FISHER'S INFORMATION IN TERMS OF THE HAZARD RATE

BY

B. EFRON AND I. JOHNSTONE

TECHNICAL REPORT NO. 264
JANUARY 1987

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS86-00235

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
FISHER'S INFORMATION IN TERMS OF THE HAZARD RATE

B. EFRON AND I. JOHNSTONE

TECHNICAL REPORT NO. 264
JANUARY 1987

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS86-00235

ALSO PREPARED UNDER PUBLIC HEALTH SERVICE GRANT 2 RO1 GM21215-12
AND ISSUED AS TECHNICAL REPORT NO. 114, DIVISION OF BIOSTATISTICS,
STANFORD UNIVERSITY.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Fisher's Information in Terms of the Hazard Rate

B. Efron and I. Johnstone*

Abstract

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 \( \theta \) can be expressed in terms of the hazard rate as follows,

\[
\mathcal{I}_\theta \equiv \int \frac{\dot{g}(\theta)}{g(\theta)}^2 g(\theta) = \int \frac{\dot{h}(\theta)}{h(\theta)}^2 h(\theta), \quad \theta \in \mathbb{R}
\]

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, its connection with martingale theory, and its relation to statistical issues in the theory of life-time distributions and censored data.

*Research partially supported also by National Science Foundation Grant DMS 8451750.
Fisher's Information in Terms of the Hazard Rate

B. Efron and I. Johnstone

Abstract

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 \( \theta \) can be expressed in terms of the hazard rate as follows,

\[
\mathcal{J}_\theta = \int \left( \frac{\dot{g}_\theta}{g_\theta} \right)^2 g_\theta = \int \left( \frac{\dot{h}_\theta}{h_\theta} \right)^2 h_\theta , \quad \theta \in \mathbb{R}
\]

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, its connection with martingale theory, and its relation to statistical issues in the theory of life-time distributions and censored data.
Fisher's Information in Terms of the Hazard Rate

B. Efron and I. Johnstone

1. Introduction.

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

$$\mathcal{I}_\theta = \int_{-\infty}^{\infty} g_\theta(t) \left( \frac{\dot{g}_\theta(t)}{g_\theta(t)} \right)^2 \, dt,$$  \hspace{1cm} (1.1)

where the dot indicates differentiation with respect to $\theta$,

$$\dot{g}_\theta(t) = \frac{\partial}{\partial \theta} g_\theta(t).$$  \hspace{1cm} (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) or Section 2.6 of Lehmann (1983). The parameter $\theta$ may be a p-dimensional vector, in which case $\dot{g}_\theta(t) = \left( \frac{\partial g_\theta(t)}{\partial \theta_1} \ldots \frac{\partial g_\theta(t)}{\partial \theta_p} \right)'$ and $\mathcal{I}_\theta$ is the $p \times p$ Fisher information matrix. The results which follow hold for the matrix case, see Remark N, Section 5, but for the sake of simple exposition we will take $\theta$ real-valued.

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

$$G_\theta(t) = \int_{t}^{\infty} g_\theta(s) \, ds = \text{Prob}_\theta(T > t),$$  \hspace{1cm} (1.3)

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

$$h_\theta(t) = \frac{g_\theta(t)}{G_\theta(t)}.$$  \hspace{1cm} (1.4)
Hazard rates are useful in discussing life-time distributions. They have the interpretation
\[ h_{\theta}(t)\Delta = \frac{g_{\theta}(t)\Delta}{G_{\theta}(t)} = \text{Prob}\{T \in (t, t+\Delta) | T > t\} \] , 
(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:
\[ \mathcal{I}_\theta = \int_{-\infty}^{\infty} g_{\theta}(t) \left[ \frac{\dot{h}_{\theta}(t)}{h_{\theta}(t)} \right]^2 dt , \quad [\dot{h}_{\theta}(t) \equiv \frac{\partial}{\partial \theta} h_{\theta}(t)] . \] 
(1.6)
In other words one can replace the usual score function \( \frac{3}{3\theta} \log g_{\theta}(t) = \dot{g}_{\theta}(t)/g_{\theta}(t) \) in (1.1) by its hazard rate analogue \( \frac{3}{3\theta} \log h_{\theta}(t) \). [However \( \int_{-\infty}^{\infty} g_{\theta}(t) \dot{h}_{\theta}(t)/h_{\theta}(t)dt \) does not usually equal zero, while \( \int_{-\infty}^{\infty} g_{\theta}(t)[g_{\theta}(t)/h_{\theta}(t)]dt = 0 \) under mild regulatory conditions.] Formula (1.6) is simple and easy to derive, but it has interesting statistical and probabilistic connections, which are the main topic of this paper.

**Example 1.** \( g_{\theta}(t) = \frac{1}{\theta} e^{-t/\theta} \) for \( t \geq 0 \). Then \( h_{\theta}(t) = 1/\theta, \dot{h}_{\theta}(t)/h_{\theta}(t) = -1/\theta, \) and (1.6) gives \( \mathcal{I}_\theta = \int_{0}^{\infty} g_{\theta}(t) \left( \frac{1}{\theta^2} \right) dt = 1/\theta^2. \)

**Example 2.** \( g_{\theta}(t) = \phi(t-\theta), \) where \( \phi(t) = (2\pi)^{-1/2} e^{-t^2/2}. \) Then (1.2) gives \( \mathcal{I}_\theta = 1 = \int_{-\infty}^{\infty} \phi(t-\theta)(t-\theta)^2 dt. \) In this case (1.6) produces an identity involving Mills' ratio,
\[ 1 = \int_{-\infty}^{\infty} \phi(t-\theta) \left[ (t-\theta) - \frac{\phi(t-\theta)}{1-\phi(t-\theta)} \right]^2 dt , \] 
(1.7)
where \( \phi(t) \) is the standard normal cumulative \( \int_{-\infty}^{t} \phi(s)ds. \)
Equality (1.6) says that the functional transformation 
\[ \frac{\dot{g}(t)}{g(t)} \to \frac{\dot{h}(t)}{h(t)} \] preserves length in the \( L_2 \) norm (1.7) \( \|b\|^2 = \int_{-\infty}^{\infty} \frac{g_0(t)b(t)}{g(t)} dt \).

We will discuss a general class of length-preserving linear transformations on the \( L_2 \) space defined by \( g_0 \). Equality (1.6) is then seen as a special case of a more general result, holding for \( L_2 \) functions \( b(T) \) of a continuous variate \( T \):

\[ \text{Var}[b(T)] = E[b(T)-\bar{b}(T)]^2 \quad \text{where} \quad \bar{b}(t) = E[b(T) | T > t]. \quad (1.8) \]

The choice \( b(t) = \frac{\dot{g}_0(t)}{g_0(t)} \) has

\[ \bar{b}(t) = \frac{1}{g_0(t)} \int_{t}^{\infty} g_0(s) \frac{\dot{g}_0(s)}{g_0(s)} ds = \frac{\dot{g}_0(t)}{g_0(t)}, \quad (1.9) \]

so

\[ b(t)-\bar{b}(t) = \frac{\dot{g}_0(t)}{g_0(t)} - \frac{\dot{g}_0(t)}{g_0(t)} = \frac{\dot{h}_0(t)}{h_0(t)}, \quad (1.10) \]

and (1.8) gives (1.6).

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 the theory of censored data, for example in the Kaplan-Meier estimate, the Mantel-Haenszel test, and Cox's proportional hazard model. (See Miller (1986), Cox and Oakes (1984), Kalbfleisch and Prentice (1980).) Our results will be proved using the left-to-right conditioning argument. Section 4 briefly discusses an important statistical point relating to this argument. Because of the connection with censored data, the generic random variables \( T \) will be referred to as "life-times", though of course this has no bearing on the results.

In fact it is not difficult to prove (1.8) directly using integration by parts, see Remark J of Section 5. In addition to verifying (1.6), (1.8), our discussion has four purposes: (i) to more fully understand the mapping \( \frac{\dot{g}(t)}{g(t)} \to \frac{\dot{h}(t)}{h(t)} \).
and its inverse; (ii) to discuss more general length-preserving transformations; (iii) to derive the discrete analogue of (1.8), which helps clarify the relationship with martingale theory; (iv) and to understand the statistical basis of (1.6), especially its connection to familiar arguments from the theory of life-time distributions and censored data.

2. Left-to-Right Identity.

Our central result, (1.6), holds as stated only for continuous random variables. It turns out to be more informative to begin with discrete distributions, for which a small correction is required to (1.6).

Suppose then that \( T \) can take on \( N \) possible values,

\[
\text{Prob}_\theta\{T=i\} = g_{\theta,i} \quad i = 1, 2, \ldots, N ,
\]

(2.1)

\( \sum_{i=1}^{N} g_{\theta,i} = 1 \). The parameter \( \theta \) will usually be omitted from the notation, \( g_i \equiv g_{\theta,i} \). This situation is indicated in Figure 1, where the values \( 1, 2, \ldots, i, \ldots, N \) are ordered in the obvious way since they are intended to represent increasing realizations of the discretized life-time variate \( T \). (We could specify the endpoints of the \( i \)th time interval, but for the discrete calculations it is sufficient to know only that \( T \) falls into the \( i \)th interval, e.g. "\( T = i \).") The right cdf

\[
G_i \equiv G_{\theta,i} = \sum_{j \geq i} g_{\theta,j}
\]

(2.2)

is the probability \( T \geq i \). The \( i \)th discrete hazard is

\[
h_i \equiv h_{\theta,i} = g_{\theta,i}/G_{\theta,i} .
\]

(2.3)

Let \( T_1, T_2, \ldots, T_n \) represent an independent and identically distributed (iid) sample from the discrete distribution (2.1), and let
(2.4)

be the number of counts in category $i$. Also define

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

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

\[
\begin{array}{cccccccccc}
  n &=& n_1 &\rightarrow& n_2 &\rightarrow& n_3 &\rightarrow& \cdots & n_i &\rightarrow& n_{i+1} &\rightarrow& \cdots & n_N &\rightarrow& s_N = n_N \\
  g_1 &=& g_2 &=& g_3 &=& g_i &=& g_N \\
  l &=& G_1 &\rightarrow& G_2 &\rightarrow& G_3 &\rightarrow& G_i &\rightarrow& G_N \\
  h_1 &=& h_2 &=& h_3 &=& h_i &=& h_N = 1 \\
\end{array}
\]

Figure 1. The discrete situation: $g_i$ is the probability that $T$ falls into the
ith time interval, $i = 1, 2, \ldots, N$; $G_i$ is the right cdf, $\Sigma_{j \geq i} g_i$; $h_i$ is the
ith discrete hazard $g_i/G_i$. A random sample of size $n$ has been drawn from this
distribution, resulting in $s_i$ observations in the ith time interval, $i = 1, 2,
\ldots, N$. In terms of the left-to-right construction, $n_i \equiv \sum_{j \geq i} s_j$ items are at
risk at the beginning of the ith time interval.

We then have the following elementary result:

**Left-to-Right Identity.** The probability of observing counts $\xi = (s_1, s_2, \ldots, s_N)$
is

$$\left(\frac{n!}{s_1! s_2! \cdots s_N!}\right)^N \prod_{i=1}^N g_i^{s_i} = \prod_{i=1}^N \left(\frac{n_i}{s_i}\right) \left(1-h_i\right)^{n_i-s_i}. \quad (2.6)$$
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 \( n_1 \) 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,

\[
s_i \mid n_i \sim \text{Bi}(n_i, h_i). \tag{2.7}
\]

Multiplying the successive binomial densities (2.7) gives the right side of (2.6).

Figure 1 illustrates the left-to-right process: beginning with all \( n = n_1 \) items available ("at risk", in the usual censored data terminology) at the beginning of the time scale, \( s_1 \sim \text{Bi}(n_1, h_1) \) die at the first opportunity, leaving \( n_2 = n_1 - s_1 \) at risk at the beginning of the second time interval; \( s_2 \sim \text{bi}(n_2, h_2) \) die in the second interval, leaving \( n_3 = n_2 - s_2 \) at risk for interval 3, etc.

Remark A. It is easy to verify (2.6) directly from definitions (2.2), (2.3), (2.5). Barlow et al (1972) gives a likelihood-based derivation on pages 104-105.

Remark B. The \( N \)th term on the right side of (2.6) equals 1 (since \( h_N = 1 \) and \( s_N = n_N \)), and so can be omitted.

Remark C. For censored data, \( n_{i+1} = n_i - s_i - c_i \), where \( c_i \) is the number of items "lost to follow-up" during the \( i \)th interval. The left-to-right construction becomes crucial in this situation, as discussed briefly in Section 4.

Remark D. Result (1.6), as it applies to discrete variables, can be proved easily by evaluating the Fisher information from each side of (2.6). Instead we will derive a more general result, essentially (1.8) for the discrete case, and then return to (1.6).

3. Conditional and Unconditional Deviations.

Let

\[
D_i = s_i - n_i \quad \text{and} \quad d_i = s_i - n_i h_i, \quad (3.1)
\]
so \( D_i \) is the deviation of the count \( s_i \), (2.4), from its unconditional expectation \( n g_i \); and \( d_i \) is the deviation of \( s_i \) from its conditional expectation \( n_i h_i \), (2.7). (Notice that \( \Sigma D_i = 0 \) and \( d_N = 0 \).) Our results follow from a lemma relating the \( D_i \) to the \( d_i \):

**Deviations Lemma.** For any vector \( \tilde{b} = (b_1, b_2, \ldots, b_N) \), we have

\[
\sum_{i=1}^{N} D_i b_i = \sum_{i=1}^{N} d_i a_i,
\]

where \( \tilde{a} = (a_1, a_2, \ldots, a_{N-1}, a_N) \) is given by

\[
a_i = b_i - \tilde{b}_{i+1} \quad \left[ \tilde{b}_{i+1} = \frac{1}{g_{i+1}} \sum_{j \geq i+1} g_j b_j \right],
\]

for \( i = 1, 2, \ldots, N-1 \), while \( a_N \) is arbitrary. Moreover, letting \( \tilde{b} = \sum_{i=1}^{N} g_i b_i \), the inverse transformation from \( \tilde{a} \) to \( \tilde{b} \) is given by

\[
b_i - \tilde{b} = a_i - \hat{a}_i \quad \left[ \hat{a}_i = \sum_{j < i} h_j a_j \right]
\]

for \( i = 1, 2, \ldots, N \).

(The proof of the deviations lemma appears later in this section.)

**Corollary.** For vectors \( \tilde{b} \) and \( \tilde{a} \) related as in (3.3), (3.4)

\[
\sum_{i=1}^{N} g_i (b_i - \tilde{b})^2 = \sum_{i=1}^{N} g_i (1 - h_i) a_i^2.
\]

**Proof:** The variance of the left side of (3.2) is

\[
\text{var} \left( \sum_{i=1}^{N} b_i s_i \right) = n \sum_{i=1}^{N} g_i (b_i - \tilde{b})^2
\]

according to standard multinomial calculations. Let \( s_i \equiv (s_{1,i}, s_{2,i}, \ldots, s_{i,i}) \). Notice that (2.7) can be written in the stronger form.
\[ s_i | s_{i-1} \sim \text{Bi}(n_i, h_i) \quad \quad \left[ n_i = n - \sum_{j<i} s_j \right] \tag{3.7} \]

This implies

\[ \mathbb{E}(d_i a_i | s_{i-1}) = 0, \quad \mathbb{E}(d_i a_i)^2 | s_{i-1} = n_i h_i (1-h_i) a_i^2, \tag{3.8} \]

and

\[ \mathbb{E}(d_i a_i)(d_i a_i') | s_{i-1} = 0 \quad \text{for} \quad i' < i. \]

In other words the right side of (3.2), \( \sum_{i=1}^{N} d_i a_i \), is the successive sum of conditionally uncorrelated terms with conditional mean 0 and conditional variance \( n_i h_i (1-h_i) a_i^2 \). The "total conditional variance"

\[ \text{var}(n) = \sum_{i=1}^{N} n_i h_i (1-h_i) a_i^2 \tag{3.9} \]

has unconditional expectation

\[ \mathbb{E}(\text{var}(n)) = \mathbb{E}\left( \sum_{i=1}^{N} n_i h_i (1-h_i) a_i^2 \right) = \sum_{i=1}^{N} n_i h_i (1-h_i) a_i^2 \tag{3.10} \]

\[ = n \sum_{i=1}^{N} g_i (1-h_i) a_i^2. \]

But it is easy to see from (3.8) that

\[ \mathbb{E}(\text{var}(n)) = \mathbb{E}(\sum_{i=1}^{N} d_i a_i)^2 = \text{var}\left( \sum_{i=1}^{N} d_i a_i \right) \]

\[ \quad \quad \quad = \text{var}\left( \sum_{i=1}^{N} D_i b_i \right). \]

Comparing (3.10) with (3.6) verifies the corollary. \( \square \)

The corollary 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 (3.5) can be written

\[ \text{var}(b_T) = \mathbb{E}[(1-h_T) a_T^2] = \mathbb{E}\left[ (1-h_T) [b_T - \bar{b}_T]^2 \right] \tag{3.11} \]

\[ = \mathbb{E}\left\{ \left( \frac{1}{1-h_T} [b_T - \bar{b}_T] \right)^2 \right\}. \]
The last form of (3.11), which follows from the identity
\[ a_i = b_i - \beta_{i+1} = \frac{b_i - b_{i+1}}{1-h_i}, \]  
(3.12)
is exactly (1.8), except for a correction factor \(1/(1-h_i)\) necessary in the discrete case.

Returning to the Fisher information, let \( b_i = g_i / g_i = \frac{3}{30} \log(g_{0,i}) \). Then it is easy to calculate \( a_i = (1-h_i)^{-1} \frac{\partial}{\partial \theta} \log(h_{0,i}) \). In this case the corollary gives
\[ \sum_{i=1}^{N} g_i \left( \frac{\dot{g}_i}{g_i} \right)^2 = \sum_{i=1}^{N} g_i \frac{1}{1-h_i} \left( \frac{h_i}{h_i} \right)^2, \]
(3.13)
which is the discrete version of (1.6). Formula (1.6) for the continuous case follows by discretizing \( T \) into infinitesimal time intervals, in which case the correction factor \( 1/(1-h_i) + 1 \), see Remark G. A direct proof of the continuous case appears in Section 5.

Proof of Deviations Lemma: If formula (3.2) is true for \( n = 1 \) it is true in general, by additivity. Therefore, it is enough to prove (3.2) for \( n = 1 \), in which case only a single life-time \( T \in \{1, \ldots, n\} \) is chosen. The identity becomes
\[ b_T - \Sigma b_i g_i = a_T - \sum_{i\leq T} h_i a_i, \]
(3.14)
or \( b_T - \beta = a_T - \bar{a}_T \), which is just the definition (3.4) of \( \beta \) in terms of \( a \). Substituting for \( \{a_i\} \) its definition (3.3) in terms of \( \{b_i\} \), the difference between left and right sides of (3.14) becomes
\[ b_T - \beta = (b_T - \beta_{T+1}) + \sum_{i \leq T} h_i (b_i - \beta_{i+1}). \]
(3.15)
An elementary calculation gives
\[ h_i(b_i+5_i+1) = (5_i+1+5_i) \quad (3.16) \]

showing that (3.15) equals 0, so (3.2) (3.3) is true. This shows the transformation \( A : \{ b_i \} \to \{ a_i = b_i - 5_i+1 \} \) and \( B : \{ a_i \}_{i=1}^{i-1} \to \{ b_i = a_i - 5_i \} \) are inverses in the sense that \( ABA = a \) and \( BBA = b - 5 \), completing the proof. \( \square \)

**Remark E.** The result \( \text{E}(\text{var}(\varphi)) = \text{var}(\varphi) \) used to complete the proof of the corollary, is a standard martingale argument. It appears frequently in the censored data literature, for example in Section 4 of Cox (1975). In martingale terminology, \( \Sigma d_i a_i \) is a "martingale transform" of the martingale \( \Sigma d_i \) obtained by subtracting the "compensator" \( \Sigma h_i \) from the counting process \( \Sigma s_i \).

**Remark F.** The left-to-right identity yields other martingale results. For any value of \( \phi \),

\[ R_i \equiv e^{-\phi a_i(s_i-n_i h_i)} e^{-n_i k_i} \quad [k_i \equiv \log((1-h_i) e^{-\phi a_i h_i} + h_i e^{\phi a_i(1-h_i)})] \quad (3.17) \]

has conditional expectation 1 under (3.7), so \( R_1, R_1 R_2, \ldots, \Pi_1^n R_i \) is an exponential martingale with expectation 1. Evaluating \( \text{E}(\Pi_1^n R_i) = 1 \) from the unconditional multinomial distribution of \( s \) gives the following identity:

\[ 1 = \sum_{i=1}^{n} g_i e^{\phi b_i k_i} \quad [k_i \equiv \sum_{j<i} k_j] \quad (3.18) \]

Differentiating (3.21) twice with respect to \( \phi \) is another way to derive (3.5).

**Remark G.** In the continuous case we can always transform \( T \) so that \( g_0(t) \) is supported on the unit interval \( [0,1] \). The time intervals indicated in Figure 1 can be taken of equal length \( \Delta = 1/N \). Under moderate regularity conditions, \( g_i = g(t_i) \Delta + O(\Delta^2) \) for \( t_i \) the midpoint of the \( i \)th interval, and likewise \( G_i = G(t_i) + O(\Delta), h_i = h(t_i) \Delta + O(\Delta^2) \). Letting \( N \to \infty \), (3.5) has as its limit (1.8). A direct proof of the continuous case, including regularity conditions, appears in Section 5.
Remark H. The total conditional variance \( \text{var}(n) \), (3.9), is closely related to Greenwood's formula for the variance of an estimated survival curve. The "life-table" or "actuarial" estimate for \( G_i = \prod_{j<i} (1-h_j) \) is

\[
\hat{G}_i = \prod_{j<i} (1-\hat{h}_j) \quad [\hat{h}_j \equiv s_j/n_j].
\]

Therefore

\[
\log \frac{\hat{G}_i}{G_i} = \sum_{j<i} \log \left[ 1 - \frac{\hat{h}_j - h_j}{1-h_j} \right] = \sum_{j<i} \log \left[ 1 - \frac{s_j - n_j h_j}{n_j (1-h_j)} \right] \quad (3.20)
\]

\[
= \sum_{j<i} \frac{s_j - n_j h_j}{n_j (1-h_j)}.
\]

This has the form \( \sum_{j=1}^{N} d_j a_j \) of the right side of (3.2), with

\[
a_j = \begin{cases} 
-1/n_j(1-h_j) & j < i \\
0 & j \geq i 
\end{cases}
\]

(3.21)

The total conditional variance (3.9) is now

\[
\text{var}(n) = \sum_{j<i} \frac{h_j}{n_j (1-h_j)}. \quad (3.22)
\]

Formula (3.25) is, approximately, an unbiased estimator for the variance of \( \log \hat{G}_i/G_i \). [This is true even though \( a_j \) depends on \( n_j \), and even if there is censored data, see Section 4.] Substituting \( \hat{h}_j \) for the unknown \( h_j \) in (3.25) gives Greenwood's formula

\[
\hat{\text{var}}(\log \hat{G}_i/G_i) = \sum_{j<i} \frac{s_j}{n_j(n_j-s_j)}, \quad (3.25)
\]

Miller (1986), pg. 45.
Remark I. The situation we have been discussing, as illustrated in Figure 1, can be thought of as a random walk down a particularly simple binary tree shown on the left side of Figure 2. Results similar to the Deviations Lemma hold for arbitrarily complicated binary trees.

Suppose that $n$ items independently walk down a binary tree, according to the probability mechanism described in Figure 2's caption. Let

$$s_i = \#\{\text{items ending at terminal node } t_i\}$$

$$n_j = \#\{\text{items passing through parent node } p_j\}$$

$$S_j = \#\{\text{of the } n_j \text{ items that descend left from node } p_j\}$$

$$D_i = s_i - n_i \cdot h_i$$ and $$d_j = S_j - n_j \cdot h_j.$$  \hspace{1cm} (3.24)

Notice that $S_j/n_j \sim Bi(n_j, h_j)$ as in (2.7). A generalization of the Deviations Lemma applies to binary trees,

$$\sum_i D_i \cdot b_i = \sum_j d_j \cdot a_j,$$ \hspace{1cm} (3.25)

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

$$a_j = \sum_{i(<L)j} g_i b_i / \sum_{i(<L)j} g_i - \sum_{i(<R)j} g_i b_i / \sum_{i(<R)j} g_i$$

$$b_i = \sum_{j(>L)i} (1-h_j)a_j - \sum_{j(>R)i} h_j a_j.$$ \hspace{1cm} (3.26)

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.
Figure 2. 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 \( g_1, g_2, g_3, \ldots \). The left diagram is the tree representation of Figure 1, \( N = 5 \). The right diagram shows a more complicated tree structure.

Result (3.25), (3.26), which will not be proved here, leads to the following variance identity,

\[
\sum_{i} g_i b_i^2 = \sum_{j} G_j h_j (1-h_j) a_j^2 \quad \left[ G_j \equiv \sum_{i(<)j} g_i \right], \tag{3.27}
\]

analogous to (3.5).

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

\[
\hat{\text{var}}\{\log \hat{G}_j / G_j\} = \sum_{j' < j} \frac{1}{n_j-n_j'-s_j'-\delta_j'} S_j'
\]
where δ_j is -1 or +1 according as the left or right path through node j' is taken on the way to node j.

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. Closed form maximum likelihood estimates of cell probabilities are easily obtained, even under restrictions on the transition probabilities h_ij. The approach can be extended to include the class of "decomposable models" introduced by Goodman (1970) and Haberman (1974) for complete multiway contingency tables.

4. Is \( \text{var}(\hat{\theta}) \) A Conditional Variance?

This section returns to the discrete situation (2.1), and considers a statistical question concerning the total conditional variance (3.9), \( \text{var}(\hat{\theta}) = \sum_{i=1}^{N} n_i h_i (1-h_i) \sigma_i^2 \).

Let \( X = \sum_{i=1}^{N} d_i a_i = \sum_{i=1}^{N} (s_i - n_i h_i) a_i \) and

\[
\sigma^2 = \text{var}(X) . \tag{4.1}
\]

A key fact in censored data theory is that \( \text{var}(\hat{\theta}) \) is an unbiased estimator of \( \sigma^2 \),

\[
\text{E} \{ \text{var}(\hat{\theta}) \} = \sigma^2 , \tag{4.2}
\]

even if there is data censoring, as mentioned in Remark C. For instance, the justification of Greenwood's formula for the variance of a survival curve is based on (4.2), see Remark H. Equation (4.2) follows from a Martingale argument similar to the proof of (3.5). Section 5.2 of Kalbfleisch and Prentice (1980) gives a nice discussion of the allowable data censoring mechanisms.

Of course \( \text{var}(\hat{\theta}) \) will sometimes exceed its mean value \( \sigma^2 \), perhaps by a considerable amount, and sometimes will be less than \( \sigma^2 \). The question we wish
to consider is this: is $\text{var}(\mathbf{n})$ a conditional variance estimator in the sense that

$$\text{var}\{X|\text{var}(\mathbf{n})\} \neq \text{var}(\mathbf{n})$$

That is, if $\text{var}(\mathbf{n})$ is unusually big should we believe that $X$ is unusually variable, and conversely? Figure 3 shows how a plot of $X$ versus $\text{var}(\mathbf{n})$ looks if (4.3) is true.

Figure 3. A plot of $X$ versus $\text{var}(\mathbf{n})$ in a situation where (4.3) is true. Actually shown are 4000 replications of $(\text{var}(\mathbf{n}),X)$ from the censored data situation (4.9), (4.10).
It would be nice if (4.3) were true since that would support the use of $\text{var}(\bar{n})$ as an estimate of $\text{var}(X)$. In fact $\text{var}(\bar{n})$, and the analogous quantities related to the Mantel-Haenszel test, Cox's partial likelihood, etc., are the only reasonable variance estimates available in the censored data context, so they must be used. If (4.3) were true, then necessity becomes virtue. A "lucky" case of just this sort occurs in maximum likelihood estimation, where the easy-to-compute observed Fisher information gives better estimates of the conditional variance of the maximum likelihood estimator than does the expected Fisher information, see Efron and Hinkley (1978).

Unfortunately it is easy to demonstrate that (4.3) is false, even as a rough approximation, when there is no data censoring. In this case both $X$ and $\text{var}(\bar{n})$ are linear combinations of the counts $s_i$,

$$X = \sum_{i=1}^{N} s_i (a_i - \bar{a}_i) = \sum_{i=1}^{N} s_i b_i,$$  \hspace{1cm} (4.4)

(3.18), $(b_i = a_i - \bar{a}_i$ is the quantity $b_i - \bar{b}$ in (3.4), so here $\bar{b} = \Sigma_i g b_i = 0$), and

$$\text{var}(\bar{n}) = \sum_{i=1}^{N} s_i C_i \quad \quad [C_i = \Sigma_{j<i} h_j (1-h_j) a_j^2]$$  \hspace{1cm} (4.5)

since $\Sigma_i n_i h_i (1-h_i) a_i^2 = \Sigma_i [\Sigma_{j>i} s_j (h_i (1-h_j) a_j^2)] = \Sigma_i s_i [\Sigma_{j<i} h_j (1-h_j) a_j^2].$

A standard calculation based on the multinomial distribution of $\bar{s}$ gives the mean vector and covariance matrix of $(\text{var}(\bar{n}),X)$:

$$\begin{pmatrix} \text{var}(\bar{n}) \\ X \end{pmatrix} \sim \begin{pmatrix} \sigma^2 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \text{var}(C) & \text{cov}(b,C) \\ \text{cov}(b,C) & \text{var}(b) \end{pmatrix},$$  \hspace{1cm} (4.6)

where $\text{var}(C) = \Sigma_i g_i [C_i - (\Sigma_j g_j C_j)]^2$, $\text{cov}(b,C) = \Sigma_i g_i b_i C_i$, $\text{var}(b) = \Sigma g_i b_i^2$.
To a first order of approximation we would expect from (4.6) that
\[ \text{var}(X | \text{var}(n)) \approx (1 - \text{cor}_g(b,C)^2)\sigma^2, \tag{4.7} \]
\[ \text{cor}_g(b,C) \equiv \frac{\text{cov}_g(b,C)}{\text{var}_g(b) \text{var}_g(C)} \frac{1}{2}, \]
rather than (4.3). In other words, the conditional variance of \( X \) given \( \text{var}(n) \) would always be smaller than \( \sigma^2 \), much smaller if \( \text{cor}_g(b,C)^2 \) were large.

This prediction was amply demonstrated in a simulation of the case
\[ N = 10, \quad g = (0.1, 0.1, \ldots, 1), \quad a = (1, 1, \ldots, 1), \quad n = 100. \tag{4.8} \]
The true variance is \( \sigma^2 = 70.71 \), formula (3.6). The definitions above give \( \text{cor}_g(b,C) = -0.936 \), so that (4.7) predicts \( \text{var}(X | \text{var}(n)) \approx 8.76 \).

Figure 4 shows 2000 independent Monte Carlo replications of \( (\text{var}(n), X) \) for situation (4.8). The correlation structure predicted by (4.6), (4.7) is evident. Table 1 shows that the approximation \( \text{var}(X | \text{var}(n)) \approx 8.76 \) is reasonably accurate, except for the extreme deciles of \( \text{var}(n) \).

Next, suppose we change situation (4.8) by the addition of a data-censoring mechanism, as in Remark C:
\[ n_{i+1} = n_i - a_i - c_i \quad \text{where} \quad c_i | n_i, z_i \sim \text{Bi}(n_i - s_i, p) \tag{4.9} \]
for \( i = 1, 2, \ldots, N-1 = 9 \). The notation \( c_i | n_i, z_i \) means that \( c_i \) has the indicated binomial distribution, conditional on \( n_i = (n_1, n_2, \ldots, n_i) \) and \( z_i = (s_1, \ldots, s_i) \) as in (3.7). Having observed \( n_{i+1} \) from (4.9) then \( s_{i+1} \) is \( \text{Bi}(n_{i+1}, h_{i+1}) \) as in (2.7). The censoring probability \( p \) in (4.9) was constant for any one replication of \( (\text{var}(n), X) \), but varied independently from replication to replication of the simulation experiment shown in Figure 3,
\[ p \sim \text{uniform}[0,1]. \tag{4.10} \]
Figure 4. 2000 replications of $(\text{var}(\bar{z}), X)$ for situation (4.8). It is clear that (4.3) is not true; $\text{var}(X|\text{var}(\bar{z}))$ is much smaller than the unconditional variance $\sigma^2$, for any small interval of $\text{var}(\bar{z})$ values.

<table>
<thead>
<tr>
<th>$\text{var}(\bar{z})$ decile:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{var}(X</td>
<td>\text{var}(\bar{z}))$:</td>
<td>19.25</td>
<td>9.04</td>
<td>10.13</td>
<td>8.76</td>
<td>9.17</td>
<td>8.31</td>
<td>10.44</td>
<td>9.58</td>
<td>9.56</td>
</tr>
<tr>
<td>$E(X</td>
<td>\text{var}(\bar{z}))$:</td>
<td>13.42</td>
<td>7.85</td>
<td>5.17</td>
<td>3.00</td>
<td>0.96</td>
<td>-0.96</td>
<td>-3.11</td>
<td>-4.91</td>
<td>-8.39</td>
</tr>
</tbody>
</table>

Table 1. Sample variances and expectations of $X$ when the points in Figure 4 are divided into the deciles of $\text{var}(\bar{z})$. For example the $X$ values corresponding to the smallest 200 values of $\text{var}(\bar{z})$ had sample variance 19.25, sample mean 13.42.
Table 2 shows that (4.3) is now quite accurate. A reasonable objection can be raised: since we know (4.2) is true conditional on $p$, since large values of $p$ produces small values of $\text{var}(n)$ and conversely, and since (4.10) varies $p$ drastically from replication to replication, aren't the results in Table 2 a foregone conclusion?

<table>
<thead>
<tr>
<th>$\text{var}(n)$ decile:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{var}(X</td>
<td>\text{var}(n))$:</td>
<td>9.88</td>
<td>11.14</td>
<td>12.06</td>
<td>13.76</td>
<td>14.72</td>
<td>17.91</td>
<td>23.96</td>
<td>32.24</td>
<td>40.75</td>
</tr>
<tr>
<td>mean of $\text{var}(n)$:</td>
<td>9.44</td>
<td>10.58</td>
<td>11.91</td>
<td>13.79</td>
<td>16.21</td>
<td>19.48</td>
<td>24.60</td>
<td>31.82</td>
<td>42.86</td>
<td>59.78</td>
</tr>
<tr>
<td>ratio:</td>
<td>1.05</td>
<td>1.05</td>
<td>1.01</td>
<td>.998</td>
<td>.968</td>
<td>.919</td>
<td>.974</td>
<td>1.01</td>
<td>.957</td>
<td>.886</td>
</tr>
<tr>
<td>$E(X</td>
<td>\text{var}(n))$:</td>
<td>.054</td>
<td>.055</td>
<td>.102</td>
<td>.015</td>
<td>-.034</td>
<td>.119</td>
<td>.220</td>
<td>.036</td>
<td>-.231</td>
</tr>
</tbody>
</table>

Table 2. The 4000 replications of $(\text{var}(n), X)$ from the censored data situation (4.8), (4.9), (4.10); summary statistics for the replications separated into the deciles of $\text{var}(n)$. In this case (4.3) is reasonably accurate.

Table 3 shows what happens if we fix $p = .5$ in (4.9), so the objection just raised no longer applies. We see that (4.3) remains roughly true, though not as accurately as in Table 2.

The results presented here are far from conclusive, but they do suggest these conclusions: in the absence of censoring, (4.3) will usually be false; if censoring plays a major role in determining the sample sizes $n_i$, and particularly if this role is quite variable from replication to replication, then (4.3) should be roughly accurate.
\begin{table}
\begin{tabular}{cccccccccccc}
\text{var}(\eta) decile: & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
\hline
\text{var}(X|\text{var(}\eta)): & 20.33 & 18.36 & 19.75 & 26.11 & 27.87 & 38.30 & 46.28 & 50.52 & 52.53 & 51.42 \\
\text{mean of var(}\eta): & 17.87 & 20.43 & 22.82 & 25.72 & 29.79 & 34.18 & 39.51 & 46.18 & 58.84 & 66.56 \\
\text{ratio:} & 1.14 & .897 & .866 & 1.02 & .935 & 1.12 & 1.17 & 1.09 & .958 & .772 \\
\text{E}(X|\text{var(}\eta)): & .869 & .056 & -.018 & .564 & .178 & .439 & .113 & .321 & -.184 & -.2885 \\
\end{tabular}
\caption{2000 replications of \((\text{var}(\eta),X)\) from (4.8), (4.9), with censoring probability fixed at \(p = .50\). Relation (4.3) is still roughly accurate.}
\end{table}

5. Continuous Case.

A direct proof of the identity (1.6) expressing Fisher's information in terms of the hazard rate is not difficult. It is given together with convenient regularity conditions in Remark J. The hazard rate identity states that the score function \(\dot{h}_\theta/h_\theta\) derived from the "hazard transform" \(h_\theta = \mathcal{H}(g_\theta)\) has the same second moment as the basic score function \(\dot{g}_\theta/g_\theta\). This property is somewhat remarkable: it is not shared by some other familiar one-to-one transforms of densities. For example, the "rootogram transform" \(h_\theta = \sqrt{g_\theta}\) used in defining the Hellinger metric and in data analysis has \(\int (\dot{h}_\theta/h_\theta)^2 g_\theta = (1/4)\int (\dot{g}_\theta/g_\theta)^2 g_\theta\), and for the quantile transform \(h_\theta(u) = G_\theta^{-1}(u)\), Fisher's information does not have a simple expression in terms of \(\dot{h}_\theta/h_\theta\).

Insight into this isometry property of the hazard transform flows from the more general result (1.8), which we now study. Fix the probability density \(g_\theta(t)\); the subscript \(\theta\) will be omitted from now on.

We consider two linear transformations operating on the space of functions \(M\) which are square-integrable relative to the density \(g(t)\) with
\[
\|b\|^2 = \int_{-\infty}^{\infty} b^2(t)g(t)dt.
\]
The first transformation is defined as in (1.8):
\[
(\mathbf{A}b)(t) = b(t) - \ddot{b}(t) = b(t) - \int_{t}^{\infty} b(s)g(s)ds/G(t) , \quad b \in M .
\] (5.1)
(Alternatively, this may be regarded as the continuous analog of (3.3).) Since $b$ is a conditional expectation, it is easily seen that $A$ is a bounded linear transformation of $M$ into itself.

To see that $A$ is in fact length-preserving, introduce the second transformation on $M$, defined by

$$ (Ba)(s) = a(s) - \int_{-\infty}^{s} a(t) h(t) dt , \quad a \in M , $$

where $h(t) = g(t)/G(t)$. The definition of $B$ may be motivated either by analogy with the discrete case, or by noting that it is the adjoint (or transpose) transformation to $A$. Using $\langle a, a' \rangle = \int_{-\infty}^{\infty} a(t) a'(t) g(t) dt$ to denote the inner product in $M$, we have

$$ \langle b, Ba \rangle = \int_{-\infty}^{\infty} b(t) \left[ a(t) - \int_{-\infty}^{t} a(s) h(s) ds \right] g(t) dt $$

$$ = \int_{-\infty}^{\infty} \left[ b(s) - \frac{1}{G(s)} \int_{s}^{\infty} b(t) g(t) dt \right] a(s) g(s) ds $$

$$ = \langle Ab, a \rangle , \quad a, b \in M . $$

(This manipulation is an abstraction of the main step of the hazard function identity in Remark J below, where we set $b = \frac{\dot{\alpha}}{\alpha_{0}}/\alpha_{0}$ and $a = h_{0}/h_{0}$.)

Not only is $B = A^{*}$, but also $B$ is (essentially) $A^{-1}$. Since the adjoint of $A$ equals its inverse, it must be length-preserving. More precisely, a calculation using integration by parts shows

$$ (BA)(b) = b - Eb , \quad b \in M $$

$$ (AB)(a) = a , \quad a \in M . $$

Consequently
$$\Var{b(T)} = \langle b, B a b \rangle = \langle a b, a b \rangle = E[b(T) - \bar{b}(T)]^2,$$

which is (1.8).

We have seen that the hazard rate identity is a special case of the isometry relation (1.8) in which $b = \dot{g}/g, \ a = \dot{h}/h$. There is, however, a more fundamental connection between the linear transformations $B$ (and $A^{-1}$) and the non-linear hazard transform $\mathcal{H}: g(t) \rightarrow g(t)/G(t)$ defined on the set of all probability density functions on $\mathbb{R}$. Although there may be value in articulating this connection in the language of (infinite dimensional) Riemannian geometry, we give only an informal, heuristic account here. The hazard transform $\mathcal{H}$ maps the space of probability density functions onto the space of hazard rates, which are nonnegative functions. Now fix a density $g$ and linearize $\mathcal{H}$ about $g$. If $b(t)$ is a bounded function with mean zero, then $g(t)(1 + \varepsilon b(t))$ is again a probability density for small $\varepsilon$. (If $\{g_0\}$ is a one parameter family of densities, as in Section 1, then it is natural to consider $b = \dot{g}_0/g_0$.) A simple calculation then shows that

$$\mathcal{H}[g(1 + \varepsilon b)] = \mathcal{H}[g][1 + \varepsilon a b] + o(\varepsilon).$$

The linear hazard transform $A$ arises as a form of logarithmic Gateaux derivative of $\mathcal{H}$ about the fixed density $g$.

By working with logarithms, we may write (5.6) as an ordinary Taylor expansion. Assume for convenience that $g(t)$ is everywhere positive, so that $\lambda(t) = \log g(t)$ is defined. If an arbitrary real valued function $\lambda(t)$ has $e^{\Lambda} = \int e^{\lambda(t)} dt < \infty$, then $g(t) = e^{\lambda(t) - \Lambda}$ defines a probability density. If we identify functions $\lambda$ that differ by an additive constant, we obtain a class $\mathcal{B}$ of functions in one-to-one correspondence with the class of positive probability densities.

Let $\mathcal{L}_\mathcal{H}$ denote the transform that takes $\log g$ into $\log h$. More precisely, $\mathcal{L}_\mathcal{H}(\lambda)(t) = \lambda(t) - \log \int_t^\infty e^{\lambda(s) - \Lambda} ds$ for a function of $\lambda \in \mathcal{B}$. Now let $b(t)$ be,
say, a bounded function not necessarily with mean 0. The linearisation (5.6) now takes the form

\[ \mathcal{L}_H (\log g + \varepsilon b) = \log h + \varepsilon Ab + o(\varepsilon) . \]

Thus the transformation \( A \) is just the (infinite dimensional) Jacobian 'matrix' of the mapping \( \mathcal{L}_H \).

Exactly analogous statements may be made for the inverse transforms \( \mathcal{L}_g \) and \( \mathcal{L}_f \) which map \( h \) to \( g \) and \( \log h \) to \( \log g \) respectively. For example

\[ \mathcal{L}_g (\log h + \varepsilon a) = \log g + \varepsilon Ba + o(\varepsilon) . \] (5.7)

These expansions indicate that the mutual invertibility of \( A \) and \( B \) follows from the mutual invertibility of \( \mathcal{L}_g \) and \( \mathcal{L}_H \) (inverse function theorem). Of course, other smooth one-to-one transforms of probability densities admit expansions such as (5.6) and (5.7), with \( A \) and \( B \) being linear, mutually inverse transformations. (For \( g(t) = \sqrt[3]{g(t)} \), for example, \( Ba = 2a, Ab = b/2 \).) The unusual property of the hazard transform is that \( B \) and \( A \) are also adjoints, and hence length-preserving. Thus, roughly speaking, the spaces of log densities (\( \mathcal{L} \) above) and log-hazards are not merely diffeomorphic, but also isometric. Hence the Riemannian distance between two (log) densities is not changed under transformation to (log) hazards, even though the transforms are far from rigid. (Some extra detail in the finite case is given in Remark N.)

**Remark J.** Proof of hazard rate identity (1.6): Recalling that

\[ \frac{\dot{h}_\theta(t)}{h_\theta(t)} = \frac{\dot{g}_\theta(t)}{g_\theta(t)} - \frac{\dot{\theta}(t)}{\theta(t)} \] (5.8)

and interchanging orders of integration, one finds
\[ \int_{-\infty}^{\infty} \left( \frac{\dot{g}_\theta(s)}{h_\theta(s)} \right)^2 g_\theta(s) ds = \int_{-\infty}^{\infty} \left[ \frac{\dot{h}_\theta(s) g_\theta(s)}{h_\theta(s) g_\theta(s)} - \frac{1}{G_\theta(s)} \int_{s}^{\infty} \frac{g_\theta(t)}{g_\theta(t)} g_\theta(t) dt \right] g_\theta(s) ds \]

\[ = \int_{-\infty}^{\infty} \frac{\dot{g}_\theta(t)}{g_\theta(t)} \frac{h_\theta(t)}{h_\theta(t)} - \int_{-\infty}^{t} \frac{h_\theta(s)}{h_\theta(s)} h_\theta(s) ds \] \[ g_\theta(t) dt . \]

(5.9)

In the second identity we have replaced \( g_\theta(s)/G_\theta(s) \) by \( h_\theta(s) \). Now interchange \( \partial/\partial \theta \) and integration in the inner integral:

\[ \int_{-\infty}^{t} \frac{\dot{h}_\theta(s)}{h_\theta(s)} ds = \frac{\partial}{\partial \theta} \int_{-\infty}^{t} \frac{\partial}{\partial s} [-\log G_\theta(s)] ds = -\frac{\dot{G}_\theta(s)}{G_\theta(s)} . \]

(5.10)

From (4.10) and (4.8), the extreme right hand integral in (5.9) reduces to Fisher's information \( \int (\dot{g}_\theta/g_\theta)^2 g_\theta \). This proof of (1.6) will be valid at \( \theta = \theta_0 \) under the following assumptions, which are chosen for convenience rather than generality.

1. There is a neighborhood \( N \) of \( \theta_0 \) in which the distributions \( G_\theta(t) \) have common support, are absolutely continuous with respect to \( t \) and twice differentiable with respect to \( \theta \).

2. \( \int [\dot{g}_\theta(t)/g_\theta(t)]^2 g_\theta(t) dt < \infty \) and \( \int [\dot{h}_\theta(t)/h_\theta(t)]^2 g_\theta(t) dt < \infty \).

3. Differentiation (w.r.t. \( \theta \)) may be performed under the integral sign for \( \int_{-\infty}^{\infty} g_\theta(t) dt \) and \( \int_{-\infty}^{t} h_\theta(t) dt \).

Familiar examples involving distributions whose support depends on the parameter show that the conditions may not be removed entirely. Thus if \( G_\theta(t) = e^{-(t-\theta)} I\{t>\theta\} \), then \( \dot{g}_\theta(t)/g_\theta(t) = I(t>\theta) \) but \( \dot{h}_\theta(t)/h_\theta(t) = 0 \) for \( t > \theta \). Secondly, if \( g_\theta(t) = 1/\theta I\{0<t<\theta\} \), then \( \dot{g}_\theta(t)/g_\theta(t) = -1/\theta^2 \) has finite variance whereas \( \dot{h}_\theta(t)/h_\theta(t) = -(\theta-t)^{-1} \) does not.
Remark K. Identity (1.8) is written for functions of a real-valued continuous
variante $T$. There is a simple extension for the case when one wishes to make ex-
licit the dependence of $T = T(x)$ on an abstract sample point $x \in \mathfrak{X}$. Altema-
tively, we may think of $T(x)$ as imposing a (semi) ordering on $\mathfrak{X}$. If $R(X)$ is
a random variable with finite variance, if the density of $T(X)$ is absolutely con-
tinuous, and if $\bar{R}(t) = E[R(X)|T(X)\geq t]$, then

$$\text{Var } R(X) = E[(R(X) - \bar{R}(X))^2] \quad (5.11)$$

where, by an abuse of notation, $\bar{R}(X) = \bar{R}(T(X))$. (4.11) is verified by bringing in
the conditional mean $\bar{R}(t) = E[R|T=t]$ , noting that $\bar{R}(t) = E[\bar{R}|T\geq t]$ ; and applying
(1.8) to $\bar{R}(T)$:

$$\text{Var } R = E \text{ Var}(R|T) + \text{ Var } \bar{R}(T)$$
$$= E E[(R-\bar{R})^2|T] + E[\bar{R}(T)-\bar{R}(T)]^2$$
$$= E(R-\bar{R})^2 .$$

Remark L. Stochastic deviations lemma. Let $T_1, \ldots, T_n$ be n i.i.d.
observations from an arbitrary distribution $F$ on $(-\infty, \infty)$. The deviations lemma
of Section 3 can be extended using stochastic integrals, yielding a version of the
variance identity (1.8) that includes both discrete and continuous cases. Write
$F_t = P_t(T\leq t)$ , and for the left and right continuous versions of the survivor
function, write respectively $G_t = P(T> t)$ and $G^+_t = P(T> t)$. The cumulative
hazard up to time $t$ is

$$H_t = \int_{(-\infty,t]} \frac{dF(s)}{G(s)} .$$

The number of deaths after time $t$ is $N_t = \#\{T_i< t\}$ and the number at risk before
time $t$ is $Y_t = \#\{T_i \geq t\}$.
Corresponding to the earlier unconditional and conditional deviations \( D_i \) and \( d_i \) are cumulative processes

\[
U_t = N_t - nF_t \\
C_t = N_t - \int_{(-\infty, t]} Y_s \, dH_s
\]

tracking unconditional and conditional deviations respectively. Now \( U_t \) is a zero mean process tied down to be 0 at \(-\infty\) and \(+\infty\), but \( C_t \) is more, namely a martingale. The deviations lemma states for a function \( b(t) \in L_1(dF) \) that

\[
\int b \, dU_t = \int a \, dC_t \tag{5.12}
\]

where \( a(t) = (Ab)(t) = b(t) - \bar{b}(t) = b(t) - (G_t^+)^{-1} \int_{(t, \infty)} b(s) \, dF_s \). Identity (5.12) is also valid if \( b(t) = (Ba)(t) \) is defined in terms of \( a(t) \) by the inverse transformation \( b(t) = a(t) - \int_{(-\infty, t]} a(s) \, dH_s \). The proof, similar in spirit to that of the earlier deviations lemma, but requiring some care with integration by parts, is omitted.

Evaluating the variance of the left side of (5.12) directly, and the right side by martingale theory leads to the generalization of (1.8) and (3.5):

\[
\int (b(s) - \bar{b})^2 \, dF_s = \int a^2(s)(1 - \Delta H(s)) \, dF_s ,
\]

where \( \Delta H(s) = H(s) - H(s-) \) measures jumps in the cumulative hazard, and \( \bar{b} = \int b \, dF \).

We will not discuss asymptotics here, save to remark, for example, that in the case where \( F \) is uniform on \([0,1]\), the relation (5.12) goes over, as \( n \to \infty \), to

\[
\int b \, dW_0 = \int a \, dW_t \quad \text{(equality in distribution)}
\]

where \( W_0 \) and \( W \) are respectively standard Brownian bridge and standard Brownian motion on \([0,1]\).
Remark M. Fisher's information is a local version of the Kullback-Leibler information distance between two densities \( g_\theta \) and \( g_{\theta + s} \) belonging to a one-parameter family: 

\[
K(g_\theta, g_{\theta + s}) = \int g_\theta \log \frac{g_{\theta + s}}{g_\theta} \, \frac{s^2}{2} \int H_\theta + o(s^2).
\]

Since the hazard rate identity (1.6) is a property of the derivatives of the hazard transform and its inverse (here \( g_\theta = h_\theta e^{-H_\theta} \), where \( H_\theta(t) = \int_0^t h_\theta(v) \, dv \) is the cumulative hazard), it is natural to seek a proof of (1.6) in terms of the Kullback-Leibler distance. Indeed, under appropriate smoothness conditions

\[
K(g_\theta, g_{\theta + s}) = \int h_\theta e^{-H_\theta} \left\{ \log \frac{h_{\theta + s}}{h_\theta} - (H_{\theta + s} - H_\theta) \right\}
\]

\[
= \int h_\theta e^{-H_\theta} \left\{ \log \left(1 + \frac{h_{\theta + s} - h_\theta}{h_\theta} \right) - \frac{h_{\theta + s} - h_\theta}{h_\theta} \right\}
\]

\[
= -\frac{1}{2} \int h_\theta e^{-H_\theta} \left( \frac{h_{\theta + s} - h_\theta}{h_\theta} \right)^2 + \ldots
\]

\[
= -\frac{s^2}{2} \int g_\theta \left( \frac{h_\theta}{h_\theta} \right)^2 + \ldots
\]

where at the third equality, an integration by parts was used.

Remark N. Multiparameter case. Equality (5.5) extends immediately to the covariance matrix \( \text{Cov}(b(T)) \) of a vector function \( b(T) \) of the scalar \( T \):

\[
\text{Cov}(b(T)) = E[b(T) - \bar{b}(T)][b(T) - \bar{b}(T)]^T.
\]

In particular, for a nice family of probability densities \( \{ g_\theta(t); t \in \mathbb{R} \} \) depending on a vector parameter \( \theta \), the basic identity (1.6) extends to

\[
\mathcal{J}_\theta = \text{Cov}_\theta \begin{bmatrix} \nabla g_\theta \\ g_\theta \end{bmatrix} = \begin{bmatrix} \nabla h_\theta \\ h_\theta \end{bmatrix}^T \int \frac{\nabla h_\theta}{h_\theta} \, g_\theta(t) \, dt,
\]

where \( \nabla = \left( \frac{\partial}{\partial \theta_1} \right) \) denotes gradient with respect to the components of \( \theta \).
Extension to multidimensional $\tau$ is not straightforward (except in the trivial sense of Remark K) 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" $\nabla_t \log P(T > t)$, discussed for example by Marshall (1975).

Remark 0. The role of the linear transformations $A$ and $B$ as linearisations of the transforms $H, \mathcal{L} H$, etc., can be detailed more precisely in the discrete case. Our goal is to interpret the variance identity (3.5) as just the expression of the 'natural' Riemannian distance metric in the log-density and log-hazard coordinate systems. Let $\mathcal{S}$ denote the (open) simplex of probability distributions $g = \{g_i\}_{i=1}^N$ on $N$ cells. This is thought of as an abstract space, onto which coordinates must be introduced in order to do computations. Thus if $\Theta: \mathcal{S} \rightarrow \mathbb{R}^{N-1}$ is a particular choice of coordinates, then the Fisher information metric is defined at a particular distribution $g$ by $\gamma_{ij}(g) = E_g \left[ \frac{\partial}{\partial \theta} \log p(X, \theta) \frac{\partial}{\partial \theta} \log p(X, \theta) \right]$, where in our present case $p(i, \theta(g)) = g_i$. For example, if $\alpha(t)$ is a parametrisation of a curve $C$ in $\mathcal{S}$ connecting $g_1$ and $g_2$, then the length of $C$ is given (symbolically) by $\int_{\gamma_{ij}} \sum_{i,j} \frac{\partial}{\partial \theta} \gamma_{ij} \, d\theta$. (It will be seen that we abuse notation here by switching some subscripts to superscripts according as the emphasis is on the approach of earlier sections or on the geometric perspective.)

The log-density coordinate system is specified by $\theta^i(g) = \log g_i$, $i = 1, \ldots, N-1$, so that Fisher's information metric is given by the $(N-1) \times (N-1)$ matrix with entries $\gamma_{ij}(g) = g_i \delta_{ij} - g_i g_j / g_N$. The corresponding quadratic form $\langle b, b \rangle_g = \sum_{i,j=1}^{N-1} b_i \gamma_{ij}(g) b_j$, simplifies to the familiar $L^2$ norm $\sum_{i=1}^N b_i^2 g_i$ if $b_N$ is defined so that $\sum_{i=1}^N b_i g_i = 0$. 

28
The log-hazard coordinate system is given by \( \xi^\alpha(g) = \log h_\alpha = \log g_\alpha - \log g_{\alpha+1}, \)
\( \alpha = 1, \ldots, N-1. \) The slightly modified hazard \( \tilde{h}_\alpha \) used here is related to the transform \( h_\alpha \) of Section 2 by \( \tilde{h}_\alpha = h_\alpha/(1-h_\alpha). \) Fisher's information metric in this coordinate system, \( \gamma^\xi_{\alpha\beta}(g) \) say, could be calculated directly from the definition. Our point is better illustrated, however, by deriving it from the log-density coordinate metric by the usual rules for change of coordinates:

\[
\gamma_{\alpha\beta} = \sum_{i,j=1}^{N-1} B^i_{\alpha} \gamma_{ij} B^j_{\beta}, \tag{5.12}
\]

where \( B^i_{\alpha} = \partial \xi^i / \partial \xi^\alpha = \delta_{i\alpha} - \chi_{\alpha<1} h_\alpha \) is the \((N-1) \times (N-1)\) Jacobian matrix of the coordinate transformation \( \theta \rightarrow \xi. \) The metric in log-hazard coordinates is then calculated to be \( \gamma_{\alpha\beta} = \delta_{\alpha\beta} g_\alpha (1-h_\alpha), \) with corresponding quadratic form

\[
<\tilde{a}, \tilde{a}>_h = \sum_{i=1}^{N-1} a^2_\alpha (1-h_\alpha) g_1. \]

This is the modified \( L^2 \) norm occurring on the right side of Corollary 3.5. Thus the perhaps mysterious "discreteness correction" \( 1-h_\alpha \) arises as a natural consequence of the change to hazard coordinates.

It is now easy to read off the variance identity (3.5). The change of basis matrix \( B^i_{\alpha} \) is precisely the matrix of the transformation from \( \sim \) to \( \tilde{b} \) occurring in the deviations lemma (3.4). Using summation convention, we therefore have from (5.12)

\[
<\sim, \sim>_h = a^\alpha \gamma_{\alpha\beta} a^\beta = a^\alpha B^i_{\alpha} \gamma_{ij} B^j_{\beta} a^\beta = b^i \gamma_{ij} b^j = \langle b, b \rangle_g
\]

which is (3.5). Of course, one can also go in the opposite direction: the transformation \( A \) from \( \tilde{b} \) to \( \sim \) given in (3.3) is just the change of basis matrix \( \tilde{B}^\alpha_i = \partial \xi^\alpha / \partial \xi^i, \) the inverse of \( B^i_{\alpha}. \)
The metric relations (5.12) show that an identity analogous to (3.5) will be available for arbitrary transformations of coordinates (for example, Amari, 1985, pp. 31-32 discusses the square root transformation). The special character of the hazard transformation lies in two points: it is an isometry (the discrete version of (5.3) says that $(B_{\alpha}^i)$ and $(\tilde{B}_{\alpha}^i)$ are not merely inverses, but also adjoints relative to the appropriate inner products) and the resulting metric still has a simple (diagonal) form.
References


