SOME GENERAL RESULTS ON TIME-ORDERED CLASSIFICATION

BY

D. V. HINKLEY

TECHNICAL REPORT NO. 4

JULY 30, 1971

PREPARED UNDER CONTRACT

N00014-67-A-0112-0030 (NR-042-034)

FOR THE OFFICE OF NAVAL RESEARCH

THEODORE W. ANDERSON, PROJECT DIRECTOR

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

SOME GENERAL RESULTS ON TIME-ORDERED CLASSIFICATION

BY

D. V. HINKLEY
Stanford University

TECHNICAL REPORT NO. 4

JULY 30, 1971

PREPARED UNDER THE AUSPICES
OF
OFFICE OF NAVAL RESEARCH CONTRACT #N00014-67-A-0112-0030

THEODORE W. ANDERSON, PROJECT DIRECTOR

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

SUMMARY

The paper examines some properties of inference about the point in a sequence of random variables at which a distribution change occurs. Three particular aspects are considered: asymptotic theory, approximations to test power, and use of discriminant functions other than the likelihood ratio. The discussion is illustrated by some examples.

1. INTRODUCTION

In three recent papers, Hinkley (1970, 1971) and Hinkley and Hinkley (1970) have discussed likelihood inference for the unknown change-point parameter ξ in the model

$$\mathcal{L}(X_i) = \begin{cases} F(x, \theta) & i=1, \ldots, \xi \\ G(x, \psi) & i=\xi+1, \ldots, T \end{cases}$$

for the sequence $(X_1, \ldots, X_T)$ of independent random variables. Particular emphasis was placed on the normal mean-shift case and the binomial proportion-shift case, for which detailed numerical results were given. However, some simple general results were overlooked. In the present paper we attempt to remedy that situation.

Section 2 is concerned primarily with likelihood ratio tests of significance about ξ, and a brief summary of previous work is given in Section 2.1. Then we derive simple approximations to the power functions of the tests using approximations for the tails of the null distributions. For this discussion we assume θ and ψ to be known, with ξ and $T-\xi$ assumed large. Asymptotically in ξ and $T-\xi$ the same results apply with θ and ψ unknown, as we prove in Section 2.3 using standard asymptotic results for likelihood estimation. Some examples of the power approximations are given in Section 2.4.

We may regard the change-point model as defining a classification problem, in which we have to classify all observations $X_1, \ldots, X_T$. We have a strong ordering principle through which ξ identifies the

correct classification. This point of view suggests consideration of discriminant functions other than the likelihood ratio, since the latter may be inconvenient for certain situations. One familiar discriminant function is the cumulative sum chart developed by Page (1954). In Section 3 we generalize the likelihood results of Section 2 to general discriminant functions, with some examples in Section 3.4.

Some brief remarks on relevant aspects of the problem are made in Section 4.

2. LIKELIHOOD INFERENCE ABOUT ξ

The general parametric single-change model for a sequence of independent vector random variables $(X_1, \ldots, X_T)$ may be written as

$$\mathcal{L}(X_j) = \begin{cases} F(x, \theta) & j=1, \ldots, \xi \\ G(x, \psi) & j=\xi+1, \ldots, T, \end{cases} \tag{2.1}$$

where ξ is unknown, F and G are known distribution functions with possibly unknown vector parameters θ and ψ respectively, and $F(x, \theta) \neq G(x, \psi)$. We should note that (2.1) may be an alternative expression for the model

$$\mathcal{L}(X|u) = \begin{cases} F(x, \theta) & u \leq \eta \\ G(x, \psi) & u > \eta, \end{cases} \tag{2.2}$$

where observations on X are taken at values $u_1, \ldots, u_\xi \leq \eta < u_{\xi+1} < \ldots < u_T$ of an independent variable u. Estimation of ξ or η implies classification of the observations, which is one possible objective of statistical analysis. Another objective is a test of significance on ξ or η, or a confidence interval for ξ or η. When η is the parameter of interest in this latter case, our discussion of the model (2.1) is directly relevant to spacing of the u_j to meet prior requirements on power of inference.

In this section we consider properties of likelihood inference about ξ. The discussion incorporates a summary of previous work (Hinkley, 1970) together with new results on large-sample theory and approximations to distributions of likelihood ratio statistics.

2.1 Likelihood inference for known θ and ψ

Suppose for the moment that F and G have discrete or continuous densities f and g respectively, and that θ and ψ are known. Then if ξ_o is the true value of ξ, the log likelihood $L(\xi)$ of ξ under the model (2.1) satisfies

$$L(\xi) - L(\xi_o) = \begin{cases} \sum_{j=1}^{\xi-\xi_o} Y_j & (\xi_o < \xi < T) \\ \sum_{j=1}^{\xi_o-\xi} Z_j & (1 \leq \xi < \xi_o) , \end{cases} \tag{2.3}$$

where

$$\begin{aligned} Y_j &= \log\{f(X_{\xi_o+j}, \theta)/g(X_{\xi_o+j}, \psi)\} \\ Z_j &= \log\{g(X_{\xi_o+1-j}, \psi)/f(X_{\xi_o+1-j}, \theta)\} . \end{aligned} \tag{2.4}$$

Thus $L(\xi)$ generates two independent random walks each with iid increments. It is apparent from (2.4) that $E(Y_j) < 0$ and $E(Z_j) < 0$ if these expectations exist. In what follows it is convenient to assume their existence, and so we shall assume f and g to have identical support; that is, the set $\{x: f(x, \theta) = 0\}$ has zero probability under $g(x, \psi)$ and vice versa. An example where this is not the case will be described in Section 2.4.

The likelihood ratio statistics for testing $H: \xi = \xi^*$ versus the one-sided alternatives $H^-: \xi < \xi^*$ and $H^+: \xi > \xi^*$ are respectively

$$S^- = \max_{\xi<\xi^*} L(\xi) - L(\xi^*) \quad , \quad S^+ = \max_{\xi>\xi^*} L(\xi) - L(\xi^*) . \tag{2.5}$$

For the two-sided alternative the test statistic is $S = \max(S^-, S^+)$. Therefore when H is true, that is $\xi^* = \xi_o$, we can write

$$S^- = U = \max_{n \geq 1} \sum_{j=1}^{n} Z_j$$

$$S^+ = V = \max_{n \geq 1} \sum_{j=1}^{n} Y_j \,, \tag{2.6}$$

and

$$S = \max(U, V) \,.$$

The maximum likelihood estimate $\hat{\xi}$ will satisfy

$$\hat{\xi} = \begin{cases} \xi_o - n & \left(U = \sum_{j=1}^{n} Z_j > 0, \quad U > V\right) \\ \xi_o & (U \leq 0, \quad V \leq 0) \\ \xi_o + n & \left(V = \sum_{j=1}^{n} Y_j > 0, \quad V > U\right) \,, \end{cases} \tag{2.7}$$

with necessary modification if multiple maxima of $L(\xi)$ can occur. The distribution of $\hat{\xi}$ is awkward to calculate, and the inefficiency of $\hat{\xi}$ as a test statistic (Hinkley, 1971; Hinkley and Hinkley, 1970) has encouraged us to pay little attention to it in this paper.

We shall assume that ξ and $T-\xi$ are infinitely large, so that U and V are maxima of interminate random walks. The distributions of U and V converge exponentially in ξ and $T-\xi$. In terms of the model (2.2), we are assuming the u_j to become increasingly dense on either side of $u = \eta$ for u in some bounded interval.

We shall denote the distribution functions of Y and Z by $H_Y(y)$ and $H_Z(z)$, while the distribution functions of U and V will be

denoted by $P_U(u)$ and $P_V(v)$. Then since $E(Y) < 0$ and $E(Z) < 0$, both U and V are finite with probability one and their distribution functions satisfy

$$P_V(v) = \int_0^\infty P_V(y)\, dH_Y(v-y)$$
$$P_U(u) = \int_0^\infty P_U(y)\, dH_Z(u-y) . \tag{2.8}$$

We also have that

$$P_V(0) = \exp\left\{ - \sum_{n=1}^{\infty} n^{-1} \, pr\left(\sum_{j=1}^{n} Y_j > 0 \right) \right\}$$
$$P_U(0) = \exp\left\{ - \sum_{n=1}^{\infty} n^{-1} \, pr\left(\sum_{j=1}^{n} Z_j > 0 \right) \right\} \tag{2.9}$$

A result central to our discussion is the following. If there exists a positive ω_Y such that $E\{\exp(\omega_Y Y)\} = 1$, then

$$P_V(v) \sim 1 - c_Y \exp(-\omega_Y v) \qquad (v \to \infty) \tag{2.10}$$

for some constant c_Y; see Feller (1966, p. 392). Therefore if we denote the (asymptotic) test size for S^+ by $\beta_o^+(v)$, then

$$\beta_o^+(v) = 1 - P_V(v) \sim c_Y \exp(-v) , \tag{2.11}$$

because $\omega_Y = 1$ when Y is the log likelihood ratio (2.4). We should note that S^+ is stochastically increasing in $T-\xi_o$, so that the test size increases monotonically to $\beta_o^+(v)$ as $T-\xi_o$ tends to infinity. The corresponding result for asymptotic test size of S^- is

$$\beta_o^-(u) = 1 - P_U(u) \sim c_Z^* \exp(-u) .$$

In general the constant c_Y in (2.10) cannot be evaluated analytically, but a method of numerical computation due to W. M. Gentleman may be summarized as follows. Consider (2.8) in the form

$$(2.12)\quad 1-P_V(v) = 1-H_Y(v) + \int_0^A \{1-P_V(x)\}dH_Y(v-x) + c_Y \int_A^\infty e^{-x}dH_Y(v-x),$$

which for large A will be a good approximation. Now for some finite set of numbers $x_0, \ldots, x_n$ such that $0 = x_0 < x_1 < \ldots < x_n = A$, replace the first integral in (2.12) by the approximating trapezoidal sum over the interval $(x_0, x_1), \ldots, (x_{n-1}, x_n)$. Then given c_Y and A, (2.12) induces a set of linear equations for $1-P_V(x_i)$ $(i=0, 1, \ldots, n)$. Suppose we solve these equations for trial values of c_Y and A, and denote the solutions by $a_0, a_1, \ldots, a_n$. If $a_j \exp(x_j)$ is approximately constant for x_i near A, then A is suitably large. The trial value of c_Y is approximately correct if $a_n \exp(A)$ is equal to the trial value. When $a_n \exp(A)$ is not close to the trial value of c_Y, a second estimate of c_Y is made close to $a_n \exp(A)$, and the "solution-inspection" process is repeated. For a numerical example of this procedure the reader is referred to Hinkley (1970).

We should emphasize that (2.11) will provide a good approximation only for the upper tail probabilities of S^+. So for small test size α, the critical value of the statistic S^+ is approximately $\log(c_Y/\alpha)$. Some numerical values of c_Y for the normal mean-shift case are given in Section 2.4.

The classical inversion of significance tests to construct confidence intervals is non-trivial for ξ, since the log likelihood $L(\xi)$ does not decrease monotonically away from $\xi = \hat{\xi}$. Hence if a lower confidence limit ξ_ℓ is defined in the usual way to be

$$\xi_\ell = \inf\{\xi: \max_{\tau>\xi} L(\tau) - L(\xi) \leq v\} ,$$

with nominal confidence coefficient $P_V(v)$, then $pr(\xi_\ell \leq \xi_o) > P_V(v)$. In fact we find that

$$pr(\xi_\ell > \xi_o) = pr(V \geq v , U \leq V-v)$$

$$= \int_v^\infty P_U(x-v)dP_V(x) .$$

For large v, (2.10) implies that

$$pr(\xi_\ell > \xi_o) \cong \beta_o^+(v) \int_0^\infty P_U(x)\, e^{-x}dx$$

$$= \beta_o^+(v)\{P_U(0) + \int_0^\infty e^{-x}dP_U(x)\}$$

$$= \beta_o^+(v)\, P_U(0)\, \{1 + \exp \sum_{n=1}^{\infty} d_n/n)\} ,$$

where $$d_n = \int_{0+}^\infty e^{-x}\, pr(x \leq \sum_{j=1}^{n} Y_j < x+dx) .$$

This last expression follows from (2.13) of Hinkley (1970) and can also be deduced from (2.18). The point is that the lower-bounded confidence set $C_\ell = \{\xi: \max_{\tau>\xi} L(\tau) - L(\xi) \leq v\}$ does not contain all ξ greater than ξ_ℓ. Incidentally, we should observe that

$$\xi_\ell = \inf\{\xi: L(\hat{\xi}) - L(\xi) \leq v \quad \text{and} \quad \xi \leq \hat{\xi}\}.$$

For large confidence coefficient $1-\alpha$, $v \doteq \log(c_Y/\alpha)$.

Similar remarks apply to confidence intervals derived from the test statistics S^- and S.

Finally we note that the asymptotic probability of no misclassification is

$$\text{pr}(\hat{\xi} = \xi_o) = P_U(0)\, P_V(0), \tag{2.14}$$

which can be calculated from (2.9). One important characteristic of $\hat{\xi}$ is that $\hat{\xi} - \xi_o = O_p(1)$.

2.2 Power of likelihood inference for known θ and ψ

In the previous section we examined the null distributions of the likelihood ratio statistics S^- and S^+. Now suppose that $\xi^* = \xi_o - r$ with $r > 0$ and consider S^+. By (2.3) and (2.5) we see that

$$S^+ = \max(-Z_r, -Z_r - Z_{r-1}, \ldots, -\sum_{j=1}^{r} Z_j, -\sum_{j=1}^{r} Z_j + V). \tag{2.15}$$

Looking first at the case $r = 1$, (2.15) implies that

$$\beta^+_{-1}(v) = \text{pr}(S^+ > v \,|\, \xi^* = \xi_o - 1) = 1 - \int_{-v}^{\infty} P_V(v+z)\, dH_Z(z). \tag{2.16}$$

But if v is large, or if $E(Z)$ and $\text{var}(Z)$ are small, then $H_Z(z)$ will concentrate its mass well above $z = -v$ and (2.16) will be

$$\sim 1 - \int_{-\infty}^{\infty} \{1 - c_Y \exp(-v - z)\}\, dH_Z(z).$$

Therefore

$$\beta^{+}_{-1}(v) \sim \beta^{+}_{o}(v)\ E\{\exp(-Z)\}\ . \tag{2.17}$$

as $v \to \infty$ or $||F-G|| \to 0$. For general positive r a similar argument applies to the generalization of (2.16) and gives

$$\beta^{+}_{-r}(v) \sim \beta^{+}_{o}(v)\ [E\{\exp(-Z)\}]^{r}\ . \tag{2.18}$$

This expression is more familiar in the special case $g(x, \psi) = f(x, \theta+\delta)$ with $\delta \to 0$, when formal expansion of $E\{\exp(-Z)\}$ and substitution in (2.18) leads to

$$\beta^{+}_{-r}(v) \sim \beta^{+}_{o}(v)\ \{1 + r\delta'\ I_{f}(\theta)\delta\} + o(||\delta||^{2})\ . \tag{2.19}$$

Here $I_{f}(\theta)$ is the Fisher information matrix for $f(x, \theta)$.

The result for S^{-} corresponding to (2.18) is

$$\beta^{-}_{r}(u) = pr(S^{-} > u | \xi^{*} = \xi_{o}+r) \sim \beta^{-}_{o}(u)\ [E\{\exp(-Y)\}]^{r}$$

as $u \to \infty$ and $||F-G|| \to 0$ with $r > 0$.

It remains to consider S^{+} when $\xi^{*} > \xi_{o}$ and S^{-} when $\xi^{*} < \xi_{o}$. But it is immediate from (2.3) and (2.5) that

$$S^{+} \overset{P}{=} V\ (\xi^{*} > \xi_{o}) \quad \text{and} \quad S^{-} \overset{P}{=} U\ (\xi^{*} < \xi_{o})\ ,$$

from which it follows that the one-sided likelihood ratio tests are unbiased. We shall see in Section 3 that they are not uniformly most powerful.

Approximations such as (2.18) and (2.19) may be useful in the model (2.2). For suppose the spacing of the independent variable u is constant, that is $u_i - u_{i-1} = \epsilon$. If we wish to test $H: \eta = \eta^*$ against $H^+: \eta > \eta^*$ with (small) test size α and fixed power π at $\eta = \eta^* + \gamma$, then for example (2.19) indicates that we should have

$$\epsilon \doteq \frac{\gamma\{\delta' I_f(\theta)\delta\}\alpha}{\pi - \alpha} .$$

Some numerical examples of (2.18) are given in Section 2.4. We shall make use of these power approximations to describe the efficiency of alternative test statistics in Section 3.

2.3 Convergence of likelihood statistics for unknown θ and ψ

In the discussion thus far we assumed θ and ψ to be known, in order to take advantage of the random walk representation for $L(\xi)$. When θ and ψ are unknown, it is easy to verify that the asymptotic distributions of the likelihood statistics are unchanged. To demonstrate this we shall use familiar results on consistency of maximum likelihood estimates.

For simplicity let θ and ψ be one-dimensional. Then we may summarize our assumptions on F and G as follows.

Assumption 2.1. The family $\{F(x, \theta), \theta \in \Theta ; G(x, \psi), \psi \in \Psi\}$ satisfies the consistency assumptions given by Wald (1949), with the understanding that continuity conditions apply to the sub-families $\{F(x, \theta), \theta \in \Theta\}$ and $\{G(x, \psi), \psi \in \Psi\}$ if these have no intersection. Included in this assumption is the fact that F and G are either both discrete or both absolutely continuous.

Assumption 1 could be relaxed to include the more general cases discussed byKiefer and Wolfowitz (1956), for example. The proofs of results in this section can be generalized, just as Wald's consistency theorem was generalized.

With θ and ψ unknown, let $\overline{\theta}_\xi$ and $\overline{\psi}_\xi$ be the maximum likelihood estimates of θ and ψ conditional on ξ. Then the marginal log likelihood of ξ is

$$(2.20) \qquad \overline{L}(\xi) = \sum_{j=1}^{\xi} \log f(X_j, \overline{\theta}_\xi) + \sum_{j=\xi+1}^{T} \log g(X_j, \overline{\psi}_\xi) .$$

To distinguish the maximum likelihood estimates of ξ when θ and ψ are known and unknown, they will be denoted by $\hat{\xi}$ and $\overline{\xi}$ respectively. The likelihood ratio statistic for testing $H: \xi = \xi^*$ against the two-sided alternative $H_a: \xi \neq \xi^*$ is

$$\overline{S} = \max_{\xi \neq \xi^*} \overline{L}(\xi) - \overline{L}(\xi^*)$$

when θ and ψ are unknown, corresponding to S when θ and ψ are known. We are also interested in θ and ψ, whose maximum likelihood estimators are $\hat{\theta} = \overline{\theta}_{\overline{\xi}}$ and $\hat{\psi} = \overline{\psi}_{\overline{\xi}}$. The true values of θ and ψ will be denoted by θ_o and ψ_o.

Once $\hat{\theta}$ and $\hat{\psi}$ are determined, the log likelihood may be modified to

$$\tilde{L}(\xi) = \sum_{j=1}^{\xi} \log f(X_j, \hat{\theta}) + \sum_{j=\xi+1}^{T} \log g(X_j, \hat{\psi}) ,$$

and corresponding statistics $\tilde{\xi}$ and $\tilde{S}$ calculated. We shall use $\tilde{L}(\xi)$ to establish the asymptotic properties of $\overline{\xi}$, $\overline{S}$, $\hat{\theta}$ and $\hat{\psi}$.

Without loss of generality we suppose that $\xi_o = \lambda T$ for some fixed $\lambda (0<\lambda<1)$, so that asymptotic results are for $T \to \infty$. The random walk structure of $L(\xi)$ implies $\hat{\xi} - \xi_o = O_p(1)$ and $S = O_p(1)$ if $\xi^* - \xi_o = O(1)$, so the asymptotic distributions of $(\hat{\xi}, S)$, $(\overline{\xi}, \overline{S})$ and $(\tilde{\xi}, \tilde{S})$ are the same if we prove the following theorem:

Theorem 2.1. Under Assumption 1, (i) $\overline{\xi} - \hat{\xi} = o_p(1)$ and (ii) $\overline{S} - S = o_p(1)$ for $\xi^* - \xi_o = O(1)$ as $T \to \infty$. Also (iii) $\tilde{\xi} - \hat{\xi} = o_p(1)$ and (iv) $\tilde{S} - S = o_p(1)$ for $\xi^* - \xi_o = O(1)$ as $T \to \infty$.

To prove the theorem, we show first that $\overline{\xi} - \xi_o = O_p(1)$. By the definition of $\overline{L}(\xi)$ in (2.20), we have that for $r > 0$

$$\overline{L}(\xi_o+r) \leq \overline{L}(\xi_o) + \sup_{\theta\in\Theta} \sum_{i=\xi_o+1}^{\xi_o+r} \log\{f(X_i,\theta)/g(X_i,\psi_o)\}$$

$$+ \sum_{i=\xi_o+1}^{T} \log\{g(X_i,\psi_o)/g(X_i,\overline{\psi}_{\xi_o})\} + \sum_{i=\xi_o+r+1}^{T} \log\{g(X_i,\overline{\psi}_{\xi_o+r})/g(X_i,\psi_o)\}$$

$$= \overline{L}(\xi_o) + A_r + B_T + C_{r,T},$$

say. Now $B_T \leq 0$, $C_{r,T}$ is finite with probability one, and $A_r \to -\infty$ with probability one by Assumption 1. It follows that if $r_o(T) \to \infty$ as $T \to \infty$, then

$$\lim_{T\to\infty} \text{pr}\{\overline{L}(\xi_o+r) < \overline{L}(\xi_o), \quad r = r_o(T), \ldots, T-\xi_o\} = 1.$$

Similarly we find that if $r_1(T) \to \infty$ as $T \to \infty$ then

$$\lim_{T\to\infty} \text{pr}\{\overline{L}(\xi_o-r) < \overline{L}(\xi_o), \quad r = r_1(T), \ldots, \xi_o-1\} = 1.$$

But since $r_o(T)$ and $r_1(T)$ are arbitrary and $\overline{L}(\overline{\xi}) \geq \overline{L}(\xi_o)$, we deduce that

$$\overline{\xi} - \xi_o = o_p(1) . \tag{2.21}$$

Next we need to prove that $\hat{\theta}$ and $\hat{\psi}$ are consistent. For $\hat{\theta}$ this means that for any closed set $\omega \subset \Theta - \theta_o$,

$$\sup_{\theta \in \omega} \sum_{j=1}^{\overline{\xi}} \log\{f(X_j, \theta)/f(X_j, \theta_o)\} = \log o_p(1) . \tag{2.22}$$

By definition, (2.22) is less than or equal to

$$\sup_{\theta \in \omega} \sum_{j=1}^{\xi_o} \log\{f(X_j,\theta)/f(X_j,\theta_o)\} + \sup_{\theta \in \omega} \sum_{j=\xi_o+1}^{\overline{\xi}} {}^{*} \log\{f(X_j,\theta)/f(X_j,\theta_o)\} , \tag{2.23}$$

where $\sum_{j=a+1}^{b}{}^{*}$ means $\sum_{j=a+1}^{b}$ if $b > a$, and $\sum_{j=b+1}^{a}$ if $b < a$. The first term in (2.23) is $\log o_p(1)$ by Assumption 1, and (2.21) implies that the second term is $o_p(1)$. This proves (2.22) and with the corresponding argument for ψ we get

$$\hat{\theta} - \theta_o = o_p(1) \quad \text{and} \quad \hat{\psi} - \psi_o = o_p(1) . \tag{2.24}$$

We note without proof that

$$\overline{\theta}_\xi - \theta_o = o_p(1) \quad \text{and} \quad \overline{\psi}_\xi - \psi_o = o_p(1) \qquad (\xi - \xi_o = O(1)) . \tag{2.25}$$

We can now prove (i) and (ii) of Theorem 2.1. The definitions of $\overline{L}(\xi)$, $\tilde{L}(\xi)$ and $L(\xi)$ imply that for any ξ

$$\overline{L}(\overline{\xi}) - \overline{L}(\xi) \leq \overline{L}(\overline{\xi}) - \tilde{L}(\xi) \leq L(\hat{\xi}) - L(\xi) + \sum_{j=\xi+1}^{\overline{\xi}} {}^{*} \log\{f(X_j,\hat{\theta})g(X_j,\psi_o)/f(X_j,\theta_o)g(X_j,\hat{\psi})\} \tag{2.26}$$

and

$$(2.27)\quad \overline{L}(\overline{\xi}) - \overline{L}(\xi) \geq \overline{L}(\hat{\xi}) - \overline{L}(\xi) \geq \sum_{j=1}^{\hat{\xi}} \log\{f(X_j,\overline{\theta}_\xi)/g(X_j,\overline{\psi}_\xi)\} - \overline{L}(\xi)$$

$$= L(\hat{\xi}) - L(\xi) + \sum_{j=\xi+1}^{\hat{\xi}} {}^{*} \log\{f(X_j,\overline{\theta}_\xi)g(X_j,\psi_o)/f(X_j,\theta_o)g(X_j,\overline{\psi}_\xi)\} .$$

But if $\xi = \xi^* = \xi_o + O(1)$, the sums in (2.26) and (2.27) contain $O_p(1)$ terms, each of which is $o_p(1)$ by (2.24) and (2.25). It follows that for any $\epsilon > 0$ and $\xi^* = \xi_o + O(1)$,

$$(2.28)\quad \lim_{T\to\infty} \mathrm{pr}\{|\overline{L}(\overline{\xi}) - \overline{L}(\xi^*) - L(\hat{\xi}) + L(\xi^*)| < \epsilon\} = 1 .$$

A similar argument shows that for any $\epsilon > 0$ and $\xi^* = \xi_o + O(1)$,

$$(2.29)\quad \lim_{T\to\infty} \mathrm{pr}\{\overline{L}(\hat{\xi}) - \overline{L}(\xi^*) > L(\hat{\xi}) - L(\xi^*) - \epsilon\} = 1 ,$$

which together with (2.28) gives

$$\lim_{T\to\infty} \mathrm{pr}\{\overline{L}(\hat{\xi}) - \overline{L}(\xi^*) > \overline{L}(\overline{\xi}) - \overline{L}(\xi^*) - 2\epsilon\} = 1 .$$

But since $\overline{L}(\overline{\xi}) \geq \overline{L}(\hat{\xi})$, with equality if and only if $\overline{\xi} = \hat{\xi}$, part (i) of Theorem 2.1 is proved. Also (2.28) proves part (ii).

Parts (iii) and (iv) may be proved by applying the same arguments to $\tilde{L}(\xi)$ and $\tilde{\xi}$, once it is shown that $\tilde{\xi} - \xi_o = O_p(1)$. We omit the details here.

Finally we have the following result for $\hat{\theta}$ and $\hat{\psi}$.

Theorem 2.2. If Assumption 1 holds, and if f and g are such that the limiting distributions of $\sqrt{\xi_o}\,(\overline{\theta}_{\xi_o} - \theta_o)$ and $\sqrt{T-\xi_o}\,(\overline{\psi}_{\xi_o} - \psi_o)$ are normal, then $\sqrt{\xi_o}\,(\hat{\theta} - \theta_o)$ and $\sqrt{T-\xi_o}\,(\hat{\psi} - \psi_o)$ also have those limiting normal distributions.

It suffices to consider $\hat{\theta}$. From the definitions of $\hat{\theta}$ and $\overline{\theta}_{\xi_o}$ we get

$$\sum_{j=\xi_o+1}^{\overline{\xi}}{}^{*} \log\{f(X_j, \overline{\theta}_{\xi_o})/f(X_j, \hat{\theta})\} \leq \sum_{j=1}^{\xi_o} \log\{f(X_j, \hat{\theta})/f(X_j, \overline{\theta}_{\xi_o})\} \leq 0 .$$

Hence by (2.21), (2.24) and (2.25) we see that

$$\sum_{j=1}^{\xi_o} \log\{f(X_j, \hat{\theta})/f(X_j, \overline{\theta}_{\xi_o})\} = o_p(1) .$$

But the asymptotic normality of $\overline{\theta}_{\xi_o}$ implies only

$$\sum_{j=1}^{\xi_o} \log\{f(X_j, \overline{\theta}_{\xi_o})/f(X_j, \theta_o)\} = o_p(1) ,$$

so that $\hat{\theta} - \overline{\theta}_{\xi_o} = o_p(\xi_o^{-\frac{1}{2}})$. This proves the desired result.

Note that by (2.21) the asymptotic normality in Theorem 2.2 holds with ξ_o replaced by $\overline{\xi}$.

2.4 Some examples

In Sections 2.1 and 2.2 we derived some simple limiting expressions for the power and size of likelihood ratio tests. To illustrate these results we look at the normal, binomial and rectangular distributions.

When F and G are the multivariate normal distributions $MN(\theta, \Sigma)$ and $MN(\psi, \Sigma)$ respectively, so that a mean-shift takes place, we have

$$(2.30) \quad L(\xi) = \text{sample constant} + \sum_{j=1}^{\xi} (\theta-\psi)' \Sigma^{-1} (X_j - \tfrac{1}{2}\theta - \tfrac{1}{2}\psi) .$$

The random walks (2.3) each have iid increments which are $N(-d^2/2, d^2)$ with $d^2 = (\theta-\psi)' \Sigma^{-1} (\theta-\psi)$. This particular problem is discussed at length by Hinkley (1970, 1971). The approximation (2.11) for test size is very accurate for $d \leq 2$ and $\beta_o^+(v) \leq 0.05$, the maximum relative error being 0.06 at $d = 2.0$ and $\beta_o^+(v) = 0.05$. For this reason it seems useful to record the coefficients c_Y in Table 2.1. Turning to the power function approximation, we get

$$E\{\exp(-Y)\} = E\{\exp(-Z)\} = \exp(d^2) .$$

Some exact values of $\beta_{-1}^+(v)/\beta_o^+(v)$ were computed from (2.16) and numerical solution of (2.12). These are given in Table 2.2 together with the approximation $\exp(d^2)$. Evidently the approximation is excellent for $d \leq 1$ and $\beta_o^+(v) \leq 0.05$, but poor for d as large as 2 .

Table 2.1. Coefficients c_Y in (2.5) for the normal case with mean-shift distance d^2

d	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
c_Y	0.70549	0.62800	0.56030	0.49990	0.46646	0.39917	0.35735	0.32037

Table 2.2. Exact and approximate values of $\beta_{-1}^{+}(v)/\beta_{0}^{+}(v)$ for the normal case with mean-shift distance d^2

d	$\beta_0^+(v)$	.05	.02	.01	.005	.001
1.0	exact	2.72	2.72	2.72	2.72	2.72
	approx.	2.72	2.72	2.72	2.72	2.72
1.5	exact		8.50	9.20	9.45	9.49
	approx.		9.49	9.49	9.49	9.49
2.0	exact				45.50	49.50
	approx.				54.60	54.60

As a second example, consider the binomial case

$$f(x, \theta) = \binom{n}{x} \theta^x (1-)^{n-x} , \ (x=0, ., \ldots n) , \ g(x, \psi) \equiv f(x, \psi) .$$

The case $n = 1$ was examined in detail by Hinkley and Hinkley (1970), who derived simple recurrence relations for the exact (asymptotic) distributions of S^-, S^+, and $\hat{\xi}$. In the notation of Section 2.1 we have

$$Y_j = \lambda X_{\xi_o+j} - \mu(n - X_{\xi_o+j})$$

$$Z_j = -\lambda X_{\xi_o-j+1} + \mu(n - X_{\xi_o-j+1}) ,$$

where $\lambda = \log(\theta/\psi)$ and $\mu = \log\{(1-\psi)/(1-\theta)\}$. Therefore

$$E\{\exp(-Y)\} = \{\psi^2 \theta^{-1} + (1-\psi)^2 (1-\theta)^{-1}\}^n$$

(2.31)

$$E\{\exp(-Z)\} = \{\theta^2 \psi^{-1} + (1-\theta)^2 (1-\psi)^{-1}\}^n .$$

Some exact values of $\beta^+_{-1}(v)/\beta^+_o(v)$ for the case $n = 1$ are given in Table 2.3 with the corresponding values of the approximation (2.17); the nominal test sizes $\beta^+_o(v)$ in the table are not quite achieved, because S^+ has a discrete distribution. The approximation (2.17) seems to be very good when its value is less than 1.6 ; the square of (2.31) is equally good as an approximation to $\beta^+_{-2}(v)/\beta^+_o(v)$. Tables of percentage points and power of the two-sided test statistic may be found in Hinkley and Hinkley (1970).

Note from (2.18) and (2.31) that the power at $\xi^* = \xi_o + r$ is approximately determined by nr , as we should expect.

Table 2.3. Exact and approximate values of $\beta^+_{-1}(v)/\beta^+_o(v)$ for the binomial case

θ	ψ \ $\beta^+_o(v)$		.05	.02	.01
0.95	0.80	exact	1.143	1.140	1.141
		approx.	1.138	1.138	1.138
0.90	0.70	exact	1.190	1.188	1.192
		approx.	1.190	1.190	1.190
0.90	0.50	exact	1.661	1.615	1.630
		approx.	1.640	1.640	1.640
0.80	0.40	exact	1.629	1.646	1.654
		approx.	1.667	1.667	1.667

The final example is the simple uniform case

$$f(x, \theta) = 1(\theta - \tfrac{1}{2} \leq x \leq \theta + \tfrac{1}{2}) \ , \quad g(x, \psi) = f(x, \psi) \quad (\psi > \theta) \ .$$

This is a degenerate case in which distribution support changes; we note that the convergence results of Section 2.3 hold. The log likelihood $L(\xi)$ is simply

$$L(\xi) = \begin{cases} 0 & (X_1, \ldots, X_\xi \leq \theta + \tfrac{1}{2} \; ; \; X_{\xi+1}, \ldots, X_T \geq \psi - \tfrac{1}{2}) \\ -\infty & (\text{otherwise}) \end{cases}$$

if $0 < \psi - \theta < 1$. Therefore we get an interval estimate of ξ which always covers ξ_o. The distribution of the interval length is easily seen to be the convolution of two identical geometric distributions, and the average interval length is $1 + 2(\psi-\theta)^{-1}$.

3. GENERAL DISCRIMINANT APPROACH

The likelihood method of inference discussed in Section 2 has several possible faults, two of which concern us here. First, when θ and ψ are unknown the log likelihood $\overline{L}(\xi)$ can be very inconvenient in actual data analysis, since it requires calculation of the sequence of conditional estimates $\{\overline{\theta}_\xi, \overline{\psi}_\xi\}$. For example, in the binomial case where X is 0 or 1 with probabilities θ and $1-\theta$ or ψ and $1-\psi$,

$$\overline{L}(\xi) = \xi\{\overline{\theta}_\xi \log \overline{\theta}_\xi + (1-\overline{\theta}_\xi) \log(1-\overline{\theta}_\xi)\} + (T-\xi)\{\overline{\psi}_\xi \log \overline{\psi}_\xi + (1-\overline{\psi}_\xi) \log(1-\overline{\psi}_\xi)\}$$

This difficulty remains even when one parameter is known. Second, even when θ and ψ are known, the likelihood statistics and their distributional properties (Section 2.1) can be more complicated than we might wish. A specific case is the generalization of the multivariate normal example of Section 2.4 in which the covariance matrix changes as well as the mean vector. Then, rather than use the log likelihood with quadratic forms in the X_j, one might prefer to use a compound linear discriminant such as (2.30). This is especially true in the normal case, because a linear discriminant defines random walks with normally distributed increments, so that appropriate inference statistics will have known distributions. Suitable choice of linear discriminant in the normal case has been studied by Anderson and Bahadur (1962).

These remarks suggest that we consider a general class of discriminants $d(X)$ to discriminate between $F(x, \theta)$ and $G(x, \psi)$, so that

$$D(\xi) = \sum_{j=1}^{\xi} d(X_j) \tag{3.1}$$

corresponds to $L(\xi)$ or $\overline{L}(\xi)$. To make any progress we must have

$$(3.2) \qquad \int d(x)\, dF(x, \theta_o) > 0 \quad , \quad \int d(x)\, dG(x, \psi_o) < 0$$

or vice versa, which implies that $d(X)$ be a consistent discriminant. In practice this may be difficult to guarantee unless we know one of the parameters. However, a suitable $d(X)$ might be suggested by an initial look at the data.

One other important situation where $D(\xi)$ might be useful is when $F(x, \theta_o)$ is known but G is unknown. Then the likelihood cannot be used, and $d(X)$ is chosen to detect departures from $F(x, \theta_o)$. A particular example is departure from randomness.

We shall assume (3.2) to hold. Then $D(\xi)$ has the same random walk structure as $L(\xi)$, which we shall use in Sections 3.1 and 3.2 to generate results corresponding to those in Sections 2.1 and 2.2. The use of $D(\xi)$ in constructing an alternative to $\overline{L}(\xi)$ is discussed in Section 3.3. Some examples are given in Section 3.4.

3.1 Inference from the discriminant $D(\xi)$

The general discriminant $D(\xi)$ has properties that correspond directly to those of $L(\xi)$, described in Section 2.1, provided (3.2) holds. We can therefore proceed by analogy with Section 2.1.

Under the model (2.1), $D(\xi)$ satisfies

$$(3.3)\qquad D(\xi) - D(\xi_o) = \begin{cases} \sum_{j=1}^{\xi-\xi_o} Y_{D,j} & (\xi_o < \xi < T) \\ \sum_{j=1}^{\xi_o-\xi} Z_{D,j} & (1 \le \xi < \xi_o) \end{cases}$$

where $Y_{D,j} = d(X_{\xi_o+j})$ and $Z_{D,j} = -d(X_{\xi_o+1-j})$. Thus $D(\xi)$ generates two random walks each with iid increments of negative mean.

The discriminant statistics for testing $H: \xi = \xi^*$ versus the one-sided alternatives $H^-: \xi < \xi^*$ and $H^+: \xi > \xi^*$ are respectively

$$(3.4)\qquad S_D^- = \max_{\xi<\xi^*} D(\xi) - D(\xi^*) \ , \quad S_D^+ = \max_{\xi>\xi^*} D(\xi) - D(\xi^*) \ .$$

For the two-sided alternative the test statistic is $S_D = \max(S_D^-, S_D^+)$.

When H is true, so that $\xi^* = \xi_o$, we have from (3.3) and (3.4)

$$(3.5)\qquad S_D^- = U_D = \max_{n\ge 1} \sum_{j=1}^{n} Z_{D,j}$$

$$S_D^+ = V_D = \max_{n\ge 1} \sum_{j=1}^{n} Y_{D,j}$$

and

$$S_D = \max(U_D, V_D) \ .$$

The discriminant estimate $\hat{\xi}_D$, which maximizes $D(\xi)$, satisfies the equation corresponding to (2.7) with subscript D everywhere. Both S_D and $\hat{\xi}_D - \xi_o$ are $O_p(1)$. As in Section 2.1 we assume ξ_o and $T-\xi_o$ to be infinitely large, so that U_D and V_D are maxima of interminate random walks.

Let the distribution functions of Y_D and Z_D be $H_{Y_D}(y)$ and $H_{Z_D}(z)$ respectively, and denote the distribution functions of U_D and V_D by $P_{U_D}(u)$ and $P_{V_D}(v)$. Then by (3.2), the equations corresponding to (2.8) and (2.9) are immediate. The asymptotic form of $P_{V_D}(v)$ is

$$P_{V_D}(v) \sim 1 - c_{Y_D} \exp(-\omega_{Y_D} v) \qquad (v \to \infty) \tag{3.6}$$

where ω_{Y_D} is the positive solution to

$$E\{\exp(\omega_{Y_D} Y_D)\} = 1 . \tag{3.7}$$

To proceed any further we must obviously assume such solutions exist.

Assumption 3.1. The distribution functions $F(x, \theta_o)$, $G(x, \psi_o)$ and the discriminant $d(X)$ are such that (3.2) holds and (3.7) and (3.10) have positive finite solutions.

If we denote the (asymptotic) test size for S_D^+ by $\beta_o^+(v; D)$, then by (3.6) we have

$$\beta_o^+(v; D) = 1 - P_{V_D}(v) \sim c_{Y_D} \exp(-\omega_{Y_D} v) . \tag{3.8}$$

This provides an approximation for small test size, as does (2.11) for $\beta_o^+(v) = \beta_o^+(v; L)$. The corresponding result for the asymptotic test size of S_D^- is

$$\beta_o^-(u; D) = 1 - P_{U_D}(u) \sim c_{Z_D} \exp(-\omega_{Z_D} u) . \tag{3.9}$$

where ω_{Z_D} is the positive finite solution to

$$E\{\exp(\omega_{Z_D} Z_D)\} = 1 . \tag{3.10}$$

Solutions for the constants c_{Y_D} and c_{Z_D} are determined by solving the integral equations for $P_{U_D}(u)$ and $P_{V_D}(v)$ as described in Section 2.1. Note that (3.7) and (3.10) may not have explicit solutions, although in most common cases numerical solution for ω_{Y_D} and ω_{Z_D} is quite easy.

The remarks in Section 2.1 about interval estimation apply here also.

For the discriminant estimator $\hat{\xi}_D$, as with $\hat{\xi} = \hat{\xi}_L$, the asymptotic distribution is complicated. The special case of linear discriminants for univariate normal random variables with mean-shift is discussed by Hinkley (1971). Here we note that the probability of no misclassification is

$$\text{pr}(\hat{\xi}_D = \xi_o) = P_{U_D}(0)\, P_{V_D}(0) \tag{3.11}$$

which is determined by equations corresponding to (2.9).

If θ and ψ are unknown, they can be estimated by

$$\hat{\theta}_D = \overline{\theta}_{\hat{\xi}_D} \quad \text{and} \quad \hat{\psi}_D = \overline{\psi}_{\hat{\xi}_D}\,, \tag{3.12}$$

with $\overline{\theta}_\xi$ and $\overline{\psi}_\xi$ as defined in Section 2.3. These estimates correspond to $\hat{\theta}$ and $\hat{\psi}$, and have the same asymptotic properties; see Theorems 3.1 and 3.2 in Section 3.3. Hence the percentage points of S_D^+, for example, can be consistently estimated, just as those of S^+ can.

3.2 Power and efficiency of discriminant statistics

We introduced the discriminant $D(\xi)$ because it may be easier to use than the likelihood in some cases. However, simplicity may be gained at the expense of power. Therefore we need to examine the power of statistics such as S_D^+ and compare with the equivalent likelihood statistics. Here we look at approximations of the type discussed in Section 2.2.

Following the arguments leading up to (2.18) in Section 2.2, we find the (asymptotic) power of S_D^+ at $\xi^* = \xi_o - r$ is

$$\beta_{-r}^+ (v; D) \sim \beta_o^+ (v; D) \, [E \exp(-\omega_{Y_D} Z_D)]^r \tag{3.13}$$

as $v \to \infty$ or $||F-G|| \to 0$ with $r > 0$. We shall assume the right-hand side of (3.13) to exist, although this need not always be the case. The corresponding expression for power of S_D^- at $\xi^* = \xi_o + r$ is

$$\beta_r^- (v; D) \sim \beta_o^- (v; D) \, [E \exp(-\omega_{Z_D} Y_D)]^r \tag{3.14}$$

for $r > 0$.

In particular cases one can compare (3.13) with (2.18) to see how well $D(\xi)$ performs relative to $L(\xi)$, for large v or small $||F-G||$. Aside from this, there are several ways to measure relative efficiency of S_D^+, say. One measure that is particularly relevant to the model (2.2) is the analogue of Pitman efficiency expressed in terms of observation-spacing.

Assume that in (2.2) the spacing $u_{i+1} - u_i = \epsilon_D$, a constant depending on $D(\cdot)$. Now suppose we wish to test $H: \eta = \eta^*$ against $H^+: \eta > \eta^*$ with test size α and fixed power π at $\eta = \eta^* + \gamma$. Then as $\alpha \to 0$ or $||F-G|| \to 0$, we see from (3.13) that $\epsilon_D \to 0$ also. The magnitude of ϵ_D depends on $D(\cdot)$, and from (2.18) and (3.13) we get

$$\xi_D^+ = \lim_{\epsilon_L \to 0} \frac{\epsilon_D}{\epsilon_L} = \frac{\log E\{\exp(-\omega_{Y_D} Z_D)\}}{\log E\{\exp(-Z)\}} , \tag{3.15}$$

independently of α and π. In view of the examples in Section 2.4 we might expect that ξ_D^+ approximates ϵ_D/ϵ_L well for useful values of α and $||F-G||$. Note that if ϵ_D/ϵ_L is less than one, then the power function of S_D^+ is totally dominated by that of S^+.

The corresponding efficiency ξ_D^- for S_D^-, with fixed power at $\eta = \eta^* - \gamma$, is

$$\xi_D^- = \frac{\log E\{\exp(-\omega_{Z_D} Y_D)\}}{\log E\{\exp(-Y)\}} .$$

For the two-sided test statistic $S_D = \max(S_D^-, S_D^+)$ the situation is a little more complicated. Without going into details, it can be shown that if S_D and $S = S_L$ have equal powers at $\eta = \eta_o + \gamma_1$ and $\eta = \eta_o - \gamma_2$, with lower bounds π_1 and π_2 respectively, then the required spacings ϵ_D and ϵ_L satisfy

$$\lim_{\alpha \to 0} \frac{\epsilon_D}{\epsilon_L} = \xi_D = \min(\xi_D^-, \xi_D^+) .$$

One special class of interesting problems is that of location shift with $F(x, \theta) = F(x - \theta)$ and $G(x, \psi) = F(x - \psi)$. Here a natural class of discriminators $D(\xi)$ is the linear class with

$$d(X) = \ell' (X - a) \ .$$

For simplicity we shall only consider the univariate case, and without loss of generality we assume $\theta > \psi$ so that $d(X) = X-a$. By (3.2), then, we must have $\theta > a > \psi$. The random walk increments Y_D and Z_D defined in (3.3) have distribution functions $F(y - \psi + a)$ and $F(-z - \theta + a)$ respectively. If we now define $\rho(\lambda)$ to be the positive solution of

$$\begin{aligned} &\int \exp(\omega x)\ dF(x) = \exp(\omega\lambda) && (\lambda > 0) \\ &\int \exp(-\omega x)\ dF(x) = \exp(-\omega\lambda) && (\lambda < 0) \ , \end{aligned} \tag{3.16}$$

then we find from (3.7) and (3.10) that

$$\omega_{Y_D} = \rho(a-\psi) \quad , \quad \omega_{Z_D} = \rho(a-\theta) \ .$$

It follows that

$$\begin{aligned} E\{\exp(-\omega_{Y_D} Z_D)\} &= \exp\{(\theta-\psi)\ \rho(a-\psi)\} \\ E\{\exp(-\omega_{Z_D} Y_D)\} &= \exp\{(\theta-\psi)\ \rho(a-\theta)\} \ . \end{aligned} \tag{3.17}$$

Substitution from (3.17) into (3.15) shows the spacing efficiency $\xi_{\rho_D^+}$ for S_D^+, for example, to be

$$(3.18) \qquad \mathcal{E}_D^+ = \frac{(\theta-\psi)\,\rho(a-\psi)}{\log\{\int f(x)/f(x+\psi-\theta)\}dF(x)} .$$

If $a \neq (\theta+\psi)/2$, it is possible for (3.18) to be greater than one. An example is given in Section 3.4. When $a = (\theta+\psi)/2$ we have the normal theory, or least-squares, discriminant $d(X) = X - (\theta+\psi)/2$. A little calculation shows that $\mathcal{E}_D^+$ is the efficiency of least-squares estimation of θ, as we would expect; cf. (2.19).

We should stress that these results are limiting results, and rely on the existence of quantities such as ω_{Y_D}. Some specific examples are given in Section 3.4 to illustrate the usefulness, and limitations, of the results as practical approximations.

Rather than look at properties of test statistics based on $D(\xi)$, we might be more interested in how well the data can be classified. This might involve the comparison of (3.11) and (2.14), the probabilities of correct classification. We have not derived any suitable limiting results for such comparisons, but for the location-shift problem one would expect the best linear discriminant to be that which best discriminates in the classical single-observation discriminant problem. An example is given in Section 3.4.

3.3 An asymptotically efficient two-stage procedure

Suppose that at least one of θ and ψ is unknown, and that we have used $D(\xi)$ in preference to $\bar{L}(\xi)$ in order to estimate and make inference about ξ. Then our estimate $\hat{\xi}_D$ defines the estimates $\hat{\theta}_D$

and $\hat{\psi}_D$, which we may have used to estimate inference properties. Another use for $\hat{\theta}_D$ and $\hat{\psi}_D$ is in construction of the pseudo-likelihood

$$\tilde{L}_D(\xi) = \sum_{j=1}^{\xi} \log f(X_j, \hat{\theta}_D) + \sum_{j=+1}^{T} \log g(X_j, \hat{\psi}_D)$$

$$= \text{sample constant} + \sum_{j=1}^{\xi} \log\{f(X_j, \hat{\theta}_D)/g(X_j, \hat{\psi}_D)\} .$$

In a situation where $L(\xi)$ is not difficult to work with, use of $D(\xi)$ and then $\tilde{L}_D(\xi)$ might be preferable to the use of $\bar{L}(\xi)$. Of course, if $\hat{\theta}_D$ and $\hat{\psi}_D$ indicate that $D(\xi)$ is very efficient, then $\tilde{L}_D(\xi)$ would not be needed.

Let the pseudo-likelihood statistics corresponding to $\hat{\xi}$ and S be denoted by $\tilde{\xi}_D$ and $\tilde{S}_D$. Then $\tilde{\xi}_D$ and S_D are asymptotically efficient (relative to $\hat{\xi}$ and S) by the following theorem.

Theorem 3.1. If Assumption 1 and (3.2) hold, then as $T \to \infty$, (i) $\hat{\theta}_D - \theta_o$ and $\hat{\psi}_D - \psi_o$ are $o_p(1)$, (ii) $\tilde{\xi}_D - \hat{\xi} = o_p(1)$ and (iii) $\tilde{S}_D - S = o_p(1)$ if $\xi^* - \xi_o = O(1)$.

These results may be proved in much the same way as Theorem 2.1, using the fact that $\hat{\xi}_D - \xi_o = O_p(1)$. We omit the details here. A stronger result for $\hat{\theta}_D$ and $\hat{\psi}_D$, corresponding to Theorem 2.2, is

Theorem 3.2. If the conditions of Theorem 3.1 hold, then $\sqrt{\xi_o}\,(\hat{\theta}_D - \theta_o)$ and $\sqrt{T-\xi_o}\,(\hat{\psi}_D - \psi_o)$ have the same limiting normal distributions as $\sqrt{\xi_o}\,(\bar{\theta}_{\xi_o} - \theta_o)$ and $\sqrt{T-\xi_o}\,(\bar{\psi}_{\xi_o} - \psi_o)$. The proof of Theorem 2.2 can be applied with $\hat{\xi}_D$ replacing $\bar{\xi}$. Note that Theorem 3.2 holds if ξ_o is replaced by $\tilde{\xi}_D$ or $\hat{\xi}_D$ since the relative errors in these estimates are $o(1)$.

Some empirical results on the use of $\tilde{L}_D(\xi)$ in finite samples from univariate normal distributions are described by Hinkley (1971, Section 3.2).

A more extreme situation where the pseudo-likelihood might be of use is the case of unknown F and (or) G. Given the estimate $\hat{\xi}_D$ from $D(\xi)$, it is possible to obtain a consistent estimate of $\ell(X_j) = \log\{f(X_j, \theta_o)/g(X_j, \psi_o)\}$ for smooth densities f and g; for example, estimates can be based on Parzen density estimates (Parzen, 1962). If the consistent estimate for $\ell(X_j)$ were $\hat{\ell}_D(X_j)$, then the pseudo-log likelihood

$$\hat{L}_D(\xi) = \sum_{j=1}^{\xi} \hat{\ell}_D(X_j)$$

would generate statistics asymptotically equivalent to $\hat{\xi}$ and S. In practice one might restrict F and G to some suitable finite class among which discrimination is possible. For example, the class might contain the normal, gamma, log normal and Cauchy distributions. The appropriate generalization of Theorem 3.1 could presumably be proved without much additional effort. One useful area of application might be in models for departure from randomness, where the alternative to randomness is not specified exactly. Then $G(x, \psi_o)$ would be estimated or selected conditional on $\hat{\xi}_D$.

3.4 Some examples

The first example is the Laplace mean-shift case where

$$f(x, \theta) = \tfrac{1}{2} \exp(-|x-\theta|) \ , \quad g(x, \psi) = f(x, \theta-2\Delta) \ .$$

The likelihood random walk increments Y and Z defined at (2.4) both have the distribution function

$$(3.19) \qquad H_Y(y) = \begin{cases} 0 & (y < -2\Delta) \\ \tfrac{1}{2} & (y = -2\Delta) \\ 1 - \tfrac{1}{2} \exp(-\tfrac{1}{2}y-\Delta) & (-2\Delta < y \leq 2\Delta) \\ 1 & (2\Delta \leq y) \end{cases} ,$$

and hence

$$(3.20) \qquad E\{\exp(-Y)\} = E\{\exp(-Z)\} = \frac{2}{3} \exp(2\Delta) + \frac{1}{3} \exp(-4\Delta) \ .$$

Now suppose that we use the linear discriminant with

$$d(X) = X - \theta - \Delta \ .$$

Then from Feller (1966, Chapter 12) we have the exact result that

$$(3.21) \qquad \beta_0^+(v; D) = \beta_0^-(v; D) = (1-\omega_{Y_D}) \exp(-\omega_{Y_D} v) \qquad (v > 0)$$

and the equation (3.7) becomes

$$1-\omega^2 = \exp(-\omega\Delta) \ .$$

If we write $\omega_{Y_D} = \omega_{Z_D} = \omega(\Delta)$, then we get

(3.22) $$\log E\{\exp(-\omega_{Y_D} Z_D)\} = 2\Delta\omega(\Delta) .$$

Table 3.1 gives some values of the spacing efficiency $\xi_D (= \xi_D^- = \xi_D^+)$ calculated from (3.15), (3.20) and (3.22). These values are quite low, but the simple exact distribution (3.20) contrasts sharply with the corresponding likelihood distributions, which involve solution of (2.8) with $H_Y(y)$ given by (3.19).

Table 3.1. Spacing efficiency ξ_D for $d(X) = X-\theta-\Delta$ in the Laplace case with mean-shift 2Δ .

Δ	0+	0.1	0.2	0.3	0.4	0.5
ξ_D	0.50	0.54	0.58	0.62	0.67	0.73

As a second example consider the univariate normal case with mean and variance changing. That is, let

$$f(x, \theta) = \phi(X) \quad \text{and} \quad g(x, \psi) = \lambda\phi\{\lambda(X+2\Delta)\} .$$

The likelihood increments Y and Z are quadratic in X , and simple calculation shows that

(3.23) $$E\{\exp(-Z)\} = (2\lambda^2 - \lambda^4)^{-\frac{1}{2}} \exp\{4\lambda^2 \Delta^2/(2-\lambda^2)\} ,$$

which is only valid for $\lambda < \sqrt{2}$. Thus the approximation (2.18) only exists for $\lambda < \sqrt{2}$. For the linear discriminant $d(X) = X - \Delta$, a simple calculation via (3.7) gives

(3.24) $$E\{\exp(-\omega_{Y_D} Z_D)\} = \exp\{2\Delta^2 \lambda^2(1+\lambda^2)\} .$$

When $\Delta = 1$, the spacing efficiency ξ_D^+ given by (3.15), (3.23) and (3.24) has values 0.95 and 0.95 for $\lambda = 0.9$ and $\lambda = 1.1$. Of course $d(X) = X - \Delta$ is the likelihood discriminant when $\lambda = 1$, and so should have high efficiency for λ close to 1 .

In the normal case with constant variance ($\lambda = 1$ in the previous example), the likelihood $L(\xi)$ uses the linear discriminant $d(X) = X - \Delta$; see Section 2.4. Consider the generalization $d(X) = X - \delta$, for which Y_D and Z_D in (3.3) are $N(-\delta, 1)$ and $N(-2\Delta+\delta, 1)$ respectively. Then we get

$$E\{\exp(-\omega_{Y_D} Z_D)\} = \exp\{4\Delta(2\Delta-\delta)\} \qquad (0 < \delta < 2\Delta)$$

Hence for $0 < \delta < \Delta$, S_D^+ is <u>more</u> powerful than the likelihood statistic S^+ , while S_D^- is <u>less</u> powerful than S^- ; the reverse is true for $\Delta < \delta < 2\Delta$. For two-sided tests, however, the likelihood statistic appears preferable: some numerical results are given by Hinkley (1971). Notice that

$$\xi_D = \min\left(\frac{\delta}{\Delta}\ ,\ 2 - \frac{\delta}{\Delta}\right) \leq 1 \ .$$

The case $\delta = \Delta$ also gives the lowest misclassification probability for this problem.

4. GENERAL REMARKS

One feature of the change-point problem is the complicated nature of the distribution theory for inferential statistics. The approximations derived in this paper for likelihood ratio may simplify data analysis, and the approximations discussed in Section 3.3 will give some idea as to efficiency loss.

The two-stage procedure mentioned in Section 3.4 provides the opportunity to clean up an initial identification of the change-point, as it were. Its value lies in circumventing calculation of a complicated likelihood function with all the conditional estimates of parameters. In practice initial estimates of θ and ψ might be made by omitting the data near the change-point, if such a judgment is possible.

One could regard the linear discriminant as a non-distributional equivalent of the likelihood for location-shift problems. The results of Section 3 cover certain aspects of the discriminant, but not the important one of robustness. We have not investigated this very much, but calculations are easy for one special case. Suppose we use $d(X) = X - \theta - \Delta$, assuming X to be normal with mean-shift 2Δ and variance 1, and carry out a two-sided test of $H_0: \xi = \xi_0$ with assumed size 0.050. If in fact X has a Laplace distribution with mean-shift 2Δ and variance 1, the true rejection probability is 0.057. For assumed size 0.010, the true size is 0.015. (The true distribution of S_D in this case follows from calculations for the Laplace example in Section 3.4.) These calculations indicate reasonable robustness

against long-tailed symmetric distributions. The effect of asymmetry and dependence between observations are probably most conveniently determined by Monte Carlo studies, which we have not undertaken.

Models with more than one change-point are considerably more difficult to analyze, although the formal theory of likelihood inference generalizes quite easily. If the number of change-points is known, and large numbers of observations are taken between each one, one might be able to break the data into segments each with one change. For multiple location-shifts, the linear discriminant is appropriate because it can be used as a sequential detector (Page, 1954; Hinkley, 1971).

When the number of change-points is not known, sequential analysis with the discriminant is still appropriate but not necessarily good. Sclove, in an unpublished Stanford University Technical Report, has considered the use of finite moving averages. This type of analysis was indicated by Chernoff and Zacks (1964), who introduced prior distributions on the change-points. One relevant class of models to consider is that where $p \geq 2$ populations can generate observations and the sequence of populations forms a Markov chain, possibly with diagonal elements of the transition probability matrix close to unity.

REFERENCES

[1] Anderson, T. W. and Bahadur, R. R. (1962), "Classification into two multivariate normal distributions with different covariance matrices", Ann. Math. Statist. 33, 420-31.

[2] Chernoff, H. and Zacks, S. (1964), "Estimating the current mean of a normal distribution which is subjected to changes in time", Ann. Math. Statist. 35, 999-1018.

[3] Feller, W. (1966), An Introduction to Probability Theory and its Applications, Vol. 2. New York: Wiley.

[4] Hinkley, D. V. (1970), "Inference about the change-point in a sequence of random variables", Biometrika 57, 1-17.

[5] Hinkley, D. V. (1971), "Inference about the change-point from cumulative sum tests", Biometrika 58 (to appear).

[6] Hinkley, D. V. and Hinkley, E. A. (1970), "Inference about the change-point in a sequence of binomial variables", Biometrika 57, 477-88.

[7] Kiefer, J. and Wolfowitz, J. (1956), "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters", Ann. Math. Statist. 27, 887-906.

[8] Page, E. S. (1954), "Continuous inspection schemes", Biometrika 41, 100-114.

[9] Parzen, E. (1962), "On estimation of a probability density function and mode", Ann. Math. Statist. 33, 1065-76.

[10] Wald, A. (1949), "Note on the consistency of the maximum likelihood estimate", Ann. Math. Statist. 20, 595-601.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) DEPT. OF STATISTICS STANFORD UNIVERSITY STANFORD, CALIFORNIA	2a. REPORT SECURITY CLASSIFICATION 2b. GROUP

3. REPORT TITLE

SOME GENERAL RESULTS ON TIME-ORDERED CLASSIFICATION

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5. AUTHOR(S) (Last name, first name, initial)

HINKLEY, D.V.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
July 30, 1971	37	10

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-67-A-0112-0030 b. PROJECT NO. NR-042-034 c. d.	#4 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. AVAILABILITY/LIMITATION NOTICES

Reproduction in whole or in part is permitted for any purpose of the United States Government

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research Arlington, VA

13. ABSTRACT

The paper examines some properties of inference about the point in a sequence of random variables at which a distribution change occurs. Three particular aspects are considered: asymptotic theory, approximations to test power, and use of discriminant functions other than the likelihood ratio. The discussion is illustrated by some examples.

DD FORM 1 JAN 64 1473

UNCLASSIFIED
Security Classification

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Discriminant Change-Point						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.

DD FORM 1 JAN 64 1473 (BACK)

Unclassified

Security Classification