Math. Proc. Camb. Phil. Soc. (1979), **86**, 85
Printed in Great Britain

On sequential versions of the generalized likelihood ratio test

BY ANDREW D. BARBOUR
Gonville and Caius College, Cambridge

(*Received* 6 *November* 1978)

Abstract. It is shown that the Wilks large sample likelihood ratio statistic λ_n, for testing between composite hypotheses $\Theta_0 \subset \Theta_1$ on the basis of a sample of size n, behaves as n varies like a diffusion process related to an equilibrium Ornstein-Uhlenbeck process, whenever the null hypothesis is true. This fact is used to construct large sample sequential tests based on λ_n, which are the same whatever the underlying distributions. In particular, the underlying distributions need not belong to an exponential family.

The classical weak convergence theory of partial sums of independent identically distributed random variables concentrates on the convergence of $S_{[nt]}/\sigma\sqrt{n}$ to the Wiener process. In some respects, this is a rather unnatural way of formulating the result, since one is more usually interested in the behaviour of the sequence $S_n/\sigma\sqrt{n}$ than that of $S_{[nt]}/\sigma\sqrt{n}$ as t varies. A functional limit theorem for $S_n/\sqrt{n}$ is, however, easy to deduce from the classical theorem, and can be expressed in many ways, one of which is as follows.

Let $(X_n)_{n\geqslant 1}$ be a sequence of independent identically distributed random variables with mean zero and variance σ^2, and set $S_n = \sum_{j=1}^n X_j$. Let $D[0,\infty)$ be the space of all right continuous functions with left limits on $[0,\infty)$, and D' the subspace consisting of those functions x which also satisfy

$$\sup_{t\geqslant 0} |x(t)|/\sqrt{\log(t\vee 3)} < \infty.$$

Define a metric m' on D' by taking $m'(x,y)$ to be the infimum of those $\epsilon > 0$ for which there exists a continuous and strictly increasing real function λ with $\lambda(0) = 0$ such that

$$\sup_{t\geqslant 0} |x(t) - y(\lambda(t))|/\sqrt{\log(t\vee 3)} < \epsilon$$

and

$$\sup_{t+s\geqslant 0} |\log[(\lambda(t)-\lambda(s))/(t-s)]| < \epsilon.$$

Finally, set $Y_N(u) = S_{[Ne^u]}/\sigma\sqrt{(Ne^u)}$ for each $u \geqslant 0$.

THEOREM 1. *As $N\to\infty$, $Y_N \Rightarrow Y$ in (D', m'), where Y is an equilibrium Ornstein-Uhlenbeck diffusion process with drift coefficient $\alpha(x) = -x/2$ and infinitesimal variance $\beta(x) = 1$.*

Proof. The theorem follows from the analogue in $D[0,\infty)$ of Theorem 2 of Müller (6), using the fact that the map $\phi: D_2 \to D': \phi(x)(u) = e^{-u/2}x(e^u)$ is continuous. Here, D_2 is the subspace of $D[0,\infty)$ consisting of those functions for which

$$\limsup_{t\to\infty} |x(t)|/\sqrt{(t \log\log t)} < \infty,$$

endowed with the appropriate metric. ∎

The description of $(S_n/\sigma\sqrt{n})_{n\geq 1}$ given by Theorem 1 is rather interesting. Y is a strictly stationary process, with $Y(t)$ distributed as a standard normal random variable. However, because $Y_N(u)$ corresponds to $S_n/\sigma\sqrt{n}$ with $n = Ne^u$, the sequence $(S_n/\sigma\sqrt{n})_{n\geq N}$ behaves like $[Y(\log(n/N))]_{n\geq N}$, so that, as measured in n-time, the fluctuations get progressively slower as n increases. Qualitatively, the picture is attractive: quantitatively, however, the Ornstein-Uhlenbeck process is somewhat more difficult to work with than the Wiener process.

As an illustration of the use of such a description of $(S_n/\sigma\sqrt{n})$, consider the generalized likelihood ratio statistic

$$\Lambda_n = \sup_{\theta\in\Theta_1} L_n(X^{(n)};\theta)/\sup_{\theta\in\Theta_0} L_n(X^{(n)};\theta),$$

where $X^{(n)}$ is a sample of n independent identically distributed random vectors $(X_i)_{i=1}^n$ from a distribution on R^m with density $f(x;\theta)$, and where $\Theta_0 \subset \Theta_1$. Wilks' theorem states that, under suitable regularity conditions, $2\log\Lambda_n$ is approximately distributed as χ_d^2 as n gets large, where $d = t_1 - t_0 = \dim\Theta_1 - \dim\Theta_0$, whenever θ^*, the true value of θ, belongs to Θ_0; and is stochastically larger when $\theta^* \in \Theta_1\backslash\Theta_0$. This result is used to provide a widely useful fixed sample test of H_0: $\theta^* \in \Theta_0$ against H_1: $\theta^* \in \Theta_1\backslash\Theta_0$. What about sequential analogues?

The main line of development stems from a paper by Schwarz (7). The densities $f(x;\theta)$ are assumed to come from an exponential family, so that the log-likelihood depends on $X^{(n)}$ only through $S_n = \Sigma_{j=1}^n X_j$. Sequential tests are then derived, by analogy with the usual sequential probability ratio test, by stopping when S_n first reaches an appropriate boundary. There are, however, difficulties with this procedure when testing contiguous or nested hypotheses, since, for certain values of the alternative, the average sample number can become large, as observed, for example, by Bechhofer (2).

A more natural approach is to consider tests of the form 'stop when $2\log\Lambda_n$ gets large'. This has been suggested, for instance, by Armitage (1), and some large deviation results associated with such a test, again in the context of exponential families, have been derived by Woodroofe (9). Here, we start by showing that, under local regularity conditions on the family of densities of the sort given in Cramér (4), Chapter 33.3, there is a process $U_{(d)}$ which approximates the sequence $(2\log\Lambda_n)_{n\geq N}$ as N gets large whenever $\theta^* \in \Theta_0$, and which does not depend on f. Large sample tests can then be constructed, and significance levels tabulated, by using the properties of $U_{(d)}$.

In proving convergence to $U_{(d)}$, we essentially follow the asymptotic theory developed in detail for fixed sample tests (albeit in the more general context of Markov chains) in Billingsley (3), noting such modifications as are required. The main assumptions are that, for any i, j and k,

(1) For almost all x, the derivatives $f_i(x;\theta)$, $f_{ij}(x;\theta)$ and $f_{ijk}(x;\theta)$ exist and are continuous in θ throughout Θ_1, where f_i denotes $\partial f/\partial\theta_i$, etc.

(2) For any $\theta \in \Theta_1$, there exists a neighbourhood N of θ such that

$$\int_{R^m} \sup_{\theta'\in N} |f_i(x;\theta')|\,dx < \infty,$$

$$\int_{R^m} \sup_{\theta'\in N} |f_{ij}(x;\theta')|\,dx < \infty,$$

$$E_\theta[\sup_{\theta'\in N} |g_{ijk}(X_1;\theta')|] < \infty,$$

where $g(x;\theta) = \log f(x;\theta)$.

(3) For any $\theta \in \Theta_1$,

$$E_\theta[|g_i(X_1;\theta)|^2] < \infty,$$

and the matrix $\sigma(\theta)$ defined by

$$\sigma_{ij}(\theta) = E_\theta[g_i(X_1;\theta)g_j(X_1;\theta)]$$

is non-singular.

(4) $\Theta_0 \subset \Theta_1$ admits of a coordinate system $(\phi_1,\ldots,\phi_{t_0})$ with respect to which it is an open subset of R^{t_0}, and for which the injection map $\theta = \theta(\phi)$ is thrice continuously differentiable, and such that the matrix $K(\phi)$ with components

$$K_{ij}(\phi) = \partial\theta_i(\phi)/\partial\phi_j \quad (1 \leq i \leq t_1,\ 1 \leq j \leq t_0)$$

has rank t_0 for all ϕ such that $\theta(\phi) \in \Theta_0$.

Under these assumptions, it follows, much as in theorem 2·1 of Billingsley (3), but using the strong law of large numbers in place of a weak law, that there exists an essentially unique sequence $\hat\theta_n = \hat\theta_n(X^{(n)})$ of random vectors in Θ_1 such that $\lim_{n\to\infty}\hat\theta_n = \theta^*$ a.s. and such that each $\hat\theta_n$ is a local maximum of $L_n(X^{(n)},\theta)$. This in itself does not guarantee that the global maxima of $L_n(X^{(n)};\theta)$ form a consistent sequence of estimators of θ^*, though conditions for this to be so can be found in Wald (8). However, in what follows it is assumed that it is possible to distinguish the consistent sequence $(\hat\theta_n)$, and that $\hat\theta_n$ is used to compute the generalized likelihood ratio statistic.

THEOREM 2. *Suppose that $\theta^* \in \Theta_0$, and that $(\hat\theta_n)_{n\geq 1}$ and $(\hat\theta_{n0})_{n\geq 1}$ are the consistent sequences of local likelihood maxima in Θ_1 and Θ_0 respectively. Let*

$$\lambda_n = 2\log\{L_n(X^{(n)};\hat\theta_n)/L_n(X^{(n)};\hat\theta_{n0})\},$$

and define the process U_N by $U_N(u) = \lambda_{[Ne^u]}$. Then, for any $T > 0$, $U_N \Rightarrow U_{(d)}$ in $D[0,T]$ as $N \to \infty$, where $U_{(d)}$ is the equilibrium diffusion process with drift coefficient $\alpha(x) = -x+d$ and infinitesimal variance $\beta(x) = 4x$. In particular, $U_{(d)}$ can be represented as $\Sigma_{j=1}^d Y_j^2$, where $(Y_j)_{j=1}^d$ are independent copies of the Ornstein-Uhlenbeck process Y of Theorem 1.

Proof. A simple modification of Theorem 2·2 of Billingsley (3) shows that

$$2\log\{L_n(X^{(n)};\hat\theta_n)/L_n(X^{(n)};\theta^*)\} = n^{-1}\sum_{i,j=1}^{t_1} S_i(n)S_j(n)\sigma_{ij}^{-1}(\theta^*) + \epsilon_n, \tag{1}$$

where

$$S_j(n) \equiv \Sigma_{r=1}^n g_j(X_r;\theta^*) \quad (1 \leq j \leq t_1),$$

and where, for all $\delta > 0$ and $c > 1$,

$$\lim_{N\to\infty} P[\sup_{N\leq n\leq cN} |\epsilon_n| > \delta] = 0.$$

To obtain this last statement, a slight strengthening of Billingsley's conclusion, a multi-dimensional analogue of Theorem 1 is used to show that, for any $T > 0$, y_N defined by

$$y_N(u) = (Ne^u)^{-\frac{1}{2}} \{S_1([Ne^u]), \ldots, S_{t_1}([Ne^u])\}$$

converges in $\{D[0,T]\}^{t_1}$. Then, for any $\eta_n = \eta_n(X^{(n)})$ such that $\lim_{n\to\infty} \eta_n = 0$ a.s., it follows that, for all $\delta > 0$ and $c > 1$,

$$\lim_{N\to\infty} P[\sup_{N\leqslant n\leqslant cN} |\eta_n y_N(\log(n/N))| > \delta] = 0.$$

The remaining modification is immediate.

Applying this result also with Θ_0 as the parameter space, it follows, as in the proof of Theorem 3·1 of Billingsley(3), that

$$2\log\{L_n(X^{(n)}; \hat{\theta}_{n0})/L_n(X^{(n)}; \theta^*)\} = n^{-1} \sum_{i,j=1}^{t_1} S_i(n) S_j(n) \{K(K^T\sigma K)^{-1} K^T\}_{ij} + \epsilon_n', \tag{2}$$

where $\sigma = \sigma(\theta^*)$, $K = K(\phi^*)$ where $\theta(\phi^*) = \theta^*$, and, for all $\delta > 0$ and $c > 1$,

$$\lim_{N\to\infty} P[\sup_{N\leqslant n\leqslant cN} |\epsilon_n'| > \delta] = 0.$$

Now, if B is any non-singular $t_1 \times t_1$ matrix such that $B\sigma B^T = I$, and if

$$Z_N(t) = N^{-\frac{1}{2}} BS([Nt]),$$

it follows from the multivariate version of Donsker's theorem that $Z_N \Rightarrow (W_1, \ldots, W_{t_1})$ in $\{D[0,e^T]\}^{t_1}$, where $(W_j)_{j=1}^{t_1}$ are independent standard Brownian motions. Furthermore, the symmetric matrix

$$(B^{-1})^T \sigma^{-1} B^{-1} - (B^{-1})^T K(K^T\sigma K)^{-1} K^T B^{-1} = I - (B^{-1})^T K(K^T\sigma K)^{-1} K^T B^{-1}$$

is easily seen to be idempotent with trace $t_1 - t_0$. Hence, and by Theorem 1,

$$e^{-u} Z_N^T(e^u) \{(B^{-1})^T \sigma^{-1} B^{-1} - (B^{-1})^T K(K^T\sigma K)^{-1} K^T B^{-1}\} Z_N(e^u)$$

converges in $D[0,T]$ as $N \to \infty$ to the process $U_{(d)}$ defined in the statement of the theorem. Subtracting (2) from (1) and expressing S in terms of Z, the theorem follows easily. ∎

There are a variety of stopping rules that could be chosen for testing H_0 against H_1 using λ_n. Here, two very simple extensions of the usual fixed sample test are considered in which an upper limit N_1 is imposed upon the number of observations to be taken: the first strategy has the form:

$S1$: reject H_0 if $\max_{N_0\leqslant n\leqslant N_1}\{\lambda_n - A(\alpha, N_1/N_0; d)\} > 0$, and accept H_0 otherwise.

and the second the form

$S2$: reject H_0 if $\max_{n\leqslant N_1}\{\lambda_n - d - n^{-1}N_1 A(\alpha, d)\sqrt{2d}\} > 0$, and accept H_0 otherwise.

In either case, A is chosen to give an appropriate asymptotic size α, and is computed using the statistics of $U_{(d)}$: some selected values of A are presented in Tables 1 and 2. The values for $d \to \infty$ are computed from the limit result

$$(U_{(d)} - d)/\sqrt{(2d)} \Rightarrow Y_2 \quad \text{as} \quad d \to \infty,$$

where $Y_2(u) \equiv Y(2u)$, and Y is the equilibrium Ornstein-Uhlenbeck process of Theorem 1.

Table 1. *Approximate values of* $c = c(\alpha, k; d)$ *such that, on* H_0, $P[\max_{N\leqslant n\leqslant kN} (\lambda_n - d)/\sqrt{(2d)} \geqslant c] = \alpha$. *In test* $S1$, $A(\alpha, k; d) = d + c(\alpha, k; d)\sqrt{(2d)}$

$k =$	2		4		8	
$\alpha =$	10 %	5 %	10 %	5 %	10 %	5 %
$d = 1$	3·1	3·7	3·3	4·4	3·6	4·9
2	2·8	3·4	3·05	4·05	3·4	4·3
3	2·6	3·15	2·9	3·85	3·25	4·0
4	2·5	3·0	2·8	3·65	3·1	3·8
5	2·4	2·95	2·75	3·5	3·0	3·65
6	2·3	2·9	2·7	3·4	2·95	3·6
7	2·3	2·85	2·7	3·35	2·9	3·5
8	2·3	2·85	2·65	3·25	2·85	3·45
9	2·25	2·8	2·65	3·2	2·85	3·4
10	2·25	2·8	2·6	3·15	2·8	3·35
11	2·25	2·8	2·6	3·1	2·8	3·3
12	2·25	2·75	2·55	3·1	2·8	3·25
13	2·2	2·75	2·55	3·05	2·75	3·25
14	2·2	2·7	2·55	3·0	2·75	3·2
15	2·2	2·7	2·5	3·0	2·75	3·2
16	2·2	2·7	2·5	3·0	2·7	3·15
17	2·2	2·7	2·5	2·95	2·7	3·15
18	2·2	2·7	2·5	2·95	2·7	3·1
19	2·15	2·65	2·45	2·9	2·7	3·1
20	2·15	2·65	2·45	2·9	2·7	3·1
∞	2·1	2·5	2·4	2·8	2·55	3·0

Table 2. *Approximate values of* $A = A(\alpha, d)$ *such that, on* H_0, $P[\max_{1\leqslant n\leqslant N} (n/N)(\lambda_n - d)/\sqrt{(2d)} \geqslant A] = \alpha$

$\alpha =$	10 %	5 %	$\alpha =$	10 %	5 %
$d = 1$	2·1	3·0	$d = 11$	1·8	2·2
2	2·05	2·85	12	1·8	2·2
3	2·0	2·7	13	1·8	2·15
4	2·0	2·6	14	1·8	2·1
5	1·95	2·5	15	1·8	2·1
6	1·9	2·4	16	1·75	2·1
7	1·9	2·4	17	1·75	2·1
8	1·85	2·3	18	1·75	2·1
9	1·85	2·3	19	1·75	2·1
10	1·8	2·2	20	1·75	2·1

For large d, A (10 %) = 1·64 and A (5 %) = 1·96.

The rejection regions are sensible in that, if $\theta^* \in \Theta_1 \backslash \Theta_0$, λ_n is stochastically larger than its null hypothesis distribution. Indeed, a similar analysis based on Taylor's series shows that λ_n, in addition to stochastic fluctuation of order 1, has a component which, for fixed θ^*, increases linearly with n, in marked contrast to the activity of the null hypothesis random fluctuation, which evolves as $\log n$. This can informally be seen by observing that, for large n, $n^{-1} \log L_n(X^{(n)}; \theta) \sim E_{\theta^*} g(X_1; \theta)$, which attains its maximum in θ at $\theta = \theta^*$, and that, for θ near θ^*,

$$E_{\theta^*} g(X_1; \theta) = E_{\theta^*} g(X_1; \theta^*) - \tfrac{1}{2}(\theta - \theta^*)^T \sigma(\theta^*)(\theta - \theta^*) + O(|\theta - \theta^*|^3),$$

so that, if θ^* is near Θ_0, λ_n has bias approximately $\frac{1}{2} n(\theta' - \theta^*)^T \sigma(\theta^*)(\theta' - \theta^*)$, where θ' minimizes this last expression among $\theta \in \Theta_0$. Thus the bias is a rough measure of the squared distance of θ^* from Θ_0. The choice of rejection region for a particular problem depends very much on what balance between expected sample sizes and power is appropriate. The strategies $S1$ and $S2$ are chosen because they are simple to use, $S2$ involving a slightly more complicated statistic than $S1$ but being easier to tabulate, and they form, together with the fixed sample test, a reasonable selection from which to choose.

As a postscript to Theorem 1 it is interesting to consider what happens when, in addition, $E(e^{s|X|}) < \infty$ for some $s > 0$. Under these circumstances the Corollary to Theorem 1 of Komlós, Major and Tusnády(5) shows that it is possible to construct, on the same space as $(X_j)_{j \geq 1}$, an equilibrium Ornstein-Uhlenbeck process Y as in Theorem 1, in such a way that, for some constant C,

$$\limsup_{n \to \infty} (n^{\frac{1}{2}}/\log n) \, |S_n/\sigma\sqrt{(n)} - Y(\log n)| \leq C \text{ a.s.}$$

REFERENCES

(1) ARMITAGE, P. Restricted sequential procedures. *Biometrika* **44**, (1957), 9–26.

(2) BECHHOFER, R. E. A note on the limiting relative efficiency of the Wald sequential probability ratio test. *J. Amer. Statist. Assoc.* **55** (1960) 660–663.

(3) BILLINGSLEY, P. *Statistical inference for Markov processes* (University of Chicago Press, 1961).

(4) CRAMÉR, H. *Mathematical methods of statistics* (Princeton University Press, 1946).

(5) KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. An approximation of partial sums of independent RV'-s, and the sample DF. I. *Z. Wahrscheinlichkeitstheorie* **32** (1975), 111–131; II. *Z. Wahrscheinlichkeitstheorie* **34**, (1976), 33–58.

(6) MÜLLER, D. W. Verteilungs-Invarianzprinzipien für das starke Gesetz der grossen Zahl. *Z. Wahrscheinlichkeitstheorie* **10**, (1968), 173–192.

(7) SCHWARZ, G. Asymptotic shapes for Bayes sequential testing regions. *Ann. Math. Stat.* **33**, (1962) 224–236.

(8) WALD, A. Note on the consistency of the maximum likelihood estimator. *Ann. Math. Stat.* **20**, (1949), 595–601.

(9) WOODROOFE, M. Large deviations of likelihood ratio statistics, with applications to sequential testing. *Ann. Statist.* **6**, (1978), 72–84.