dot{l}_i$ indicates the score function (derivative of the log density with respect to θ) of the ith term on the right side of (3.6), then

$$\dot{l}_i = \frac{s_i - n_i h_i}{1 - h_i}\left(\frac{\dot{h}_i}{h_i}\right). \tag{3.9}$$

Given n_i, $\dot{l}_i$ has conditional mean 0 and variance $n_i h_i (1 - h_i)^{-1}(\dot{h}_i/h_i)^2$, according to (3.7).

Let $v(\mathbf{n})$ indicate the "total conditional variance"

$$v(\mathbf{n}) = \sum_{i=1}^{N} \operatorname{var}\left(\dot{l}_i | n_i\right) = \sum_{i=1}^{N} n_i h_i (1 - h_i)^{-1}\left(\dot{h}_i/h_i\right)^2. \tag{3.10}$$

Since $En_i = nG_i$, the unconditional expectation of $v(\mathbf{n})$ is

$$E\{v(\mathbf{n})\} = n \sum_{i=1}^{N} G_i h_i (1 - h_i)^{-1}\left(\frac{\dot{h}_i}{h_i}\right)^2 = n \sum_{i=1}^{N} \frac{g_i}{1 - h_i}\left(\frac{\dot{h}_i}{h_i}\right)^2. \tag{3.11}$$

$\sum_{i=1}^{N} \dot{l}_i$ is the derivative of the log density of the entire sample $\mathbf{s}$ with respect to θ.

The Fisher information

$$n\mathcal{I}_\theta = \operatorname{var}\left\{\sum_{i=1}^{N} \dot{l}_i\right\} = n \sum_{i=1}^{N} g_i \left(\frac{\dot{g}_i}{g_i}\right)^2$$

equals $E\{v(\mathbf{n})\}$, by a standard left-to-right martingale conditioning argument. [This point is explained below, in the derivation of the discrete analog (3.16) of (1.8).]

Equating (3.11) with $n\mathcal{I}_\theta$ gives result (3.8). □

The discrete version (3.16) of the general result (1.8) requires one more lemma. Let

$$D_i \equiv s_i - ng_i \quad \text{and} \quad d_i \equiv s_i - n_i h_i, \tag{3.12}$$

so D_i is the deviation of the count s_i, (3.4), from its *unconditional* expectation ng_i; and d_i is the deviation of s_i from its *conditional* expectation $n_i h_i$, (3.7). (Notice that $\sum D_i = 0$ and $d_N = 0$.) Our results follow from a lemma relating the D_i to the d_i.

LEMMA 3 (Deviations lemma). *For any vector* $\mathbf{b} = (b_1, b_2, \ldots, b_N)$*, we have*

$$\sum_{i=1}^{N} D_i b_i = \sum_{i=1}^{N} d_i a_i, \tag{3.13}$$

where $\mathbf{a} = (a_1, a_2, \dots, a_{N-1}, a_N)$ *is given by*

$$a_i \equiv b_i - \bar{b}_i \qquad \left[\bar{b}_i \equiv \frac{1}{G_{i+1}} \sum_{j \geq i+1} g_j b_j\right], \tag{3.14}$$

for $i = 1, 2, \dots, N-1$, *while* a_N *is arbitrary. Moreover, letting* $Eb \equiv \Sigma_{i=1}^N g_i b_i$, *the inverse transformation from* $\mathbf{a}$ *to* $\mathbf{b}$ *is given by*

$$b_i - Eb = a_i - \tilde{a}_i \qquad \left[\tilde{a}_i \equiv \sum_{j \leq i} h_j a_j\right], \tag{3.15}$$

for $i = 1, 2, \dots, N$.

(The indeterminacies could be removed, e.g., by decreeing that $a_N = 0 = Eb$. The proof of the deviations lemma appears later in this section.)

THEOREM. *For vectors* $\mathbf{b}$ *and* $\mathbf{a}$ *related as in* (3.14) *and* (3.15),

$$\sum_{i=1}^{N} (b_i - Eb)^2 g_i = \sum_{i=1}^{N} a_i^2 g_i (1 - h_i). \tag{3.16}$$

PROOF. The variance of the left side of (3.13) is

$$\operatorname{var}\left\{\sum_{i=1}^{N} b_i s_i\right\} = n \sum_{i=1}^{N} g_i (b_i - Eb)^2, \tag{3.17}$$

according to standard multinomial calculations. Let $\mathbf{s}_i \equiv (s_1, s_2, \dots, s_i)$, and turn now to the right side of (3.13). From (3.7) $s_i | \mathbf{s}_{i-1} \sim \operatorname{Bi}(n_i, h_i)$, so that $\{d_i a_i\}$ are martingale differences and hence conditionally uncorrelated. Since $En_i h_i = nG_i h_i = ng_i$, it follows that

$$\begin{aligned} \operatorname{var}(\Sigma a_i d_i) &= E \Sigma \operatorname{var}(a_i d_i | \mathbf{s}_i) \\ &= \Sigma E n_i h_i (1 - h_i) a_i^2 \\ &= n \Sigma g_i (1 - h_i) a_i^2. \end{aligned}$$

□

The theorem is a discrete version of the length-preserving identity (1.8), which we saw was a generalization of our main result (1.6). Let b_T indicate the discrete variate taking value b_i with probability g_i, and likewise a_T, h_T, etc. Then the theorem can be written as

$$\begin{aligned} \operatorname{var}\{b_T\} &= E\{[1 - h_T] a_T^2\} = E\{[1 - h_T][b_T - \bar{b}_T]^2\} \\ &= E\left\{\frac{1}{1 - h_T}[b_T - \bar{b}_{T-1}]^2\right\}. \end{aligned} \tag{3.18}$$

The last form of (3.18), which follows from the identity

$$a_i = b_i - \bar{b}_i = \frac{b_i - \bar{b}_{i-1}}{1 - h_i},$$

is exactly (1.8), except for a correction factor $1/(1 - h_T)$ necessary in the discrete case. [The Fisher information result (3.8) uses $b_i = \dot{g}_i/g_i$, $a_i = (1 - h_i)^{-1}(\dot{h}_i/h_i)$.]

Proof of deviations lemma. If formula (3.13) is true for $n = 1$ it is true in general, by additivity. Therefore, it is enough to prove (3.13) for $n = 1$, in which case only a single life-time $T \in \{1, \dots, N\}$ is chosen. The identity becomes

$$b_T - \Sigma b_i g_i = a_T - \sum_{i \le T} h_i a_i, \tag{3.19}$$

or $b_T - Eb = a_T - \tilde{a}_T$ which is just the definition (3.15) of $\mathbf{b}$ in terms of $\mathbf{a}$. Substituting for $\{a_i\}$ its definition (3.14) in terms of $\{b_i\}$, the difference between the left and right sides of (3.19) becomes

$$b_T - Eb - \left(b_T - \bar{b}_T\right) + \sum_{i \le T} h_i\left(b_i - \bar{b}_i\right). \tag{3.20}$$

An elementary calculation gives

$$h_i\left(b_i - \bar{b}_i\right) = -\left(\bar{b}_i - \bar{b}_{i-1}\right),$$

showing that (3.20) equals 0, so (3.13) and (3.14) are true. This shows the transformation A: $\{b_i\}_{i=1}^N \to \{a_i = b_i - \bar{b}_i\}_{i=1}^{N-1}$ and B: $\{a_i\}_{i=1}^{N-1} \to \{b_i = a_i - \tilde{a}_i\}_{i=1}^N$ are inverse in the sense that $AB\mathbf{a} = \mathbf{a}$ and $BA\mathbf{b} = \mathbf{b} - Eb$, completing the proof. □

We have so far supposed the n observations to be identically distributed. Now suppose each observation has its own fixed covariate vector z_i, $i = 1, \dots, n$. Let $s_i = \Sigma_{j=1}^n z_j I(T_j = i)$, where $I(T_j = i)$ equals 1 or 0 as T_j does or does not equal i; s_i is the sum of the z_j assigned to index i. A simple generalization of the deviations lemma is obtained by multiplying (3.19) by z_j (with T replaced by T_j) and summing over j.

Lemma 4. *Let $e_i = s_i - h_i\Sigma_{j \ge i} s_j$ and $E_i = s_i - g_i\Sigma_{j=1}^n z_j$. Then for sequences* $\mathbf{a}$ *and* $\mathbf{b}$ *related as in* (3.14) *and* (3.15), *we have*

$$\sum_{i=1}^N E_i b_i = \sum_{i=1}^N e_i a_i. \tag{3.21}$$

Lemma 3 *is the special case of Lemma* 4 *where $z_1 = z_2 = \cdots = z_n = 1$.*

We illustrate Lemma 4 using hypothesis tests in the proportional hazards model $h_\theta(t, z_i) = e^{\theta' z_i} h_0(t)$ and an example derived from Cox and Oakes (1984), page 98. For brevity, assume that there are no ties in the observed life-time variables T_i. Hence $N = n$ and $s_i = z_{j(i)}$, where $j(i)$ is the individual with $T_j = i$.

Let g_i be the empirical distribution putting probability $g_i = 1/n$ on the n distinct outcomes, and let a_j equal the censoring indicator δ_j [$\delta_j = 0$ or 1 as the observation (z_j, T_j) is or is not censored]. Let $\mathscr{D}$ be the set of exact failure times and $\mathscr{R}_i = \{j\colon T_j \geq i\}$ the risk set just before i. The size of $\mathscr{R}_i$ is $n_i = n - i + 1$. Then (3.21) becomes

$$\sum_{i=1}^{n} z_{j(i)} b_i = \sum_{i \in \mathscr{D}} \left(z_{j(i)} - \frac{1}{n_i} \sum_{j \in \mathscr{R}_i} z_j \right) = U, \tag{3.22}$$

say, where $b_i = a_i - \tilde{a}_i = \delta_i - \Sigma_{j \leq i,\, j \in \mathscr{D}} 1/n_j$. (Recall that $\Sigma b_i = 0$.)

The right side of (3.22), corresponding to conditional deviations, is the usual score statistic for testing the null hypothesis $\theta = 0$. In the two-sample case, where $z_i = 0$ or 1 according as i belongs to group 0 or 1, U reduces to the *log-rank* or *Mantel–Haenszel* statistic.

If the censoring pattern is independent of the covariates, a permutation test of H_0 is possible. The mean and variance (and other aspects) of the null-hypothesis permutation distribution can be read off from identity (3.22): For example, $EU = 0$ and $\mathrm{Cov}(U) = (n-1)^{-1}\Sigma(z_i - \bar{z})(z_i - \bar{z})^t \Sigma b_i^2$.

For the two-sample problem, Prentice and Marek (1979), following Prentice (1978), use a version of (3.22) to define and interpret a class of censored data rank tests beginning with interesting sets of coefficients b_i (considered only at exact failure times).

Andersen, Borgan, Gill and Keiding (1982) discuss the connection of the Prentice–Marek statistics with the counting process approach to censored survival data. The calculations in their Section 3.4 are closely related to those leading to Lemma 4.

4. Extensions and remarks. We have taken two distinct approaches to interpreting identities (1.6) and (1.8). An analytic viewpoint was illustrated in the continuous case in Section 2, while a probabilistic approach was presented for the discrete case in Section 3. Although important simplifications justify the separate discussion of discrete and continuous cases, we summarize here for completeness the direct generalization (of both approaches) to general distributions $F_t = P(T \leq t)$ on $\mathbb{R}$.

Write $G_t = P(T \geq t)$ and $G_t^+ = P(T > t)$, respectively, for the left and right continuous versions of the survivor function, and

$$H_t = \int_{(-\infty, t]} \frac{dF_s}{G_s}, \tag{4.1}$$

for the cumulative hazard up to time t. The definitions (2.1) become

$$\bar{b}(s) = \frac{1}{G_s^+} \int_{(s, \infty)} b(t)\, dF_t, \qquad \tilde{a}(t) = \int_{(-\infty, t]} a(s)\, dH_s. \tag{4.2}$$

(We take $0/0 = 0$.) For Lemma 1, we replace the two inner products in (2.3a), respectively, by

$$\langle b, b' \rangle_{ld} = \int b(t) b'(t)\, dF_t, \qquad \langle a, a' \rangle_{lh} = \int a(s) a'(s)(1 - \Delta H(s))\, dF_s,$$

where $\Delta H(s) = H(s) - H(s-)$ measures jumps in the cumulative hazard. [The subscripts "ld" and "lh" identify Fisher's information metric expressed respectively in log-density and log-hazard coordinates at the distribution F: For more details in the discrete case see Remark O of Efron and Johnstone (1987).] Identities (1.8) and (3.16) become

$$\int (b(s) - Eb)^2\, dF_s = \int a^2(s)(1 - \Delta H(s))\, dF_s. \tag{4.3}$$

The probabilistic approach rests on the methods of counting processes [see, e.g., surveys by Andersen, Borgan, Gill and Keiding (1982) and Andersen and Borgan (1985)]. For iid observations $T_1, \dots, T_n$ from F, the number of "deaths" just after t is $N_t = \#\{T_i \le t\}$ and the number at risk just before t is $Y_t = \#\{T_i \ge t\}$. Corresponding to the earlier D_i and d_i, we now track unconditional and conditional deviations via the cumulative processes

$$U_t = N_t - nF_t, \qquad M_s = N_s - \int_{(-\infty, s]} Y_t\, dH_t. \tag{4.4}$$

The zero mean process U_t is tied down to 0 at $\pm\infty$, while M_s is the fundamental martingale of counting process theory [Aalen (1975, 1978) and Gill (1980)]. The deviations lemma states

$$\int b_t\, dU_t = \int a_s\, dM_s, \tag{4.5}$$

for functions b, a related by either $a = b - \bar{b}$ or $b = a - \tilde{a}$. The basic identity (4.3) is obtained by taking variances in the deviations lemma, using the expectation of the predictable variance process of M_s on the right side.

The discrete information identity (3.8) is proved by decomposing the multinomial experiment into a sequence of binomial subexperiments and adding the expected conditional information in each subexperiment given its predecessors. The referees note that this approach extends to more general dominated families $\{F_\theta\} \ll \nu$, where now $F_\theta(dt) = g_\theta(t)\, d\nu(t)$. For $n = 1$, if $T \ge t$, calculate the information (conditional on the strict past) in the Bernoulli experiment from t to $t + dt$ "observe $T \in dt$ with probability $dH_\theta(t) = h_\theta(t)\, d\nu$ and $T \notin dt$ with probability $1 - dH_\theta(t)$." Adding the infinitesimal experiments gives an extension of (1.6) and (3.8):

$$\mathscr{I}_\theta = \int \left(\frac{\dot{h}_\theta}{h_\theta} \right)^2 \frac{h_\theta G_\theta\, d\nu}{1 - \Delta H_\theta} = \int \left(\frac{\dot{h}_\theta}{h_\theta} \right)^2 \frac{f_\theta\, d\nu}{1 - \Delta H_\theta}.$$

Basawa and Rao (1980) give references to the large literature on such conditional variance calculations in inference for stochastic processes. Related material leading to the identity (1.6) using information processes and locally asymptotically normal experiments appears in Jacod (1989) and Greenwood and Wefelmeyer (1989), respectively.

We will not discuss asymptotics here, save to remark that in the case where F is uniform on $[0, 1]$, the relation (4.5) goes over, as $n \to \infty$, to

$$\int b\, dW_t^0 = \int a\, dW_t \quad (\text{equality in } \textit{distribution}),$$

where W^0 and W are, respectively, standard Brownian bridge and standard Brownian motion on $[0, 1]$. This identity may be derived directly using the Doob–Meyer decomposition of Brownian bridge [see, e.g., Khamaladze (1989) for suggestive related material]. A standard tool for results of this sort are the limit theorems of Rebolledo (1978, 1980).

REMARK A (Regularity conditions). Equality (1.6) and its analog for $\theta \in \mathbb{R}^p$ hold if the score function exists in $L^2(g_\theta(t)\, dt)$ and the interchange

$$\int_t^\infty \frac{\partial}{\partial \theta} g_\theta(s)\, ds = \frac{\partial}{\partial \theta} \int_t^\infty g_\theta(s)\, ds$$

in (1.10) is valid. A standard set of sufficient conditions [cf. Ibragimov and Has'minskii (1981), pages 65 and 67] is that on a neighborhood of θ, the function $\phi \to \sqrt{g_\phi(t)}$ be continuous for Lebesgue almost all t and continuously root-mean-square differentiable in $L^2(dt)$.

REMARK B (Multiparameter case). As remarked in Section 1, the extension of (1.6) to vector-valued $\boldsymbol{\theta}$ is trivial. By contrast extension to multidimensional $\mathbf{t}$ is not straightforward (except in the limited sense of Remark C) because the total ordering is lost. Perhaps the most obvious choice for the multivariate failure rate of a density $g(t_1, t_2)$, namely $r(t_1, t_2) = g(t_1, t_2)/P\{T_1 > t_1, T_2 > t_2\}$, fails because $r(t_1, t_2)$ does not even uniquely determine $g(t_1, t_2)$ [see, e.g., Puri and Rubin (1974)]. An alternative is to consider vector-valued analogs of the hazard, such as the "hazard gradient" $\nabla_t \log P(\mathbf{T} > \mathbf{t})$, discussed for example by Marshall (1975).

REMARK C. Identity (1.8) is written for functions of a *real-valued* continuous variate T. There is a simple extension for the case when one wishes to make explicit the dependence of $T = T(x)$ on an abstract sample point $x \in \mathscr{X}$. Alternatively, we may think of $T(x)$ as imposing a (semi) ordering on $\mathscr{X}$. If $R(X)$ is a random variable with finite variance, if the density of $T(X)$ is absolutely continuous, and if $\overline{R}(t) = E[R(X) | T(X) \geq t]$, then

$$\operatorname{Var} R(X) = E\left[R(X) - \overline{R}(X)\right]^2,$$

where, by an abuse of notation, $\overline{R}(X) = \overline{R}(T(X))$. This equation is verified by considering the conditional mean $\tilde{R}(t) = E[R | T = t]$, noting that $\overline{R}(t) =$

$E[\tilde{R}|T \geq t]$; and applying (1.8) to $\tilde{R}(T)$:

$$\begin{aligned}\operatorname{Var} R &= E \operatorname{Var}(R|T) + \operatorname{Var} \tilde{R}(T) \\ &= EE\left[(R - \tilde{R})^2 \middle| T\right] + E\left[\tilde{R}(T) - \overline{R}(T)\right]^2 \\ &= E(R - \overline{R})^2.\end{aligned}$$

REMARK D. The left-to-right identity yields other martingale results. For any value of ϕ,

$$R_i \equiv e^{\phi a_i(s_i - n_i h_i) - n_i k_i}, \qquad k_i \equiv \log\left\{(1 - h_i)e^{-\phi a_i h_i} + h_i e^{\phi a_i(1 - h_i)}\right\},$$

has conditional expectation 1 under (3.7), so $R_1, R_1R_2, \dots, \prod_1^N R_i$ is an exponential martingale with expectation 1. Evaluating $E\{\prod_1^N R_i\} = 1$ from the unconditional multinomial distribution of $\mathbf{s}$ gives the following identity:

$$(4.6) \qquad 1 = \sum_{i=1}^{N} g_i e^{\phi b_i - K_i}, \qquad K_i \equiv \sum_{j \leq i} k_j.$$

Differentiating (4.6) twice with respect to ϕ is another way to derive the theorem (3.16). [Note: This exponential martingale should be distinguished from the Doléans martingale $\mathscr{E}(M)_t = \prod_{s \leq t}(1 + \Delta M_s)\exp(M_t^c)$ [see, e.g., Rogers and Williams (1987)] which would yield here $\prod_{j \leq i}[1 + \phi a_j(s_j - n_j h_j)]$.]

REMARK E. The total conditional variance $v(\mathbf{n})$, (3.10), is closely related to Greenwood's formula for the variance of an estimated survival curve. The deviations lemma leads at least in the uncensored case to a simple derivation of Greenwood's formula that avoids the usual appeal to Taylor series approximations and the "delta method." The "life-table" or "actuarial" estimate for $G_i = \prod_{j<i}(1 - h_j)$ is $\hat{G}_i = \prod_{j<i}(1 - \hat{h}_j) = n_i/n$, where as before $\hat{h}_j = s_j/n_j$. By introducing $b_j = n^{-1}I\{j \geq i\} - n^{-1}G_i$, we find

$$\hat{G}_i - G_i = \sum_j D_j b_j, \qquad D_j = s_j - n g_j.$$

Apply the deviations lemma (with $a_j = -G_i I\{j < i\}/nG_{j+1}$):

$$(4.7) \qquad \hat{G}_i - G_i = -G_i \sum_{j<i} \frac{s_j - n_j h_j}{nG_j(1 - h_j)}.$$

By taking conditional variances

$$(4.8) \qquad v_i(\mathbf{n}) = G_i^2 \sum_{j<i} \frac{n_j h_j}{(nG_j)^2(1 - h_j)}.$$

Substituting $\hat{h}_j$ and $\hat{G}_j$ for the unknowns in (4.8) gives Greenwood's formula

$$\widehat{\operatorname{var}}(\hat{G}_i) = \hat{G}_i^2 \sum_{j<i} \frac{s_j}{n_j(n_j - s_j)},$$

Miller (1981), page 45. Formula (4.7) is a special case of the stochastic integral

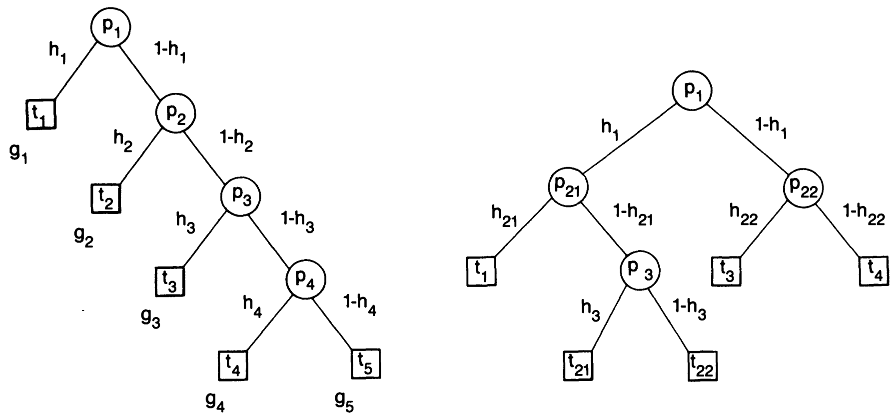

FIG. 1. *Random walks down a binary tree. Each item begins its walk at the top parent node* p_1. *At each successive parent node* p_j *(circles), the item descends left or right with probability* h_j *or* $1 - h_j$, *until it finally arrives at a terminal node* t_i *(squares). The marginal probabilities of the terminal nodes are* $g_1, g_2, g_3, \dots$. *The left diagram is the tree representation of a five-cell multinomial distribution. The right diagram shows a more complicated tree structure.*

identity $\hat{G} - G = -G\int(\hat{G}_-/G)\,dM/Y$ [in the notation of (4.4) with $\hat{G}$ the Kaplan–Meier estimate] derived by Aalen and Johansen (1978), Gill (1980) and via product integrals in Gill and Johansen (1987) and applied to Greenwood's formula.

REMARK F. The discrete situation discussed in Section 3, can be thought of as a random walk down a particularly simple binary tree shown on the left side of Figure 1. Results similar to the deviations lemma hold for arbitrarily complicated binary trees. Such trees could model survival distributions in experiments where the group of experimental subjects is subdivided (perhaps repeatedly and perhaps according to treatment or covariates) as observation proceeds over time.

Suppose that n items independently walk down a binary tree, according to the probability mechanism described in the caption of Figure 1. Let

$$
\begin{aligned}
s_i &= \#\{\text{items ending at terminal node } t_i\},\\
n_j &= \#\{\text{items passing through parent node } p_j\},\\
S_j &= \#\{\text{of the } n_j \text{ items that descend left from node } p_j\},\\
D_i &= s_i - ng_i \quad \text{and} \quad d_j = S_j - n_j h_j.
\end{aligned}
\tag{4.9}
$$

Notice that $S_j|n_j \sim Bi(n_j h_j)$ as in (3.7). A generalization of the deviations lemma applies to binary trees,

$$
\sum_i D_i b_i = \sum_j d_j a_j, \tag{4.10}
$$

the sums being over all terminal nodes i and parent nodes j, respectively, with the vectors **b** and **a** related as follows:

$$(4.11)\qquad \begin{aligned} a_j &= \sum_{i(<L)j} g_i b_i \Big/ \sum_{i(<L)j} g_i - \sum_{i(<R)j} g_i b_i \Big/ \sum_{i(<R)j} g_i,\\ b_i &= \sum_{j(>L)i} (1-h_j)a_j - \sum_{j(>R)i} h_j a_j. \end{aligned}$$

The notation $i(<L)j$ indicates terminal nodes t_i that can be reached by descending left from parent node p_j; likewise $j(>L)i$ indicates parent nodes p_j such that a left descent can lead to t_i, etc.

Results (4.10) and (4.11), which will not be proved here, lead to the following variance identity:

$$(4.12)\qquad \sum_i g_i b_i^2 = \sum_j G_j h_j (1-h_j) a_j^2, \qquad G_j \equiv \sum_{i(<)j} g_i,$$

analogous to the theorem, (3.16).

The probability $G_j = \sum_{i(<)j} g_i$ of all terminal nodes descended from the parent j can be estimated by $\hat{G}_j$ in the obvious way. A generalization of Greenwood's formula (that applies even when the random walks are censored) is

$$\widehat{\mathrm{Var}}\left\{\log \frac{\hat{G}_j}{G_j}\right\} = \sum_{j'<j} \frac{1}{n_{j'}} \left[\frac{s_{j'}}{n_{j'} - s_{j'}}\right]^{\delta_{j'}},$$

where $\delta_{j'}$ is -1 or $+1$ according as the left or right path through node j' is taken on the way to node j.

The covariates version of the deviations lemma (Lemma 4) generalizes in an obvious way and can be exploited for permutation tests and construction of a Prentice–Marek class of statistics as described after (3.22).

A further virtue of the binary (and general) tree viewpoint is that many complex probability models can be quite simply represented in terms of random walks down trees. A trivial example occurs in Figure 1(b) if we replace parent node p_3 by a terminal node t_2. Adding the constraint $h_2 = h_{21} = h_{22}$ produces the independence model for a 2×2 contingency table. The standard MLEs of the terminal cell probabilities are obtained from $\hat{h}_1 = S_1/n_1$, $\hat{h}_2 = (S_{21} + S_{22})/(n_{21} + n_{22})$. In general, closed-form maximum likelihood estimates of cell probabilities are easily obtained, even under more general equality restrictions on the transition probabilities h_i. The approach can be extended to include the class of "decomposable models" introduced by Goodman (1970) and Haberman (1974) for complete multiway contingency tables.

5. A geometric view. We have seen that the hazard rate identity (1.6) is a special case of the isometry relation (1.8) in which $b = \dot{g}_\theta / g_\theta$, $a = \dot{h}_\theta / h_\theta$. There is, however, a more fundamental connection between the linear transformations A and B and the nonlinear hazard transform $\mathbf{H}$: $g(t) \to g(t)/G(t)$ defined on the set of probability density functions on $\mathbb{R}$. First, a heuristic account.

We fix a density g and linearize $\mathbf{H}$ about g. Suppose that b has mean 0, $\int_{-\infty}^{\infty} b(t)g(t)\,dt = 0$, and that for sufficiently small ε, $g \cdot (1 + \varepsilon b)$ is a probability density. [If $\{g_\theta\}$ is a one-parameter family of densities, then $b = \dot{g}_{\theta_0}/g_{\theta_0}$ is a particularly interesting choice for b.] A simple calculation shows that

$$\mathbf{H}[g \cdot (1 + \varepsilon b)] = \mathbf{H}(g) \cdot [1 + \varepsilon A b] + o(\varepsilon), \tag{5.1}$$

where $Ab = b - \bar{b}$ as in (2.2). In other words, A is a logarithmic Gateaux derivative [e.g., Huber (1981), Chapter 2] of $\mathbf{H}$ about the fixed density g.

Now let $\mathbf{G}$ indicate the inverse of $\mathbf{H}$, mapping a hazard rate h into the corresponding density g. Another simple calculation shows that

$$\mathbf{G}[h \cdot (1 + \varepsilon a)] = \mathbf{G}(h) \cdot [1 + \varepsilon B a] + o(\varepsilon), \tag{5.2}$$

for $\varepsilon \to 0$, where $a(t)$ is any function for which $h \cdot (1 + \varepsilon a)$ is a hazard rate when ε is sufficiently small. In other words, B of (2.2) is a logarithmic Gateaux derivative of $\mathbf{G} = \mathbf{H}^{-1}$.

These expansions indicate that the mutual invertibility of A and B follows from the mutual invertibility of $\mathbf{G}$ and $\mathbf{H}$ (inverse function theorem). Furthermore, B and A are also adjoints, and hence length preserving. Thus, roughly speaking, the spaces of log densities and log hazards are not merely diffeomorphic, but also isometric. The Riemannian distance between two log densities is not changed under transformation to log hazards, even though the transforms are far from rigid.

In the remainder of this section we attempt a more formal geometric description of these phenomena. The (log) hazard function representation of a probability measure provides an alternative set of "coordinates" on the space of density functions. The hazard transformation applies to any density function; being thus nonparametric, it suggests an infinite-dimensional treatment. The operators A and B assume the role played by Jacobian matrices for change of coordinates in finite-dimensional situations.

Let $\mathscr{Q}$ be the set of finite (nonnegative) measures on $(\mathscr{X}, \mathscr{B})$ equivalent to a σ-finite measure μ. We assume that $\mathscr{X} \subset \mathbb{R}$. The map $Q \to 2(dQ/d\mu)^{1/2}$ identifies $\mathscr{Q}$ with the subset $\tilde{\mathscr{U}} = \{u\colon u > 0 \text{ a.e. } (\mu)\}$ of the Hilbert space $L_2(\mu)$. We recall some facts collected by Dawid (1975, 1977) (see also the references mentioned therein). The induced metric on $\mathscr{Q}$ is Hellinger distance

$$\rho^2(Q, Q') = 4\int \left[\left(\frac{dQ}{d\mu}\right)^{1/2} - \left(\frac{dQ'}{d\mu}\right)^{1/2}\right]^2 d\mu = \int (u - u')^2\, d\mu.$$

If $\theta \to Q_\theta$ is a smooth one-parameter curve in $\mathscr{Q}$, then $\rho(Q_\theta, Q_{\theta + d\theta}) = i_\theta^{1/2}\, d\theta$, where

$$i_\theta = 4\int \left[\partial\sqrt{g_\theta}\big/\partial\theta\right]^2 d\mu = \int \left[\partial(\log g_\theta)/\partial\theta\right]^2 g_\theta\, d\mu, \qquad g_\theta = \frac{dQ_\theta}{d\mu}.$$

Further the subset $\mathscr{P}$ of $\mathscr{Q}$ consisting of *probability* measures corresponds to a subset of a sphere: $\mathscr{U} = \{u \in L_2(\mu)\colon \|u\| = 2,\ u > 0 \text{ a.e. } (\mu)\}$.

We ignore technical difficulties (Remark G) and think of $\tilde{\mathscr{U}}, \mathscr{U}$ and the sets $\mathbf{F}\mathscr{U}$ to be defined below as manifolds modeled on Hilbert space [see Lang (1972) for infinite-dimensional manifolds].

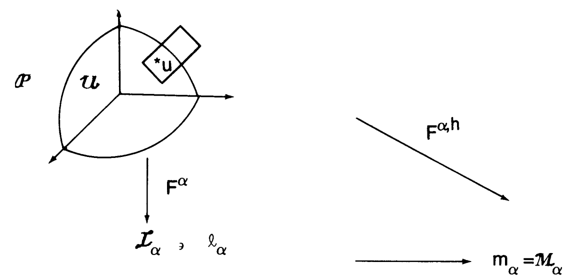

FIG. 2(a). *The $\mathscr{L}_\alpha$ and $\mathscr{M}_\alpha$ representations are defined by mappings* $\mathbf{F}^\alpha$ *of* (5.3) *and* $\mathbf{F}^{\alpha,h}$ *of* (5.4) *from the space of probability measures regarded as a subset of the sphere in* $L_2(\mu)$.

For a manifold $\mathscr{N}$ embedded in $L_2(\mu)$, define the tangent space $T_p\mathscr{N}$ as the collection of all $L_2(\mu)$ functions v for which there is a curve $\varepsilon \to u(\varepsilon)$ in $\mathscr{N}$ satisfying $\|u(\varepsilon) - p - \varepsilon v\| = o(\varepsilon)$. It is easy to check that $T_u\tilde{\mathscr{U}} = L_2(\mu)$ and $T_u\mathscr{U} = \{v \in L_2(\mu)\colon \int vu\, d\mu = 0\}$.

Thus, the induced information metric for $\mathscr{P}$ is just that induced on a sphere of radius 2 in Hilbert space. The geodesic (shortest path) curve in $\mathscr{P}$ between distributions P_1 and P_2 is a great circle through P_i with distance ρ^* given by $\rho^* = 2\cos^{-1}(\int(p_1 p_2)^{1/2}\, d\mu)$, where $p_i = dP_i/d\mu$.

We now describe two families of isometric representations of $\mathscr{P}$. The first family is studied by Amari (1982, 1985) in a finite-dimensional setting, and the second involves the hazard transformation and the operators A and B. Let $\mathscr{N}$ be a subset of $L^2(\mathscr{X}, \mathscr{B}, \mu)$ and consider mappings $\mathbf{F}\colon \mathscr{N} \to \mathscr{B}(\mathscr{X})$, the class of $\mathscr{B}$-measurable real-valued functions on $\mathscr{X}$. Let $\mathscr{N}' = \mathbf{F}\mathscr{N}$. The two families are defined by [see Figure 2(a)]

$$
\text{(5.3)} \qquad \mathbf{F}^\alpha u = \begin{cases} \log\left(\dfrac{u}{2}\right)^2, & \alpha = 1, \\ \dfrac{2}{1-\alpha}\left(\dfrac{u}{2}\right)^{1-\alpha}, & \alpha \neq 1, \end{cases}
$$

$$
\text{(5.4)} \qquad \mathbf{F}^{\alpha,h} u = \begin{cases} \log\left(\dfrac{u}{2}\right)^2 - \log\displaystyle\int_\cdot^\infty \left(\dfrac{u}{2}\right)^2 d\mu, & \alpha = 1, \\ \dfrac{2}{1-\alpha}\exp\left\{\dfrac{1-\alpha}{2}\mathbf{F}^{1,h}u\right\}, & \alpha \neq 1. \end{cases}
$$

(We make the convention that $\int_t^\infty g\, d\mu = \int_{(t,\infty)} g\, d\mu$.)

Table 1 summarizes the motivating special cases and the special notation used for each. The notation highlights the special role of the log-density and log-hazard representations ($\alpha = 1$). The values $\alpha = 1, 0, -1$ are most important, but other values can occur [e.g., $\alpha = \pm 1/3$ in Kass (1984)].

TABLE 1
Important special cases of the two families of representations

α	Name	Space	Definition	Typical element	Tangent space
		$\mathscr{L}_\alpha$	$\mathbf{F}^\alpha\mathscr{U}$	l_α	$T_{l_\alpha}\mathscr{L}_\alpha = L_2^0(g^\alpha\, d\mu)$
1	Log densities	$\mathscr{L}$	$\mathbf{F}^1\mathscr{U}$	$l = \log g$	$T_l\mathscr{L} = L_2^0(g\, d\mu)$
0	Root densities	$\mathscr{P}$	$\mathbf{F}^0\mathscr{U}$	u	
−1	Densities	$\mathscr{G}$	$\mathbf{F}^{-1}\mathscr{U}$	$g = u^2/4$	
		$\mathscr{M}_\alpha$	$\mathbf{F}^{\alpha,\, h}\mathscr{U}$	m_α	$T_{m_\alpha}\mathscr{M}_\alpha = L_2(h^{\alpha-1}g\, d\mu_h)$
1	log hazards	$\mathscr{M}$	$\mathbf{F}^{1,\, h}\mathscr{U}$	m	$T_m\mathscr{M} = L_2(g\, d\mu_h)$
−1	hazards	$\mathscr{H}$	$\mathbf{F}^{-1,\, h}\mathscr{U}$	h	

Discrete cases. Formula (5.4) shows that $\mathbf{F}^{-1,\, h}u(s) = g(s)/G(s+)$. When $\Delta G(s) \neq 0$, this differs from the traditional form of the hazard rate $h^*(s) = g(s)/G(s-)$, being instead equal to $h^*(s)/(1-\Delta H(s))$. Since all densities considered are equivalent to μ, this can only occur at an atom of μ. If there is an atom at the upper limit of the support of μ, then $\mathbf{F}^{\alpha,\, h}u$ is undefined there. This is important in the finite discrete case of Section 3 in which μ is supported on N points: Then $\mathscr{M}_\alpha$ becomes an $(N-1)$-dimensional manifold.

Tangent spaces. We use the mapping $\mathbf{F}$ to carry the Hilbert manifold structure of $\mathscr{N}$ to $\mathscr{N}' = \mathbf{F}\mathscr{N}$. [Thus, $\phi = (\mathbf{F})^{-1}$ is a "chart".] Fix $u \in \mathscr{N}$ and tangent vector $v \in L_2(\mu)$; the derivative of $\mathbf{F}$ at u is the linear mapping $\mathbf{F}_{*u}$ defined as usual by $\mathbf{F}_{*u}(v) = \lim_{\varepsilon\to 0}\varepsilon^{-1}[\mathbf{F}(u+\varepsilon v) - \mathbf{F}u]$, where the limit is taken pointwise to give a function in $\mathscr{B}(\mathscr{X})$. Then the tangent space to $\mathscr{N}'$ at $p = \mathbf{F}u$ is $T_p\mathscr{N}' = \mathbf{F}_{*u}(T_u\mathscr{N})$. When $\mathbf{F}_{*u}$ is one-to-one on $T_u\mathscr{N}$, the inner product of $a, a' \in T_p\mathscr{N}'$ is defined by

$$\langle a, a'\rangle_p = \langle(\mathbf{F}_{*u})^{-1}a, (\mathbf{F}_{*u})^{-1}a'\rangle_u, \tag{5.5}$$

where $\langle\cdot,\cdot\rangle_u$ is the usual inner product of $L_2(\mu)$. Thus, the length of a curve $\theta \to p_\theta$ in $\mathscr{N}'$ in the information metric is given by $\int\langle\dot p_\theta, \dot p_\theta\rangle_{p_\theta}^{1/2}\, d\theta$, where $\dot p_\theta \in T_{p_\theta}\mathscr{N}'$. [Informally, $\dot p_\theta = (\partial/\partial\theta)p_\theta$, but more carefully, if $u_\theta = (\mathbf{F})^{-1}p_\theta$, $\dot p_\theta = \mathbf{F}_{*u_\theta}((\partial/\partial\theta)u_\theta)$.]

We can now explicitly describe the tangent spaces corresponding to the various representations. First, $\mathbf{F}^\alpha_{*u}v = (u/2)^{-\alpha}v$ and $T_{l_\alpha}\tilde{\mathscr{L}}_\alpha = L_2(g^\alpha\, d\mu)$ with the associated inner product. This is the "α-expectation" of Amari (1985). For *probability* densities the tangent space $T_{l_\alpha}\mathscr{L}_\alpha$ is the subspace of $v \in L_2(g^\alpha\, d\mu)$ for which $\langle v, l_\alpha\rangle_{l_\alpha} = 0$ when $\alpha \neq 1$ and $\langle v, 1\rangle_l = \int vg\, d\mu = 0$ when $\alpha = 1$. We denote these spaces by $L_2^0(g^\alpha\, d\mu)$.

Second, the hazard maps $\mathbf{F}^{\alpha,\, h}u$ are unchanged if u is replaced by cu, so we make them one-to-one by restriction to $\mathscr{U}$. Calculation shows

$$\mathbf{F}^{1,\, h}_{*u}v = 2A(u^{-1}v) = A(\mathbf{F}^1_{*u}v), \tag{5.6}$$

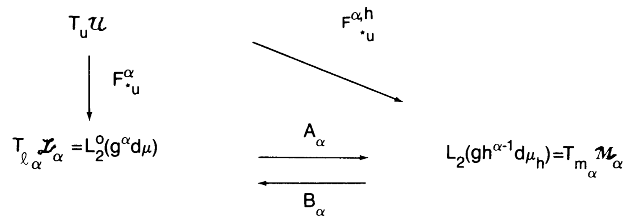

FIG. 2(b). *The tangent spaces at l_α, m_α are images of the tangent space $T_u\mathcal{U}$ at u under the derivatives $\mathbf{F}^{\alpha}_{*u}$, $\mathbf{F}^{\alpha,h}_{*u}$. Operators A and B (and their α-family generalizations) arise as derivatives of the log-hazard transform and its inverse.*

where $Ab = b - \bar{b}$ is our basic operator from (2.2). For general α, we find from (5.4) that

$$\mathbf{F}^{\alpha,h}_{*u}v = \left(\frac{1-\alpha}{2}\right)\mathbf{F}^{\alpha,h}u \cdot \mathbf{F}^{1,h}_{*u}v.$$

From Sections 2 to 4, A is an isometry of $L_2^0(g\,d\mu)$ on $L_2(g(1-\Delta H)\,d\mu)$ with inverse (and adjoint) given by the operator B of (2.2), (3.15) and (4.2). It follows that $T_m\mathcal{M} = L_2(g(1-\Delta H)\,d\mu)$ and that $\langle a, a'\rangle_m = \int aa'g(1-\Delta H)\,d\mu$ is the induced inner product. We shall write $d\mu_h = (1-\Delta H)\,d\mu$, mindful that $\mu_h = \mu$ in the continuous case.

Role of the operators A and B. The mapping from l_α to m_α in Figure 2(a) is given by $\mathbf{L}^\alpha = \mathbf{F}^{\alpha,h}\circ(\mathbf{F}^\alpha)^{-1}$. From the chain rule, the derivative of $\mathbf{L}^\alpha$ at l_α is found for $\alpha \neq 1$ to be

$$\begin{aligned}\mathbf{L}^{\alpha}_{*l_\alpha}b &= m_\alpha A\left(\frac{b}{l_\alpha}\right) \qquad b \in L_2(g^\alpha\,d\mu)\\ &= A_\alpha b.\end{aligned} \tag{5.7}$$

[Of course, $\mathbf{L}^1_{*l}b = Ab$, as shown in (5.6).] It follows that A_α is an isometry of $T_{l_\alpha}\mathcal{L}_\alpha$ on $T_{m_\alpha}\mathcal{M}_\alpha$ with inverse B_α defined in the obvious way. Figure 2(b) summarizes the tangent spaces and the isometries between them.

This section began with identities (5.1) and (5.2) suggesting that A and B were logarithmic Gateaux derivatives of $\mathbf{H}$ and $\mathbf{G}$. Relation (5.7) expresses this formally: The derivative of $\mathbf{H}$ is A_{-1}, which is related to the derivative A of $\mathbf{M}$ exactly as is suggested by (5.1). [In related work, Gill and Johansen (1987), Theorem 14, study the (compact) derivative of the map from hazard *measures* to survival (multiplicative interval) *functions*. Our mapping $\mathbf{H}$ is obtained from theirs by composition with integration and differentiation; the derivatives will be connected by the chain rule.]

To summarize, we have constructed isometric representations $\mathcal{L}_\alpha$ and $\mathcal{M}_\alpha$ of the collection of probability measures equivalent to a fixed measure μ on $\mathbb{R}$. We have described the corresponding tangent bundles (collections of tangent spaces) explicitly and shown that the inner products induced by carrying over the information metric correspond to the natural Hilbertian inner product on these tangent spaces. For the spaces $\mathscr{L}_\alpha$ this is well known; for $\mathscr{M}_\alpha$ it depends essentially on the identity (1.8) and its extensions in Sections 2 to 4.

What is the value of all these isometric representations? Dawid, Amari and others have shown that each representation $\mathscr{L}_\alpha$ suggests a different notion of straight line (affine connection) and a corresponding measure of curvature. These curvature measures are used to study loss of information and second-order efficiency of (for example) the maximum likelihood and minimum cross-entropy methods of estimation in finite-dimensional parametric models, as in Efron (1975), Efron and Hinkley (1978), Amari (1982) and much related work. [See Kass (1988) and Amari, Barndorff-Nielsen, Kass, Lauritzen and Rao (1987) for surveys.]

In the next section, we exploit the $\mathscr{M}_1$ (log-hazard) representation to derive simple formulas for statistical curvature when models are specified in terms of finite-dimensional families of hazard rates.

REMARK G. According to Bickel (1984), tangent spaces (and cones) were introduced into nonparametric statistics by Koshevnik and Levit (1976). They play an important role in the study of semiparametric models [Bickel, Klaassen, Ritov and Wellner (1989)].

A technical difficulty arises in pursuing our analogy with finite-dimensional manifold theory: The set $\mathscr{Q}$ is *not* open in the norm topology of $L_2(\mu)$, and standard forms of the inverse/implicit function theorems (used in studying submanifolds) assume smoothness in an open neighborhood of the point of interest. Note, however, that $\mathscr{Q}$ is to a first approximation open in the sense needed for *compact* differentiability: If $K \subset L_2(\mu)$ is compact and $x_0 \in \mathscr{Q}$, then

$$\sup_{k \in K} \|(x_0 + tk)_+ - x_0 - tk\| \in o(t).$$

Reeds (1976) gives relevant implicit function theorems for (and many applications of) compact differentiability.

6. Statistical curvature in terms of hazard rates. The previous section emphasized the log-hazard representation of probability distributions or models. Here we use the identities of Section 2 to compute the statistical curvature of finite-dimensional models specified in terms of hazard rates.

To keep notation simple, we focus on one-parameter families $\mathscr{F}$ of density functions $\{g_\theta(s)\}$, $\theta \in \Theta \subset \mathbb{R}$, and write $l_\theta = \log g_\theta$, $\dot{l}_\theta = \partial l_\theta / \partial\theta$, $\ddot{l}_\theta = \partial^2 l_\theta / d\theta^2$. [We assume the regularity conditions of Remark A and that $\ddot{l}_\theta$ exists in $L^2(g_\theta(t)\,dt)$.] Let M_θ denote the covariance matrix under θ of $(\dot{l}_\theta, \ddot{l}_\theta)$. Efron's (1975) curvature γ^2 is defined in the log-density representation as $\det M_\theta / \mathscr{I}_\theta^3$. Suppose now that $\mathscr{F}$ is specified in terms of the corresponding hazard rates h_θ and/or log-hazard rates $m_\theta = \log h_\theta$. Again dots denote $\partial/\partial\theta$, and the cumulative hazard $H_\theta(t) = \int_{-\infty}^t h_\theta(s)\,ds$. The next result describes how to compute curvature.

PROPOSITION 6.1. *Suppose that* $\ddot{m}_\theta$ *and* $\dot{m}_\theta^2 \in L^2(g_\theta\, dt)$ *and that* $\partial^k/\partial\theta^k \int^t h_\theta = \int^t \partial^k h_\theta/\partial\theta^k$, $k = 1, 2$. *Then*

$$E\dot{l}_\theta^2 = E\dot{m}_\theta^2, \tag{6.1a}$$

$$E\dot{l}_\theta \ddot{l}_\theta = E\dot{m}_\theta \ddot{m}_\theta + E\dot{m}_\theta^2 \dot{H}_\theta, \tag{6.1b}$$

$$E\ddot{l}_\theta^2 = E\ddot{m}_\theta^2 + 2E\dot{m}_\theta^2 \ddot{H}_\theta. \tag{6.1c}$$

PROOF. Of course, (6.1a) is just (1.6). For identity (6.1b), differentiate the relation

$$\dot{l}_\theta = \dot{m}_\theta - \int^{\cdot} \dot{m}_\theta e^{m_\theta} = B\dot{m}_\theta$$

with respect to θ to obtain

$$\ddot{l}_\theta = \ddot{m}_\theta - \int^{\cdot} \left(\ddot{m}_\theta + \dot{m}_\theta^2\right) e^{m_\theta} = B\ddot{m}_\theta - \widetilde{\dot{m}_\theta^2}, \tag{6.2}$$

where $\tilde{a}$ was defined at (2.1). Use (2.3a) and (2.6) (B is an isometry) to find

$$E\dot{l}_\theta \ddot{l}_\theta = E\dot{m}_\theta \ddot{m}_\theta - E\overline{B\dot{m}_\theta}\dot{m}_\theta^2.$$

Now from (2.3b),

$$-\overline{B\dot{m}_\theta} = -\overline{\dot{m}_\theta} + \bar{\tilde{\dot{m}}}_\theta = \tilde{\dot{m}}_\theta = \int^{\cdot} \dot{m}_\theta h_\theta = \int^{\cdot} \dot{h}_\theta = \dot{H}_\theta.$$

To get identity (6.1c), rewrite $\ddot{l}_\theta$ as $B(\ddot{m}_\theta + \dot{m}_\theta^2) - \dot{m}_\theta^2$ in (6.2). Thus,

$$E\ddot{l}_\theta^2 = E\left[B\left(\ddot{m}_\theta + \dot{m}_\theta^2\right)\right]^2 + E\dot{m}_\theta^4 - 2E\dot{m}_\theta^2 B\left(\ddot{m}_\theta + \dot{m}_\theta^2\right). \tag{6.3}$$

In the final term on the right side, expand $B(\ddot{m}_\theta + \dot{m}_\theta^2)$ as $\ddot{m}_\theta + \dot{m}_\theta^2 - (\ddot{m}_\theta + \dot{m}_\theta^2)^{\sim}$ and note that

$$\left(\ddot{m}_\theta + \dot{m}_\theta^2\right)^{\sim} = \int^{\cdot} \left(\ddot{m}_\theta + \dot{m}_\theta^2\right) e^{m_\theta} = \int^{\cdot} \partial^2/\partial\theta^2 e^{m_\theta} = \ddot{H}_\theta.$$

This and the isometry property (2.6) applied to the first term of the right side of (6.3) yield identity (6.1c). □

EXAMPLES. Clayton (1983) and Efron (1988) have studied estimation of parametric classes of continuous hazard functions of the form

$$h_\alpha(t) = \exp\left[\alpha^T x(t)\right],$$

where α is a $p \times 1$ vector of unknown parameters and $x(t)$ is an observed p-dimensional time-dependent covariate vector. Proposition 6.1 and its multivariate extension (Remark H) simplify curvature calculations for such models, as we illustrate in two particular cases.

1. Suppose the hazard rate is a known constant apart from a jump of unknown size at a known time t_0:

$$m_\theta(t) = \log h_\theta(t) = \begin{cases} c + \theta, & 0 < t \le t_0, \\ c, & t > t_0. \end{cases}$$

Since we assume c known, we can without loss set it to 0 (by rescaling time via $\tilde{t} = e^{c}t$). Then $\dot{m}_\theta = I\{t \le t_0\}$, $\ddot{m}_\theta = 0$ and $H_\theta = \phi t \wedge t_0 + (t - t_0)_+$, where $\phi = e^{\theta}$. It follows that $\mathscr{I}_\theta = 1 - e^{-a}$, where $a = H(t_0) = \phi t_0$ and $E\dot{l}_\theta^2 = 2E\dot{l}_\theta \ddot{l}_\theta = 2(\mathscr{I}_\theta - ae^{-a})$. As the following table shows, the curvature for even a single observation is very small here:

$a = e^{\theta}t_0$	0.2	0.4	0.6	0.8	1.0	2	4	6	8	10
γ^2	0.015	0.027	0.036	0.042	0.046	0.043	0.013	0.002	0.0003	< 0.00005

2. If the log hazard is assumed linear

$$m(t) = \log h(t) = \alpha_1 + \alpha_2 t, \qquad t > 0,$$

then $G(t)$ is the survival function of the Gompertz distribution [Johnson and Kotz (1970), page 271; and Read (1983)], corresponding to the random variable

$$T \sim \frac{1}{\alpha_2}\log\left(1 + \frac{Z}{\theta}\right), \qquad \theta = \frac{e^{\alpha_1}}{\alpha_2},$$

where Z is a standard one-sided exponential variate. If α_2 is assumed known, the curvature of the one-parameter family indexed by α_1 vanishes since $\beta = e^{\alpha_1}$ is the natural parameter of a one-parameter exponential family and curvature is invariant to reparametrizations of either parameter space or sample space. If α_1 is assumed known, the curvature in α_2 is found from the formulas $\dot{m}_{\alpha_2}(t) = t$, $\ddot{m}_{\alpha_2}(t) \equiv 0$, $\partial^k H/\partial^k \alpha_2 = \beta\int_0^t s^k e^{\alpha_2 s}\,ds$. The curvature is evaluated numerically, as it involves integrals $K(a, m)(\theta) = \int_0^\infty \theta e^{-\theta w}(1 + w)^{a-1}\log^m(1 + w)\,dw$ for $(a, m) = (1,2), (2,2), (2,3)$ and (2.4). Asymptotic values are used at 0 and ∞.

θ	0	0.05	0.10	0.2	0.5	1.0	2	5	10	20	∞
γ^2	0	0.062	0.082	0.11	0.15	0.19	0.24	0.35	0.48	0.64	1

REMARK H [Multivariate version of (6.1)]. Let $\{g_\theta\}$ be a smooth family of densities indexed by a multidimensional parameter $\theta = (\theta^i)$. Define $l_\theta = \log g_\theta$, $l_i = \partial l_\theta/\partial\theta_i$, $l_{ij} = \partial^2 l_\theta/\partial\theta_i\,\partial\theta_j$ and similarly for $m_\theta = \log h_\theta$ and $H_\theta = \int^{\cdot} h_\theta$ define m_i, m_{jk}, H_i, H_{jk}, etc. The method of Proposition 6.1 shows that

$$El_il_j = Em_im_j,$$
$$El_il_{jk} = Em_im_{jk} + EH_im_im_k,$$
$$El_{ij}l_{kl} = Em_{ij}m_{kl} + Em_im_jH_{kl} + EH_{ij}m_km_l.$$

Reeds (1975) and Amari (1987) give the extension of statistical (exponential) curvature to the multidimensional case, and the formulas above would be used in calculating these curvatures from a hazard function form of the model.

Acknowledgments. We thank the referees and Associate Editors for their thoughtful comments on successive revisions and references to the work of Prentice and Marek and to the counting process literature. We are grateful to Chitra Lele for help with the numerical integrals in Section 6, and to Ya'acov Ritov and Jon Wellner for sending us a preprint of their paper during the initial review of this manuscript.

REFERENCES

AALEN, O. O. (1975). Statistical inference for a family of counting processes. Ph.D. dissertation, Dept. Statist., Univ. California, Berkeley.

AALEN, O. O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6** 701–726.

AALEN, O. O. and JOHANSEN, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand. J. Statist.* **5** 141–150.

AMARI, S.-I. (1982). Differential geometry of curved exponential families—curvatures and information loss. *Ann. Statist.* **10** 357–387.

AMARI, S.-I. (1985). *Differential-Geometric Methods in Statistics. Lecture Notes in Statist.* **28**. Springer, New York.

AMARI, S.-I. (1987). Differential geometry theory of statistics—towards new developments. In *Differential Geometry in Statistical Inference*. Inst. Math. Statist., Monograph Series, Hayward, Calif.

AMARI, S.-I., BARNDORFF-NIELSEN, O. E., KASS, R. E., LAURITZEN, S. L. and RAO, C. R. (1987). *Differential Geometry in Statistical Inference*. Inst. Math. Statist., Monograph Series, Hayward, Calif.

ANDERSEN, P. K. and BORGAN, Ø. (1985). Counting process models for life history data: A Review (with discussion). *Scand. J. Statist.* **12** 97–158.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. and KEIDING, N. (1982). Linear nonparametric tests for comparison of counting processes, with applications to censored survival data (with Discussion). *Internat. Statist. Rev.* **50** 219–258.

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference under Order Restrictions*. Wiley, New York.

BASAWA, I. V. and PRAKASA RAO, B. L. S. (1980). *Statistical Inference in Stochastic Processes*. Academic, London.

BICKEL, P. J. (1984). Review of book by J. Pfanzagl. *Ann. Statist.* **12** 786–791.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1990). *Efficient and Adaptive Inference in Semiparametric Models*. Johns Hopkins Univ. Press, Baltimore.

BORGAN, Ø. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scand. J. Statist.* **11** 1–16.

CLAYTON, D. (1983). Fitting a general family of failure-time distributions using GLIM. *J. Roy. Statist. Soc. Ser. C* **32** 102–109.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.

COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

DAWID, A. P. (1975). Discussion to Efron's paper. *Ann. Statist.* **3** 1231–1234.

DAWID, A. P. (1977). Further comments on some comments on a paper by Bradley Efron. *Ann. Statist.* **5** 1249.

EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency) (with Discussion). *Ann. Statist.* **3** 1189–1242.

EFRON, B. (1977). The efficiency of Cox's likelihood function for censored data. *J. Amer. Statist. Assoc.* **72** 557–565.

EFRON, B. (1988). Logistic regression, survival analysis and the Kaplan–Meier curve. *J. Amer. Statist. Assoc.* **83** 414–425.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with Discussion). *Biometrika* **65** 457–487.

EFRON, B. and JOHNSTONE, I. (1987). Fisher's information in terms of the hazard rate. Technical Report 264, Dept. Statistics, Stanford Univ.

GILL, R. D. (1980). *Censoring and Stochastic Integrals*. Math. Centre Tracts 124. Mathematisch Centrum, Amsterdam.

GILL, R. D. and JOHANSEN, S. (1987). Product-integrals and counting processes. Preprint, Inst. Math. Statist., Univ. Copenhagen.

GOODMAN, L. A. (1970). The multivariate analysis of qualitative data: Interaction among multiple classifications. *J. Amer. Statist. Assoc.* **65** 226–256.

GREENWOOD, P. E. and WEFELMEYER, W. (1989). Efficiency bound for estimating functionals of stochastic processes. Preprint No. 117, Univ. Cologne.

HABERMAN, S. (1974). *The Analysis of Frequency Data*. IMS Monographs. Univ. Chicago Press.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York.

JACOD, J. (1989). Une application de la topologie d'Emery: Le processus information d'un modele statistique filtre. *Sem. Probab.* To appear.

JAMES, I. R. (1986). On estimating equations with censored data. *Biometrika* **73** 35–42.

JAMES, I. R. (1987). Tests for location with k samples and censored data. *Biometrika* **74** 599–607.

JOHNSON, N. and KOTZ, S. (1970). *Continuous Univariate Distributions* **2**. Wiley, New York.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

KASS, R. E. (1984). Canonical parametrizations and zero parameter-effects curvature. *J. Roy. Statist. Soc. Ser. B* **46** 86–92.

KASS, R. E. (1989). The geometry of asymptotic inference (with Discussion). *Statistical Science* **4** 188–234.

KHAMALADZE, E. V. (1989). An innovation approach to goodness of fit tests in $\mathbb{R}^m$. Preprint, Centre for Mathematical and Computer Science, Amsterdam.

KOSHEVNIK, Y. and LEVIT, B. (1976). On a non-parametric analogue of the information matrix. *Theory Probab. Appl.* **21** 738–753.

LANG, S. (1972). *Differential Manifolds*. Addison-Wesley, Reading, Mass.

LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

MARSHALL, A. W. (1975). Some comments on the hazard gradient. *Stochastic Process. Appl.* **3** 293–300.

MILLER, R. G., JR. (1981). *Survival Analysis*. Wiley, New York.

PRENTICE, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65** 167–179.

PRENTICE, R. L. and MAREK, P. (1979). A qualitative discrepancy between censored data rank tests. *Biometrics* **35** 861–867; correspondence (1980) **36** 353–355; correction (1980) **36** 751.

PURI, P. S. and RUBIN, H. (1974). On a characterization of the family of distributions with constant multivariate failure rates. *Ann. Probab.* **2** 738–740.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications* 2nd ed. Wiley, New York.

READ, C. B. (1983). Gompertz distribution. *Encyclopedia of Statistical Sciences* **3**. Wiley, New York.

REBOLLEDO, R. (1978). Sur les applications de la théorie des martingales à l'étude statistique l'une famille de processus ponctuels. *Lecture Notes in Math.* **636** 27–70. Springer, Berlin.

REBOLLEDO, R. (1980). Central limit theorems for local martingales. *Z. Wahrsch. verw. Gebiete* **51** 269–286.

REEDS, J. A. (1975). Discussion to Efron's paper. *Ann. Statist.* **3** 1234–1238.

REEDS, J. A. (1976). On the definition of von Mises functionals. Ph.D. thesis, Dept. Statist., Harvard Univ.

RITOV, Y. and WELLNER, J. A. (1988). Censoring, martingales and the Cox model. In *Proc. Conf. Statist. Inference for Stochastic Processes, Contemporary Mathematics*. Amer. Math. Soc.

ROGERS, L. C. G. and WILLIAMS, D. (1987). *Diffusions, Markov Processes and Martingales: Itô Calculus* **2**. Wiley, New York.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065