
MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016

SEQUENTIAL ANALYSIS
Vol. 21, No. 4, pp. 191–218, 2002

CORRECTED CONFIDENCE SETS FOR SEQUENTIALLY DESIGNED EXPERIMENTS: EXAMPLES*

Michael Woodroofe[1] and D. Stephen Coad[2]

[1]University of Michigan, Ann Arbor, Michigan
[2]University of Sussex, Falmer, Brighton, England

1. INTRODUCTION

Ford and Silvey (1980) proposed an adaptive design for estimating the point at which a regression function attains its minimum. In their model, there are potential observations of the form

$$y_k = \theta_1 x_k + \theta_2 x_k^2 + \epsilon_k, \quad k = 1, 2, \ldots, \tag{1}$$

where x_k are design points chosen from the interval $-1 \le x \le 1$, θ_1 and $\theta_2 > 0$ are unknown parameters, and $\epsilon_1, \epsilon_2, \ldots$ are i.i.d. standard normal random variables. Interest centered on estimating the value $-\theta_1/2\theta_2$ at which the regression function attains its minimum. After examining the asymptotic variance of the maximum likelihood estimator, Ford and Silvey proposed the following *adaptive design*: first take observations at $x_1 = -1$ and $x_2 = +1$; thereafter, if (x_k, y_k), $k = 1, \ldots, n$, have been determined, let y_n^- or y_n^+ denote the sum of y_k for which $k \le n$ and $x_k = -1$ or $x_k = +1$, respectively; take the next observation at $x_{n+1} = +1$ if $|y_n^+| < |y_n^-|$

*Reprinted from *Multivariate Analysis, Design of Experiments, and Survey Sampling*; Ghosh, S., Ed.; Marcel Dekker, Inc.: New York, 1999, 135–161.

DOI: 10.1081/SQA-120016300
0747-4946 (Print); 1532-4176 (Online)

MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016

and take the next observation at $x_{n+1} = -1$ otherwise. After the experiment is run, investigators may want confidence intervals, and a problem arises here. Owing to the adaptive nature of the design, it is not the case that the maximum likelihood estimators, say $\hat{\theta}_{n,1}$ and $\hat{\theta}_{n,2}$, have a bivariate normal distribution. This problem was addressed by Ford et al. (1985) and Wu (1985). The former paper proposed an exact solution which seems overly conservative. The latter proposed an asymptotic solution which, at a practical level, ignores the adaptive nature of the design. This paper contains an asymptotic solution which does not ignore the adaptive nature of the design.

The Ford–Silvey example fits nicely into the following more general model: letting $'$ denote transpose, suppose that there are potential observations of the form

$$y_k = x_k'\theta + \sigma\epsilon_k, \quad k = 1, 2, \ldots \tag{2}$$

where $x_k = (x_{k,1}, \ldots, x_{k,p})'$ are design variables which may be chosen from a subset $\mathcal{X} \subseteq \Re^p$, $\theta = (\theta_1, \ldots, \theta_p)'$ is a vector of unknown parameters, $\sigma > 0$ may be known or unknown, and $\epsilon_1, \epsilon_2, \ldots$ are i.i.d. standard normal random variables. The design vectors x_k, $k = 1, 2, \ldots$, may be chosen *adaptively*; that is, each x_k may be a (measurable) function of previous responses and auxiliary randomization, say

$$x_k = x_k(u_1, \ldots, u_k, y_1, \ldots, y_{k-1}), \quad k = 1, 2, \ldots, \tag{3}$$

where $u_1, u_2, \ldots$ are independent of $\epsilon_1, \epsilon_2, \ldots$ and have a known distribution. Letting $\mathbf{y}_n = (y_1, \ldots, y_n)'$, $X_n = (x_1, \ldots, x_n)'$, and $\mathbf{e}_n = (\epsilon_1, \ldots, \epsilon_n)'$, the model in Eq. (2) may be written in the familiar form

$$\mathbf{y}_n = X_n\theta + \sigma\mathbf{e}_n, \quad n = 1, 2, \ldots.$$

The model is very general. It includes adaptive procedures for finding the maximum or minimum of a regression function, as in Ford and Silvey (1980), models for sequential clinical trials, as in Coad (1995), adaptive biased coin designs, as in Eisele (1994), and time series and controlled time series, as in Lai and Wei (1982), among others.

The likelihood function is not affected by the adaptive nature of the design, or by the optional stopping introduced later (see, for example, Berger and Wolpert, 1984). So the maximum likelihood estimator of θ has the familiar form

$$\hat{\theta}_n = (X_n'X_n)^{-1}X_n'\mathbf{y}_n \tag{4}$$

provided that $X_n'X_n$ is positive definite. The maximum likelihood estimator of σ^2 also has a familiar form. The usual estimator of σ^2 for a nonadaptive design is

$$\hat{\sigma}_n^2 = \frac{\|\mathbf{y}_n - X_n\hat{\theta}_n\|^2}{n-p} \tag{5}$$

when $X_n'X_n > 0$ and $n > p$. This estimator is unbiased for nonadaptive designs and is nearly unbiased in the absence of optional stopping. See Sec. 2 for the bias when there is a stopping time.

The sampling distributions of these estimators may be affected by the adaptive design, however, and it is the purpose of this paper to explain how approximate expressions for the sampling distributions of $\hat{\theta}_n$ may be obtained. Approximations to the sampling distributions are presented in Secs. 2–4. These are illustrated by simple examples in Secs. 2–4, and by more complicated examples in Secs. 5 and 6. The accuracy of the approximations is assessed by simulation. The presentation in Secs. 2–4 is informal. Precise statements, with conditions, are deferred to Sec. 7. Proofs are discussed briefly in Sec. 8.

1.1. Remark on Notation

Below, $P_{\sigma,\theta}$ denotes a probability model under which Eq. (2) holds, and $E_{\sigma,\theta}$ denotes expectation with respect to $P_{\sigma,\theta}$. When σ is known and fixed, it may be omitted from the notation. For example, probability and expectation may be denoted by P_θ and E_θ.

2. BIASES

The effect of the adaptive design may be illustrated by computing the biases of $\hat{\theta}_n$ and $\hat{\sigma}_n^2$. With a view towards Example 3 below, suppose that the model in Eq. (2) is observed for $k = 1, \ldots, N$, where N is a stopping time with respect to (u_k, y_k), $k = 1, 2, \ldots$. That is, the event $\{N = n\}$ can depend only on (u_k, y_k), $k = 1, \ldots, n$. The approximate biases are derived from asymptotic expansions. Thus, suppose that N depends on a parameter a, say $N = N_a$, and that $N_a \to \infty$ in probability as $a \to \infty$. The case of nonrandom sample size $N = n \to \infty$ is not excluded here. Suppose that $X_N'X_N > 0$ *w.p.*1 for all a, and that

$$\eta(\sigma, \theta) := \lim_{a\to\infty} a(X_N'X_N)^{-1} \tag{6}$$

and

$$\rho(\sigma^2,\theta) := \lim_{a\to\infty} \frac{a}{N} \tag{7}$$

exist in $P_{\sigma,\theta}$= probability for *a.e.* θ for each $\sigma > 0$. Actually, more is required (see Sec. 8). Write $\eta(\sigma,\theta) = [\eta_{ij}(\sigma,\theta), i,j = 1,\ldots,p]$ and let $\mathbf{1} = (1,\ldots,1)'$. Then

$$E_{\sigma,\theta}(\hat{\theta}_N - \theta) \approx \frac{\sigma^2}{a}\eta^{\#}(\sigma,\theta)\mathbf{1} \tag{8}$$

and

$$E_{\sigma,\theta}(\hat{\sigma}_N^2 - \sigma^2) \approx \frac{2\sigma^4}{a}\rho'(\sigma^2,\theta) \tag{9}$$

in the very weak sense of Woodroofe (1986, 1989), where $'$ denotes differentiation with respect to σ^2 and

$$\eta_{ij}^{\#}(\sigma,\theta) = \frac{\partial}{\partial\theta_j}\eta_{ij}(\sigma,\theta) \tag{10}$$

These are among the main findings of Coad and Woodroofe (1998), who also obtain approximations to variances and covariances. Observe that the bias of $\hat{\sigma}_N^2$ is determined primarily by the optional stopping; that is, if $N = n = a$, then $\rho(\sigma^2,\theta) = 1$ for all σ^2 and θ, so that $\rho'(\sigma^2,\theta) = 0$ and, therefore, $E_{\sigma,\theta}(\hat{\sigma}_n^2 - \sigma^2) = o(1/n)$.

Example 1. *An Autoregressive Process*. To illustrate the use of Eqs. (8) and (9), consider the simple autoregressive process $y_k = \theta y_{k-1} + \sigma\epsilon_k$, $k = 1,2,\ldots$, with $y_0 = 0$. Suppose that the process is observed for n time units, so that $N = n$. Then $p = 1$, $X_n'X_n = y_0^2 + y_1^2 + \cdots + y_{n-1}^2$, and

$$\lim_{n\to\infty} n(X_n'X_n)^{-1} = \lim_{n\to\infty} \frac{n}{y_0^2 + y_1^2 + \cdots + y_{n-1}^2} = \frac{1-\theta^2}{\sigma^2}$$

*w.p.*1 for all $|\theta| < 1$ and $\sigma > 0$. It then follows from Eq. (8) that

$$E_{\sigma,\theta}(\hat{\theta}_n - \theta) \approx -\frac{2\theta}{n} \tag{11}$$

for large n, and $E_{\sigma,\theta}(\hat{\sigma}_n^2 - \sigma^2) = o(1/n)$ as $n \to \infty$, and Eq. (11) agrees well with Coad and Woodroofe's (1998) simulations for n as small as 25. □

Example 2. *The Ford–Silvey Example*. In the Ford–Silvey example in Eq. (1), let $N = n = a$ since there is no stopping time, and write

$$X_n'X_n = \begin{pmatrix} \sum x_k^2 & \sum x_k^3 \\ \sum x_k^3 & \sum x_k^4 \end{pmatrix} = \begin{pmatrix} n & s_n \\ s_n & n \end{pmatrix}$$

where $s_n = x_1 + \cdots + x_n$. Ford and Silvey (1980) showed that *w.p.*1, $s_n/n \to -\kappa(\theta)$, where $\kappa(\theta) = \theta_1/\theta_2$, if $|\theta_1| \le |\theta_2|$, and $\kappa(\theta) = \theta_2/\theta_1$ otherwise. It follows that if $|\theta_1| \ne |\theta_2|$, then

$$\eta(\theta) = \lim_{n\to\infty} n(X_n'X_n)^{-1} = \frac{1}{1-\kappa^2}\begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix}$$

where κ has been written for $\kappa(\theta)$. The expression for $\eta^{\#}(\theta)$ is complicated and will not be detailed. After some algebra, however, it is easily seen that

$$E_\theta(\hat{\theta}_n - \theta) \approx \frac{1}{n}\eta^{\#}(\theta)\mathbf{1} = \frac{1}{n\beta(\theta)}\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \tag{12}$$

where $\beta(\theta) = \max[\theta_1^2, \theta_2^2] - \min[\theta_1^2, \theta_2^2]$. Again, the approximation in Eq. (12) agrees well with Coad and Woodroofe's (1998) simulations for n as small as 25. □

3. SAMPLING DISTRIBUTIONS

In this section, σ is assumed to be known, and probability is denoted by P_θ. Let m be an integer, $1 \le m \le p$, and let A_n be an $m \times p$ matrix and B_n be a $p \times p$ matrix for which

$$A_nA_n' = I_m \quad \text{and} \quad X_n'X_n = B_nB_n', \tag{13}$$

where I_m denotes the $m \times m$ identity matrix. There are many possible choices for A_n and B_n, and some advantages of using a Cholesky decomposition for $X_n'X_n$ are described by Woodroofe and Coad (1997). The only requirements, however, are Eq. (13), Eq. (16) below, and that A_n and B_n depend measurably on X_n and $\mathbf{y}_n$. If $X_n'X_n > 0$, let

$$Z_n^o = \frac{1}{\sigma}B_n'(\theta - \hat{\theta}_n) \tag{14}$$

and

$$W_n^o = A_nZ_n^o = \frac{1}{\sigma}A_nB_n'(\theta - \hat{\theta}_n) \tag{15}$$

Here W_n^o and Z_n^o are first approximations to approximately pivotal quantities. Of course, Z_n^o would have an (exactly) standard p-variate normal distribution in the absence of an adaptive design and optional stopping.

The problem is to find an approximation to the distribution of W_N^o. Let

$$Q^{(a)} = \sqrt{a} A_N B_N^{-1}, \quad a \geq 1$$

where $N = N_a$ denotes the stopping time. The conditions for the expansions require that $Q^{(a)}$ have a limit as $a \to \infty$, say

$$\lim_{a \to \infty} Q^{(a)} = Q(\theta) \tag{16}$$

in probability, where $Q(\theta) = [q_{ij}(\theta) : i = 1, \ldots, m, \ j = 1, \ldots, p]$. Again, a stronger form of convergence is required (see Sec. 7). Suppose that the entries $q_{ij}(\theta)$ are differentiable with respect to θ, and let

$$q_{ij}^{\#}(\theta) = \frac{\partial}{\partial \theta_j} q_{ij}(\theta)$$

$$m_{ij}(\theta) = \sum_{k=1}^{p} \sum_{\ell=1}^{p} \frac{\partial^2}{\partial \theta_k \partial \theta_\ell} [q_{ik}(\theta) q_{j\ell}(\theta)]$$

$Q^{\#}(\theta) = [q_{ij}^{\#}(\theta) : i = 1, \ldots, m, \ j = 1, \ldots, p]$ and $M(\theta) = [m_{ij}(\theta) : i, j = 1, \ldots, m]$. Next, let Φ^m denote the standard m-variate normal distribution and write

$$\Phi^m h = \int_{\Re^m} h(w) \Phi^m\{\mathrm{d}w\}$$

$$\Phi_1^m h = \int_{\Re^m} w h(w) \Phi^m\{\mathrm{d}w\}, \quad (m \times 1) \tag{17}$$

and

$$\Phi_2^m h = \frac{1}{2} \int_{\Re^m} (ww' - I_m) h(w) \Phi^m\{dw\}, \quad (m \times m)$$

whenever the integrals are meaningful. If $h : \Re^m \to \Re$ is a function of quadratic growth (that is, $|h(w)| \leq C(1 + \|w\|^2)$ for all $w \in \Re^m$ for some constant $0 < C < \infty$), then

$$E_\theta[h(W_N^o)] \approx \Phi^m h - \frac{\sigma}{\sqrt{a}} (\Phi_1^m h)' Q^{\#}(\theta) 1 + \frac{\sigma^2}{a} \mathrm{tr}[(\Phi_2^m h) M(\theta)] \tag{18}$$

for large a in the very weak sense. This is Theorem 1 of Woodroofe and Coad (1997). See Sec. 7 for a precise statement with conditions. Specializing Eq. (18) to the cases $h(w) = w_i$ and $h(w) = w_i w_j$ gives an approximate mean and covariance matrix of W_N^o; that is,

$$E_\theta(W_N^o) \approx -\frac{\sigma}{\sqrt{a}} Q^{\#}(\theta)1 = \mu_a(\theta), \quad \text{say} \tag{19}$$

and

$$E_\theta(W_N^o W_N^{o'}) \approx I_m + \frac{\sigma^2}{a} M(\theta) \tag{20}$$

Relation (18) may be summarized as W_N^o is approximately normal with mean $\mu_a(\theta)$ and covariance matrix $I_m + a^{-1}\sigma^2 M(\theta) - \mu_a(\theta)\mu_a(\theta)'$ because, letting Ψ_a denote the latter distribution, a three-term Taylor series expansion of $\int_{\mathcal{R}^m} h \, d\Psi_a$ agrees with the right-hand side of Eq. (18). This is a remarkably simple description, given the generality of the basic model in Eq. (2).

The magnitude of the effect and the accuracy of the approximation are illustrated in the following example.

Example 3. Hayre and Gittins (1981) considered a problem in which two treatments, A and B say, produce normally distributed responses with unknown means μ and ν and variance σ^2, say. Such a model may be written in the form of Eq. (2) with $\theta_1 = \mu$, $\theta_2 = \nu - \mu$, and $x = (1, z)'$, where $z = 0$ or 1 accordingly as an observation is made on A or B. Motivated by clinical trials in which there was an ethical cost for giving an inferior treatment, Hayre and Gittins (1981) suggested the sampling rule $z_1 = 0$, $z_2 = 1$, and

$$z_{n+1} = 1 \quad \text{iff} \quad \frac{s_n}{n - s_n} \le w(\hat{\theta}_{n,2}), \quad n \ge 2$$

where $s_n = z_1 + \cdots + z_n$ and w is a positive function on $\Re$. For example, the function $w(\delta) = \sqrt{1 + 10\delta}$ for $\delta > 0$ and $w(\delta) = 1/\sqrt{1 + 10|\delta|}$ for $\delta \le 0$ is used in the simulations below. For the case of known σ^2, the sequential design was to be used with the stopping time

$$N = \inf\{n \ge 3 : |i_n^2 \hat{\theta}_{n,2}| > a\sigma^2\}$$

where

$$i_n^2 = \frac{s_n(n - s_n)}{n}$$

and $a > 0$ is a design parameter, chosen to control the error probabilities of a sequential probability ratio test. In this example,

$$X_n' X_n = \begin{pmatrix} n & s_n \\ s_n & s_n \end{pmatrix}, \quad n \geq 1$$

Further, there is special interest in $\theta_2 = \nu - \mu$, the difference of the two means, and it is natural to let $A_n B_n' = c_n(0, 1)$, where c_n is a constant. This may be accomplished by letting

$$A_n = \left[-\sqrt{\frac{s_n}{n}}, \sqrt{\frac{n - s_n}{n}} \right]$$

and

$$B_n = \begin{pmatrix} \sqrt{n - s_n} & \sqrt{s_n} \\ 0 & \sqrt{s_n} \end{pmatrix}$$

in which case $c_n = i_n$ and

$$Q^{(a)} = \sqrt{a} A_N B_N^{-1} = \frac{\sqrt{a}}{i_N} \left[-\frac{s_N}{N}, 1 \right]$$

To apply Eq. (18), it is necessary to find the limiting matrix $Q(\theta)$. Hayre and Gittins (1981) showed that

$$\frac{a}{i_N^2} \rightarrow \frac{|\theta_2|}{\sigma^2}, \qquad \frac{s_N}{N} \rightarrow \frac{w(\theta_2)}{1 + w(\theta_2)}$$

and

$$\frac{a}{N} \rightarrow \frac{|\theta_2| w(\theta_2)}{\sigma^2 [1 + w(\theta_2)]^2}$$

*w.p.*1. It follows that

$$Q^{(a)} \rightarrow \frac{\sqrt{|\theta_2|}}{\sigma} \left[-\frac{w(\theta_2)}{1 + w(\theta_2)}, 1 \right] = Q(\theta)$$

$$Q^{\#}(\theta) = \frac{1}{\sigma} \left[0, \frac{\text{sign}(\theta_2)}{2\sqrt{|\theta_2|}} \right]$$

and $m_{11}(\theta) = 0$ for $\theta_2 \neq 0$. Thus, $\mu_a(\theta) = -\text{sign}(\theta_2)/(2\sqrt{a|\theta_2|})$ in Eq. (19). When specialized to indicator functions, Eq. (18) asserts

$$P_\theta\{W_N^o \leq c\} \approx \Phi(c) + \frac{1}{\sqrt{a}}\varphi(c)\frac{\text{sign}(\theta_2)}{2\sqrt{|\theta_2|}} \tag{21}$$

where Φ and ϕ denote the standard normal univariate distribution and density functions.

It is clear from Eq. (21) that the correction term $\varphi(c)\text{sign}(\theta_2)/(2\sqrt{a|\theta_2|})$ can be significant, and the approximation in Eq. (21) effectively predicts the simulated values in Table 1 for the cases $\theta_1 = 0$, $\theta_2 = 0.5$, $\theta_2 = 1$, and $a = 6$. The accuracy of the approximation deteriorates as $|\theta_2|$ decreases. For $a = 6$, the approximation overcorrects when $\theta_2 = 0.25$. Woodroofe (1989) has reported simulations of a similar nature for a closely related example, due to Robbins and Siegmund (1974). In this example, the approximation does not depend on the function w that determines the allocation rule. Coad (1991) has shown that this independence holds more generally, for

Table 1. $P_\theta\{W_N^o \leq c\}$ in the Hayre–Gittins Example

		$\theta_2 = 0.25$		$\theta_2 = 0.5$		$\theta_2 = 1$	
c	Normal	MC	Approx.	MC	Approx.	MC	Approx.
−2.00	.0228	.0381	.0448	.0369	.0383	.0324	.0338
−1.75	.0401	.0623	.0753	.0622	.0650	.0536	.0577
−1.50	.0668	.0994	.1197	.0998	.1042	.0911	.0932
−1.25	.1056	.1525	.1802	.1572	.1584	.1443	.1429
−1.00	.1587	.2264	.2574	.2272	.2285	.2148	.2081
−0.75	.2266	.3199	.3496	.3084	.3186	.2748	.2881
−0.50	.3085	.4245	.4523	.4082	.4102	.3631	.3804
−0.25	.4013	.5325	.5593	.5092	.5129	.4754	.4802
0.00	.5000	.6338	.6629	.6126	.6152	.5780	.5814
0.25	.5987	.7242	.7566	.7054	.7103	.6740	.6776
0.50	.6915	.8003	.8352	.7863	.7931	.7593	.7633
0.75	.7734	.8619	.8963	.8525	.8603	.8307	.8348
1.00	.8413	.9048	.9401	.9021	.9112	.8826	.8907
1.25	.8944	.9314	.9689	.9409	.9471	.9288	.9316
1.50	.9332	.9451	.9861	.9649	.9706	.9548	.9596
1.75	.9599	.9513	.9952	.9787	.9848	.9758	.9776
2.00	.9772	.9534	.9993	.9869	.9928	.9870	.9883

Note: Based on 10,000 replications with $a = 6$, $\sigma = 1$, $\theta_1 = 0$, $w(\delta) = \sqrt{1 + 10\delta}$ for $\delta > 0$, and $w(\delta) = 1/\sqrt{1 + 10|\delta|}$ for $\delta \leq 0$.

essentially different allocation rules, and has reported simulations which indicate that the approximation in Eq. (21) works well for other allocation rules too. □

3.1. Corrected Confidence Sets

The main application of Eq. (18) is to form corrected confidence sets, confidence sets whose actual coverage probability differs from the nominal by $o(1/a)$. To do this, it is convenient to standardize W_N^o. For the case of known σ^2, let $\hat{\mu}_a^o$ denote an estimator of $\mu_a(\theta)$, for example, $\hat{\mu}_a^o = \mu_a(\hat{\theta}_N)$. Then it may be shown that

$$E_\theta[(W_N^o - \hat{\mu}_a^o)(W_N^o - \hat{\mu}_a^o)'] \approx I_m + \frac{1}{a}\Delta^o(\theta)$$

where

$$\Delta_{ij}^o(\theta) = \sigma^2 \sum_{k=1}^{p}\sum_{\ell=1}^{p} \frac{\partial}{\partial\theta_\ell} q_{ik}(\theta) \frac{\partial}{\partial\theta_k} q_{j\ell}(\theta)$$

for $i,j = 1,\ldots,m$. Let $\hat{\Delta}_a^o$ denote an estimator of $\Delta^o(\theta)$, for example, $\hat{\Delta}_a^o = \Delta^o(\hat{\theta}_N)$, and let

$$W_N^{o*} = \left(I_m + \frac{\hat{\Delta}_a^o}{2a}\right)^{-1} (W_n^o - \hat{\mu}_a^o)$$

Then $W_N^o{*}$ is approximately standard m-variate normal to a high order; that is

$$E_\theta[h(W_N^{o*})] \approx \Phi^m h \tag{22}$$

for functions h of quadratic growth. Here $\approx$ means equality up to $o(1/a)$ in the very weak sense of Woodroofe (1986, 1989). This is Theorem 2 of Woodroofe and Coad (1997). See Sec. 7 for precise statements with conditions.

It is easy to use Eq. (22) to form corrected confidence sets for θ. Given a subset $C \subseteq \Re^m$, let

$$\mathcal{C} = \{\theta : W_N^{o*} \in C\} \tag{23}$$

Then

$$P_\theta\{\theta \in \mathcal{C}\} = P_\theta\{W_N^{o*} \in C\} \approx \Phi^m(C)$$

That is, $\mathcal{C}$ is an approximate confidence set of level $\Phi^m(C)$ to order $o(1/a)$. In applications, C might be of the form $C = \{\theta : |\theta| \le c\}$, where $|\cdot|$ is a norm on $\Re^m$ and $c > 0$. There may be interest in a given set of linear functions, say $\Gamma\theta$ where Γ is an $m \times p$ matrix. For example, the rows of Γ might be a set of contrasts. In such cases, it is easy to construct A_n, so that $\mathcal{C}$ is, in fact, a confidence set for $\Gamma\theta$. To do so, let G_n be an $m \times m$ matrix for which $G_n\Gamma(X_n'X_n)^{-1}\Gamma'G_n' = I_m$, and let $A_n = G_n\Gamma B_n'^{-1}$. Then $A_nA_n' = I_m$ and $A_nB_n' = G_n\Gamma$, so that

$$\mathcal{C} = \{\theta : \Gamma\theta \in G_n^{-1}C\}$$

There is further simplification if $m = 1$. Then $G_n = 1/\sqrt{\Gamma(X_n'X_n)^{-1}\Gamma'}$, $A_nB_n^{-1} = G_n\Gamma(X_n'X_n)^{-1}$, and

$$Q^{(a)} \to \frac{\Gamma\eta(\theta)}{\sqrt{\Gamma\eta(\theta)\Gamma'}} = Q(\theta) \tag{24}$$

if Eq. (6) holds.

4. UNKNOWN σ

The procedure is similar for the case of unknown σ^2, but the answers are more complicated. For unknown σ^2, let

$$Z_n = \frac{1}{\hat{\sigma}_n} B_n'(\theta - \hat{\theta}_n) \tag{25}$$

and

$$W_n = A_nZ_n = \frac{1}{\hat{\sigma}_n} A_nB_n'(\theta - \hat{\theta}_n) \tag{26}$$

where A_n and B_n are as in Eq. (13). Then the analogue of Eq. (18) takes the following form. If h is a function of quadratic growth, then

$$\begin{aligned} E_{\sigma,\theta}[h(W_N)] \approx \Phi^m h &- \frac{\sigma}{\sqrt{a}}(\Phi_1^m h)'Q^{\#}(\sigma,\theta)\mathbf{1} \\ &+ \frac{\sigma^2}{a}\mathrm{tr}\{(\Phi_2^m h)[M(\sigma,\theta) - \sigma^2\rho'(\sigma^2,\theta)I_m]\} \\ &+ \frac{1}{a}(\Phi_4^m h)\rho(\sigma^2,\theta) \end{aligned} \tag{27}$$

where

$$\Phi_4^m h = \frac{1}{4}\int_{\Re^m} h(w)[\|w\|^4 - 2m\|w\|^2 + m(m-2)]\Phi^m\{dw\}$$

and ρ is as in Eq. (7). The final two terms on the right-hand side of Eq. (27) arise from the variability in $\hat{\sigma}_N^2$. In the absence of optional stopping, $\rho(\sigma^2, \theta) = 1$ and $\rho'(\sigma^2, \theta) = 0$, and the right-hand side of Eq. (27) simplifies.

To form corrected confidence sets, write $\mu_a(\theta) = \mu_a(\sigma, \theta)$ and $\Delta^o(\theta) = \Delta^o(\sigma, \theta)$ to emphasize the dependence on σ. Further, let

$$\Delta(\sigma, \theta) = \Delta^o(\sigma, \theta) - \sigma^2\rho'(\sigma^2, \theta)I_m$$

let $\hat{\boldsymbol{\mu}}_a$ and $\hat{\Delta}_a$ denote estimators of $\mu_a(\sigma, \theta)$ and $\Delta(\sigma, \theta)$ for example $\hat{\boldsymbol{\mu}}_a = \mu_a(\hat{\sigma}_N, \hat{\theta}_N)$ and $\hat{\Delta}_a = \Delta(\hat{\sigma}_N, \hat{\theta}_N)$, and let

$$W_N^* = \left(I_m + \frac{\hat{\Delta}_a}{2a}\right)^{-1}(W_N - \hat{\boldsymbol{\mu}}_a) \tag{28}$$

If h is a function of quadratic growth, then

$$E_{\sigma,\theta}[h(W_N^*)] \approx \Phi^m h + (\Phi_4^m h)\frac{\rho(\sigma^2, \theta)}{a} \tag{29}$$

The dependence of the right-hand side of Eq. (29) on the parameters is more apparent than real, since $\rho(\sigma^2, \theta)/a$ may be estimated by $1/N$ (see Eq. (7)). Let $\mathcal{T}_n^m$ denote the standard m-variate t-distribution on n degrees of freedom, and write $\mathcal{T}_n^m h = \int_{\Re^m} h(y)\mathcal{T}_n^m\{\mathrm{d}y\}$ for suitable functions h. Then it is not difficult to see that the right-hand side of Eq. (29) is an approximation to $\mathcal{T}_{a/\rho}^m h$ which may be estimated by $\mathcal{T}_N^m h$ or $\mathcal{T}_{N-p}^m h$.

The construction of corrected confidence sets now proceeds as in the previous section. If $C_a \subseteq \Re^m$ are measurable and

$$\mathcal{C}_a = \{\theta : W_N^* \in C_a\}$$

then

$$P_{\sigma,\theta}\{\theta \in \mathcal{C}_a\} \approx \mathcal{T}_{a/\rho}^m(C_a)$$

Example 1 revisited. In the autoregressive example with $N = n = a$,

$$Q(\sigma, \theta) = \frac{\sqrt{1-\theta^2}}{\sigma}$$

$$\mu_n(\sigma, \theta) = \frac{\theta}{\sqrt{n(1-\theta^2)}}$$

and

$$\Delta(\sigma,\theta)=\frac{\theta^2}{1-\theta^2}$$

Both μ_n and Δ are independent of σ. Let $\tilde{\theta}_n=\hat{\theta}_n$ if $|\hat{\theta}_n|\le 1-1/n$ and let $\tilde{\theta}_n=\pm(1-1/n)$ otherwise. Further, let $\hat{\mu}_n=\mu_n(1,\tilde{\theta}_n)$, let $\hat{\Delta}_n=\Delta(1,\tilde{\theta}_n)$, and define W_n^* by Eq. (28). Let $i_n^2=X_n'X_n=y_0^2+\cdots+y_{n-1}^2$. If c_n is the upper $\alpha/2$ quantile of the t-distribution on $n-1$ degrees of freedom, then

$$P_{\sigma,\theta}\left\{-\frac{\hat{\sigma}_n}{i_n}\left[c_n\left(1+\frac{\hat{\Delta}_n}{2n}\right)-\hat{\mu}_n\right]\le\theta-\hat{\theta}_n\le\frac{\hat{\sigma}_n}{i_n}\left[c_n\left(1+\frac{\hat{\Delta}_n}{2n}\right)+\hat{\mu}_n\right]\right\}\approx 1-\alpha \tag{30}$$

for $|\theta|<1$ and $\sigma>0$.

The assess the accuracy of Eq. (30), a simulation experiment was conducted in which 10,000 samples of size n were generated for selected values of α, n, and θ with $\sigma=1$. For each choice of α, n, and θ, the coverage probability on the left-hand side of Eq. (30) was estimated by Monte Carlo. The results are reported in Table 2. They show that the approximation in Eq. (30) is excellent, even for θ near to ± 1. □

Example 3 revisited. If σ^2 is unknown in Example 3 (the Hayre–Gittins example), then the stopping time N is modified as follows. Five observations are taken from each population at the beginning, and

$$N=\inf\left\{n\ge 11:|i_n^2\hat{\theta}_{n,2}|>a\left(\frac{n+6}{n-6}\right)\hat{\sigma}_n^2\right\}$$

Table 2. Coverage Probabilities for Autoregressive Processes

	$n=25$		$n=50$	
θ	$\alpha=0.05$	$\alpha=0.10$	$\alpha=0.05$	$\alpha=0.10$
0.0	.949	.899	.948	.899
0.2	.947	.896	.949	.896
0.4	.946	.894	.946	.896
0.6	.945	.895	.945	.896
0.8	.953	.894	.947	.897
0.9	.954	.904	.953	.899
±	.0044	.006	.0044	.006

Note: Based on 10,000 replications with $\sigma=1$; ± is two standard deviations.

Then, $Q^{\#}(\sigma,\theta)=[0,\text{sign}(\theta_2)/2\sqrt{|\theta_2|}]/\sigma$, $\mu_a(\sigma,\theta)=-\text{sign}(\theta_2)/(2\sqrt{a|\theta_2|})$, and $\Delta^o(\sigma,\theta)=1/4|\theta_2|$, as above. Further,

$$\rho(\sigma^2,\theta)=\frac{|\theta_2|w(\theta_2)}{[1+w(\theta_2)]^2\sigma^2}$$

and

$$\rho'(\sigma^2,\theta)=-\frac{|\theta_2|w(\theta_2)}{[1+w(\theta_2)]^2\sigma^4}$$

In view of the singularity at $\theta_2=0$, it seems prudent to smooth the estimators near $\hat{\theta}_{N,2}=0$. There are many ways to do this. Here we take

$$\hat{\mu}_a=-\frac{\text{sign}(\hat{\theta}_{N,2})}{2\sqrt{a((1/N)+|\hat{\theta}_{N,2}|)}}$$

and

$$\hat{\Delta}_a=\frac{1}{4[(1/N)+|\hat{\theta}_{N,2}|]}-\sigma_N^2\rho'(\hat{\sigma}_N^2,\hat{\theta}_N)$$

Then Eq. (29) holds. In this example, $W_N=i_N(\theta_2-\hat{\theta}_{N,2})$. So, if c_n is the upper $100\alpha/2$ percentile of the t-distribution on $n-2$ degrees of freedom, then an approximate $100(1-\alpha)$ percent confidence interval for θ_2 is

$$\mathcal{C}=\left\{\theta:-\left[c_N\left(1+\frac{\hat{\Delta}_a}{2a}\right)-\hat{\mu}_a\right]\frac{\hat{\sigma}_N}{i_N}\le\theta_2-\hat{\theta}_{N,2}\right.$$
$$\left.\le\left[c_N\left(1+\frac{\hat{\Delta}_a}{2a}\right)+\hat{\mu}_a\right]\frac{\hat{\sigma}_N}{i_N}\right\}$$

As above, the accuracy of this interval may be assessed by simulation. Simulated coverage probabilities are reported in Table 3 for selected values of α, θ_2, σ, and a, and $\theta_1=0$. These show excellent agreement with the nominal values for $a=9$ and very good agreement for $a=6$. It is interesting that the corrected confidence intervals are valid, even for small values of θ_2. □

Remark. There are other applications to clinical trials. Woodroofe and Coad (1997) have shown how Eq. (22) can be used to form simultaneous confidence intervals in Siegmund's (1993) and Betensky's (1996) procedures

Table 3. Coverage Probabilities for the Hayre–Gittins Example

				$a=6$					
σ		0.75			1.0			1.5	
α		0.05	0.10		0.05	0.10		0.05	0.10
θ_2	$E(N)$	CP	CP	$E(N)$	CP	CP	$E(N)$	CP	CP
0.100	102.5	.955	.908	165.7	.949	.905	338.2	.947	.899
0.250	69.8	.948	.898	113.0	.943	.894	234.3	.944	.892
0.375	52.9	.957	.907	85.5	.949	.901	176.0	.948	.896
0.500	43.2	.954	.902	68.9	.949	.898	141.7	.946	.891
0.750	32.8	.954	.906	51.9	.955	.907	104.6	.949	.900
1.000	27.5	.958	.905	42.9	.954	.904	85.0	.945	.898
$\pm$		.0044	.006		.0044	.006		.0044	.006
				$a=9$					
0.100	178.6	.951	.905	299.3	.952	.903	635.0	.948	.897
0.250	100.5	.954	.904	169.8	.954	.906	362.7	.951	.903
0.375	74.2	.951	.901	123.4	.948	.899	261.9	.953	.904
0.500	59.6	.951	.899	98.1	.950	.898	206.9	.952	.903
0.750	45.0	.953	.902	73.3	.948	.898	152.0	.946	.895
1.000	37.6	.952	.897	59.8	.949	.903	122.7	.949	.901
$\pm$		.0044	.006		.0044	.006		.0044	.006

Note: Based on 10,000 replications with $\theta_1=0$, $w(\delta)=\sqrt{1+10\delta}$ for $\delta>0$, $w(\delta)=1/\sqrt{1+10|\delta|}$ for $\delta\le 0$; $\pm$ is two standard deviations. The t-percentiles c_n were approximated using (26.7.8) of Abramowitz and Stegun (1964).

for comparing three treatments. In addition, Coad and Woodroofe (1997) have used similar ideas to construct confidence intervals for the ratio of two hazard rates, following a sequential test.

5. THE FORD–SILVEY EXAMPLE

In the Ford–Silvey example, Eq. (1) and Example 2 of Sec. 2, $N=a=n$,

$$X_n'X_n=\begin{pmatrix} n & s_n \\ s_n & n \end{pmatrix}$$

$$B_n=\frac{1}{\sqrt{n}}\begin{pmatrix} \sqrt{n^2-s_n^2} & s_n \\ 0 & n \end{pmatrix}$$

and

$$\eta(\theta) = \lim_{n\to\infty} n(X_n' X_n)^{-1} = \frac{1}{1-\kappa^2}\begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix}$$

if $\kappa \neq 1$, where $s_n = x_1 + \cdots + x_n$, $\kappa = \theta_1/\theta_2$ if $|\theta_1| \leq \theta_2$, and $\kappa = \theta_2/\theta_1$ if $0 < \theta_2 < |\theta_1|$.

5.1. Confidence Intervals for θ_1 and θ_2

To find a confidence interval for θ_1, let $\Gamma = (1, 0)$ in Eq. (24). Then $A_n = (1, 0)$,

$$W_n^o = \sqrt{\frac{n^2 - s_n^2}{n}}(\theta_1 - \hat{\theta}_{n,1})$$

and

$$Q(\theta) = \frac{\Gamma\eta(\theta)}{\sqrt{\Gamma\eta(\theta)\Gamma'}} = \frac{1}{\sqrt{1-\kappa^2}}[1, \kappa]$$

In the region $|\theta_1| < \theta_2$, $\kappa = \theta_1/\theta_2$, $Q(\theta) = (\theta_2, \theta_1)/\sqrt{\theta_2^2 - \theta_1^2}$, and, therefore, $Q^{\#}(\theta) = \theta_1\theta_2(1, -1)/\sqrt{(\theta_2^2 - \theta_1^2)^3}$, so that

$$\mu_n(\theta) = -\frac{1}{\sqrt{n}} Q^{\#}(\theta)\mathbf{1} = 0$$

and

$$\Delta^o(\theta) = 0$$

Similarly, in the region $0 < \theta_2 < |\theta_1|$, $\kappa = \theta_2/\theta_1$, $Q(\theta) = (\theta_1, \theta_2)/\sqrt{\theta_1^2 - \theta_2^2}$, and $Q^{\#}(\theta) = (-\theta_2^2, \theta_1^2)/\sqrt{(\theta_1^2 - \theta_2^2)^3}$, so that

$$\mu_n(\theta) = -\frac{1}{\sqrt{n(\theta_1^2 - \theta_2^2)}}$$

and

$$\Delta^o(\theta) = \frac{1}{\theta_1^2 - \theta_2^2}$$

Thus, if c is the upper $100\alpha/2$ percentile of the standard normal distribution and if $C = [-c, c]$ in Eq. (23), then

$$\mathcal{C} = \left\{\theta : -\sqrt{\frac{n}{n^2 - s_n^2}}\left[c\left(1 + \frac{\hat{\Delta}_n^o}{2n}\right) - \hat{\mu}_n^o\right] \le \theta_1 - \hat{\theta}_{n,1} \le \sqrt{\frac{n}{n^2 - s_n^2}} \times \left[c\left(1 + \frac{\hat{\Delta}_n^o}{2n}\right) + \hat{\mu}_n^o\right]\right\}$$

where $\hat{\mu}_n^o$ and $\hat{\Delta}_n^o$ denote estimators of $\mu_n(\theta)$ and $\Delta^o(\theta)$, possibly smoothed as in Example 3 of Sec. 4.

A similar procedure may be used to find approximate confidence intervals for θ_2. Letting $\Gamma = (0, 1)$ in Eq. (24), leads to

$$W_n^o = \sqrt{\frac{n^2 - s_n^2}{n}}(\theta_2 - \hat{\theta}_{n,2})$$

$$\mu_n(\theta) = \begin{cases} -1\Big/\sqrt{n(\theta_2^2 - \theta_1^2)} & \text{if } |\theta_1| < \theta_2 \\ 0 & \text{if } 0 < \theta_2 < |\theta_1| \end{cases}$$

$$\Delta^o(\theta) = \begin{cases} 1/(\theta_2^2 - \theta_1^2) & \text{if } |\theta_1| < \theta_2 \\ 0 & \text{if } 0 < \theta_2 < |\theta_1| \end{cases}$$

and

$$\mathcal{C} = \left\{\theta : -\sqrt{\frac{n}{n^2 - s_n^2}}\left[c\left(1 + \frac{\hat{\Delta}_n^o}{2n}\right) - \hat{\mu}_n^o\right] \le \theta_2 - \hat{\theta}_{n,2} \le \sqrt{\frac{n}{n^2 - s_n^2}} \times \left[c\left(1 + \frac{\hat{\Delta}_n^o}{2n}\right) + \hat{\mu}_n^o\right]\right\}$$

Monte Carlo estimates of the actual coverage probabilities of these confidence intervals are reported in Table 4. It appears that the approximation is excellent, except on the lines $|\theta_1| = \theta_2$, where it is conservative. The theoretical justification for the procedure fails on these lines, as is evident from the infinite discontinuities in $\mu_n^o(\theta)$ and $\Delta^o(\theta)$.

5.2. Confidence Intervals for θ_1/θ_2

Since the Ford–Silvey example was motivated by the problem of finding a good estimator for $\upsilon = \theta_1/\theta_2$, it is natural to consider

Table 4. Coverage Probabilities for the Ford–Silvey Example

		$n=25$		$n=50$	
θ_1	θ_2	$\alpha=0.05$	$\alpha=0.10$	$\alpha=0.05$	$\alpha=0.10$
1.0	1.0	.958	.912	.956	.909
		.956	.912	.953	.907
1.0	1.5	.954	.901	.952	.901
		.948	.897	.950	.899
1.0	2.0	.948	.896	.949	.898
		.948	.902	.951	.901
1.0	4.0	.950	.902	.950	.900
		.949	.899	.952	.899
1.5	1.0	.950	.902	.948	.898
		.949	.897	.947	.900
2.0	1.0	.948	.896	.948	.897
		.946	.897	.951	.903
2.0	2.0	.957	.911	.952	.901
		.952	.905	.952	.903
4.0	1.0	.950	.901	.949	.898
		.948	.896	.948	.898
4.0	4.0	.953	.904	.949	.902
		.953	.904	.949	.900
±		.0044	.006	.0044	.006

Note: Based on 10,000 replications; ± is two standard deviations. The upper figure is the coverage probability for θ_1, and the lower is for θ_2.

$\hat{\upsilon}_n = \hat{\theta}_{n,1}/\hat{\theta}_{n,2}$ in some detail. In some ways, approximations to the distribution of $\hat{\upsilon}_n$ are nicer than those for $\hat{\theta}_{n,1}$ and $\hat{\theta}_{n,2}$. The bias of $\hat{\upsilon}_n$ may be computed, as in Sec. 2, and

$$E_\theta(\hat{\upsilon}_n - \upsilon) \approx \begin{cases} 0 & \text{if } |\theta_1| < \theta_2 \\ \theta_1/n\theta_2^3 & \text{if } 0 < \theta_2 < |\theta_1| \end{cases} \tag{31}$$

Again the approximation is discontinuous when $|\theta_1| = \theta_2$, and the theoretical justification fails in this case.

To form confidence intervals, it is necessary to compute the probability that $\upsilon - \hat{\upsilon}_n \le b_n$, where b_n may depend on the data. Assuming that $\theta_2 > 0$, the inequality may be rewritten as

$$\theta_1 - (b_n + \hat{\upsilon}_n)\theta_2 \le 0 \tag{32}$$

Letting $\kappa_n = -s_n/n$, Eq. (32) may be written in the form

$$a_n' Z_n \le c \tag{33}$$

where

$$a_{n,1} = \frac{1 - (b_n + \hat{\upsilon}_n)\kappa_n}{\sqrt{1 - 2(b_n + \hat{\upsilon}_n)\kappa_n + (b_n + \hat{\upsilon}_n)^2}}$$

$$a_{n,2} = -\frac{\sqrt{(1 - \kappa_n^2)}(b_n + \upsilon_n)}{\sqrt{1 - 2(b_n + \hat{\upsilon}_n)\kappa_n + (b_n + \hat{\upsilon}_n)^2}}$$

and

$$c = \frac{\sqrt{n(1 - \kappa_n^2)}\, b_n \hat{\theta}_{n,2}}{\sqrt{1 - 2\kappa_n(b_n + \hat{\upsilon}_n) + (b_n + \hat{\upsilon}_n)^2}} \tag{34}$$

The probability of Eq. (33) may be approximated using Eq. (18). To do so, it is necessary to find the limit of $Q^{(n)} = \sqrt{n} a_n' B_n^{-1}$. First observe that if c remains bounded in Eq. (34), then $b_n \to 0$ in probability as $n \to \infty$. It follows easily that

$$\begin{aligned} Q^{(n)} &\to \frac{[1 - \upsilon\kappa, -\sqrt{(1 - \kappa^2)}\upsilon]}{\sqrt{(1 - 2\kappa\upsilon + \upsilon^2)(1 - \kappa^2)}} \begin{pmatrix} 1 & \kappa \\ 0 & \sqrt{1 - \kappa^2} \end{pmatrix} \\ &= \frac{[1 - \upsilon\kappa, \kappa - \upsilon]}{\sqrt{(1 - 2\kappa\upsilon + \upsilon^2)(1 - \kappa^2)}} = Q(\theta), \quad \text{say} \end{aligned}$$

in probability, if $\upsilon \ne 1$. If $|\upsilon| < 1$, then $\kappa = \upsilon$ and $Q(\theta) = [1, 0]$, and if $|\upsilon| > 1$, then $\kappa = 1/\upsilon$ and $Q(\theta) = [0, -1]$. In either case, $Q(\theta)$ does not depend on θ, and therefore the probability of Eq. (33) is

$$P_\theta\{a_n' Z_n \le c\} = \Phi(c) + o(1/n)$$

in the very weak sense. Let c be the upper $100\alpha/2$ percentile of the standard normal distribution, let b_n^+ be the positive solution to Eq. (34), and let b_n^- be the negative solution to Eq. (34) with c replaced by $-c$. Then

$$P_\theta\{b_n^- \le \upsilon - \hat{\upsilon}_n \le b_n^+\} \approx 1 - \alpha \tag{35}$$

in the very weak sense.

Table 5. Bias of $\hat{v}_n$ in the Ford–Silvey Example

		$n=25$		$n=50$	
θ_1	θ_2	Approx.	M.C.	Approx.	M.C.
1.0	1.0	.040	.034	.020	.014
1.0	2.0	.000	−.002	.000	.000
1.0	4.0	.000	.000	.000	.000
2.0	1.0	.080	.098	.040	.044
2.0	2.0	.010	.088	.006	.004
4.0	1.0	.160	.196	.080	.088
4.0	4.0	.002	.002	.002	.000

Coverage Probabilities for v

		$n=25$		$n=50$	
θ_1	θ_2	$\alpha=.05$	$\alpha=.10$	$\alpha=.05$	$\alpha=.10$
1.0	1.0	.945	.888	.948	.896
1.0	1.5	.946	.899	.953	.904
1.0	2.0	.946	.895	.947	.900
1.0	4.0	.950	.899	.950	.900
1.5	1.0	.949	.894	.952	.903
2.0	1.0	.950	.895	.952	.901
2.0	2.0	.947	.894	.950	.899
4.0	1.0	.950	.895	.951	.900
4.0	4.0	.946	.892	.952	.904
±		.0044	.006	.0044	.006

Note: Based on 10,000 replications; ± is two standard deviations.

As above, the theoretical justification for Eq. (35) fails on the lines $|\theta_1| = \theta_2$, but now the approximations compare well with simulated values, even on these lines. Monte Carlo estimates of the bias and coverage probabilities are provided in Table 5.

6. AUTOREGRESSIVE PROCESSES

Consider an autoregressive process of order two, say

$$y_k = \theta_1\, y_{k-1} + \theta_2\, y_{k-2} + \sigma\varepsilon_k, \quad k = 1, 2, \ldots,$$

with $y_{-1} = y_0 = 0$. Here $\sigma > 0$ and $\theta = (\theta_1, \theta_2)'$ is confined to a triangular region Ω in which $\theta_1 + \theta_2 < 1$, $\theta_1 - \theta_2 > -1$, and $|\theta_2| < 1$ (see, for example,

Brockwell and Davis (1991, Chap. 8). Clearly, this model is of the form of Eq. (2), and

$$X_n'X_n = \begin{pmatrix} r_{n,0} & r_{n,1} \\ r_{n,1} & r_{n-1,0} \end{pmatrix}$$

where

$$r_{n,j} = \sum_{k=2}^{n} y_{k-1}y_{k-1-j}$$

for $n \geq 3$. It may be shown that

$$\eta(\sigma,\theta) = \lim_{n\to\infty} n(X_n'X_n)^{-1} = \frac{1}{\sigma^2}\begin{pmatrix} 1-\theta_2^2 & -\theta_1(1+\theta_2) \\ -\theta_1(1+\theta_2) & 1-\theta_2^2 \end{pmatrix}$$

*w.p.*1 for all $\sigma > 0$ and $\theta \in \Omega$. Using Eq. (8), it then follows that

$$E_{\sigma,\theta}(\hat{\theta}_n - \theta) \approx -\frac{1}{n}\begin{pmatrix} \theta_1 \\ 3\theta_2 + 1 \end{pmatrix} \tag{36}$$

and Eq. (36) agrees well with Coad and Woodroofe's (1998) simulations.

To form confidence intervals for θ_1, say, let $\Gamma = (1, 0)$ in Eq. (24). Then

$$G_n = \sqrt{r_{n,0} - \frac{r_{n,1}^2}{r_{n-1,0}}}$$

$$W_n = \frac{G_n}{\hat{\sigma}_n}(\theta_1 - \hat{\theta}_{n,1})$$

and

$$Q(\sigma,\theta) = \frac{\sqrt{1-\theta_2^2}}{\sigma}\left[1, -\frac{\theta_1}{1-\theta_2}\right]$$

Differentiation then yields

$$\mu_n(\sigma,\theta) = \frac{\theta_1}{(1-\theta_2)\sqrt{n(1-\theta_2^2)}}$$

and

$$\Delta(\sigma,\theta) = \frac{\theta_1^2 + 2\theta_2(1-\theta_2)^2(1+\theta_2)}{(1-\theta_2)^3(1+\theta_2)}$$

Observe that μ_n and Δ do not depend on σ. Let $\tilde{\theta}_{n,1} = \hat{\theta}_{n,1}$, $\tilde{\theta}_{n,2} = \hat{\theta}_{n,2}$ if $|\tilde{\theta}_{n,2}| \leq 1 - 1/n$, $\tilde{\theta}_{n,2} = \pm(1 - 1/n)$ otherwise, $\hat{\mu}_n = \mu_n(1, \tilde{\theta}_n)$, and $\hat{\Delta}_n = \Delta(1, \tilde{\theta}_n)$. Then confidence intervals for θ_1 may be formed as in Eq. (30), except that now there are $n - 2$ degrees of freedom.

The procedure is similar for θ_2. Letting $\Gamma = (0, 1)$ in Eq. (24), $W_n = G_n(\theta_2 - \hat{\theta}_{n,2})/\hat{\sigma}_n$, where

$$G_n = \sqrt{r_{n-1,0} - \frac{r_{n,1}^2}{r_{n,0}}}$$

Then

$$Q(\sigma, \theta) = \frac{\sqrt{1 - \theta_2^2}}{\sigma}\left[-\frac{\theta_1}{1 - \theta_2}, 1\right]$$

$$\mu_n(\sigma, \theta) = \frac{1 + 2\theta_2}{\sqrt{n(1 - \theta_2^2)}}$$

and

$$\Delta(\sigma, \theta) = \frac{1 + 2\theta_2 + 2\theta_2^2}{1 - \theta_2^2}$$

Table 6. Coverage Probabilities for Autoregressive Processes

		$n = 25$		$n = 50$		$n = 100$	
θ_1	θ_2	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$
0.0	0.0	.949	.897	.949	.897	.957	.907
		.949	.897	.949	.902	.955	.906
0.0	0.5	.950	.900	.949	.897	.953	.906
		.952	.897	.954	.904	.953	.902
1.5	−0.6	.948	.895	.947	.898	.956	.905
		.946	.896	.948	.896	.957	.908
1.9	−0.9	.949	.895	.948	.891	.948	.899
		.942	.880	.942	.883	.948	.895
±		.0044	.006	.0044	.006	.0044	.006

Note: Based on 10,000 replications; ± is two standard deviations. The upper figure is the coverage probability for θ_1, and the lower is for θ_2.

Estimators of μ_n and Δ may be constructed as above, and approximate confidence intervals of the form of Eq. (30) may also be constructed.

A simulation study was conducted to assess the accuracy of the approximation. For $n=25$ and 50, the simulated and nominal values of the coverage probability agree well, except for θ_2 near one of the vertices of the triangle Ω. For $n=100$, the agreement is still very good, but the nominal values may be slightly too conservative. Some representative values are included in Table 6.

7. FORMALITIES

The theoretical justification for the expansions in Secs. 2–4 is in the very weak sense of Woodroofe (1986). Let $\Omega \subseteq \Re^p$ denote the parameter space, and suppose that Ω is a convex open set or a union of a countable collection of such sets, as in Coad and Woodroofe (1998). Further, let $\gamma_a(\theta)$ be a function of the unknown parameter, like a coverage probability or rescaled bias, and let $\gamma(\theta)$ be a candidate for a limiting function. If $q > 0$, then the notation

$$\gamma_a(\theta) = \gamma(\theta) + o(a^{-q}) \quad \text{very weakly} \tag{37}$$

means that

$$\lim_{a\to\infty} a^q \int_\Omega [\gamma_a(\theta) - \gamma(\theta)]\xi(\theta)\,\mathrm{d}\theta = 0 \tag{38}$$

for all sufficiently smooth compactly supported prior densities ξ on Ω. The precise amount of smoothness required of ξ may depend on q. It is argued below that very weak approximation is strong enough to support the frequentist interpretation of confidence. An example in which $\gamma_a(\theta) = \gamma(\theta) + o(a^{-1})$ holds in the very weak sense, but not in the conventional sense, is also described below.

Letting $\|\cdot\|$ denote the trace norm for matrices, the conditions under which Eq. (18) holds may now be stated quite simply: *If there are matrices* $Q(\theta) = [q_{ij}(\theta) : i = 1,\ldots,m, j = 1,\ldots,p]$ *for which* q_{ij} *are twice continuously differentiable on* Ω*, if*

$$\lim_{a\to\infty} \int_K E_\theta \|Q^{(a)} - Q(\theta)\|^2\,\mathrm{d}\theta = 0 \tag{39}$$

for all compact $K \subseteq \Omega$, and if $\approx$ is interpreted to mean equality up to $o(a^{-1})$ in the very weak sense, then Eq. (18) holds for all measurable, symmetric (sign invariant) functions $h : \Re^m \to \Re$ of quadratic growth; and if also

$$\lim_{a\to\infty} \left\| \sqrt{a} \int_K E_\theta[Q^{(a)} - Q(\theta)]\, d\theta \right\| = 0 \tag{40}$$

for all compact $K \subseteq \Omega$, then Eq. (18) holds for all measurable h of quadratic growth. These assertions are the corollary to Theorem 1 in Woodroofe and Coad (1997), and there is substantial uniformity with respect to h in that theorem that is not reported here. The condition in Eq. (39) is not restrictive.

Surprisingly, less smoothness is required for the corrected confidence sets than for Eq. (18). With $Q(\theta)$ as in Eq. (16), suppose that q_{ij} are once differentiable on Ω and that

$$\int_K [\|Q^{\#}(\theta)\mathbf{1}\|^2 + \|\Delta^o(\theta)\|]\, d\theta < \infty \tag{41}$$

for all compact $K \subseteq \Omega$. In words, $\|Q^{\#}(\theta)\mathbf{1}\|^2$ and $\|\Delta^o(\theta)\|$ must be locally integrable in θ. Let $\mathcal{C} = \mathcal{C}_a$ be a confidence set of the form of Eq. (23), and let

$$\gamma_a(\theta) = P_\theta\{\theta \in \mathcal{C}_a\} = P_\theta\{W_N^{0*} \in C\}$$

for $\theta \in \Omega$ and $a \geq a_0$. *If Eqs. (39) and (41) hold, then there exist estimators $\hat{\mu}_a^o$ and $\hat{\Delta}_a^o$ for which*

$$\gamma_a(\theta) = \Phi^m(C) + o\left(\frac{1}{a}\right) \tag{42}$$

in the very weak sense for all measurable, symmetric (sign invariant) subsets $C \subseteq \Re^m$; and if Eq. (40) holds too, then Eq. (42) holds for all measurable $C \subseteq \Re^m$. These assertions follow from Theorem 2 and Proposition 3 of Woodroofe and Coad (1997). As stated, they are unsatisfactory in that only the existence of estimators $\hat{\mu}_a^o$ and $\hat{\Delta}_a^o$ for which Eq. (42) holds is claimed. However, if $Q^{\#}(\theta)$ and $\Delta^o(\theta)$ are bounded and continuous, then it is sufficient to use $\hat{\mu}_a^o = \mu_a(\hat{\theta}_N)$ and $\hat{\Delta}_a^o = \Delta^o(\hat{\theta}_N)$.

Similar conditions are sufficient for Eqs. (8) and (9). These are included in Sec. 8, along with an outline of the proof of Eq. (8).

Very weak approximations are strong enough to support a frequentist interpretation of confidence. To see why, consider the case of known σ, let $\mathcal{C}_a$ denote a confidence set of the form of Eq. (23), and suppose that this procedure is put into routine use. If the procedure is used by a sequence of

clients, then it seems reasonable to suppose that the values of θ will vary from client to client. If these values are drawn from a density ξ, say, then the long run relative frequency of coverage is

$$\bar{\gamma}_a(\xi) = \int_\Omega \gamma_a(\theta)\xi(\theta)\,d\theta \tag{43}$$

Thus, in order to have a valid confidence procedure, it is enough to have $\bar{\gamma}_a(\xi)$ approximate a nominal value, and this is precisely the meaning of Eq. (37), assuming only that ξ is smooth and compactly supported. It is amusing to contrast the use of ξ here with conventional Bayesian uses. Here ξ has a clear frequentist interpretation. However, it is unknown to any given client and may be unknowable, since estimating ξ would require access to others' data sets and, even then, there is only indirect information about ξ.

Woodroofe and Keener (1987) developed an example in which Eq. (18) holds in a very weak sense, but not in a conventional one.

Example 4. Let $y_1, y_2, \ldots$ be independent and normally distributed with unknown mean θ and unit variance (so that $p=1$ and $x_k=1$ for all $k=1,2,\ldots$). Suppose that θ is known to be positive and let

$$N = N_a = \inf\{n \geq 1 : y_1 + \cdots + y_n > a\}$$

for $a \geq 1$. This corresponds roughly to a one-sided sequential probability ratio test. Then $X_N' X_N = N$,

$$\lim_{a\to\infty} \frac{a}{N} = \theta$$

and Eq. (18) holds in the very weak sense with $Q(\theta) = \sqrt{\theta}$, $\theta > 0$. However, Eq. (18) does not hold in the conventional sense. In fact,

$$P_\theta[Z_N^o \leq c] = \Phi(c) + \frac{1}{\sqrt{a\theta}} R_a(\theta, c) + o\left(\frac{1}{\sqrt{a}}\right)$$

for fixed $\theta > 0$, where $R_a(\theta, c)$ is bounded but oscillates wildly as a increases. The exact expression for R_a is complicated, owing to the presence of ladder variables, and is not reproduced here. □

8. OUTLINE OF A PROOF

Suppose that Ω is a convex open set or the union of a countable collection of convex open sets. If $\eta(\sigma, \theta)$ is continuously differentiable in θ for fixed σ and

$$\lim_{a\to\infty}\int_K E_{\sigma,\theta}\|a(X_N'X_N)^{-1}-\eta(\sigma,\theta)\|\,\mathrm{d}\theta=0 \tag{44}$$

for all compact subsets $K\subseteq\Omega$, then Eq. (8) holds in the very weak sense. For Eq. (9) it is sufficient that $\rho(\sigma^2,\theta)$ be differentiable in σ^2, that $\rho'(\sigma^2,\theta)$ and the derivatives of η be continuous in (σ^2,θ), that Eq. (44) holds for each fixed σ, and that

$$\lim_{a\to\infty}\int_K E_{\sigma,\theta}\left|\frac{a}{N}-\rho(\sigma^2,\theta)\right|d\sigma^2\,d\theta=0$$

for all compact subsets $K\subseteq(0,\infty)\times\Omega$.

The proof of Eq. (8) is outlined next. For simplicity, σ is assumed to be known, say $\sigma=1$, and is omitted from the notation. The meaning of Eq. (8) is that

$$\int_\Omega\left[E_\theta(\hat\theta_N-\theta)-\frac{1}{a}\eta^{\#}(\theta)\mathbf{1}\right]\xi(\theta)\,\mathrm{d}\theta=o\left(\frac{1}{a}\right) \tag{45}$$

as $a\to\infty$ for all continuously differentiable densities ξ with compact support. The first step in the proof is to write

$$\int_\Omega E_\theta(\hat\theta_N-\theta)\xi(\theta)\,\mathrm{d}\theta=E_\xi(\hat\theta_N-\theta)$$

where E_ξ denotes expectation in a Bayesian model in which θ has prior distribution ξ and Eq. (2) holds conditionally given θ. Then

$$E_\xi(\hat\theta_N-\theta)=E_\xi[E_\xi^N(\hat\theta_N-\theta)]$$

where E_ξ^N denotes conditional (posterior) expectation given the data. The next step is to approximate $E_\xi^N(\hat\theta_N-\theta)$. Let L_N denote the likelihood function,

$$L_N(\theta)=\exp\left[-\frac{1}{2}(\theta-\hat\theta_N)'(X_N'X_N)(\theta-\hat\theta_N)\right]$$

Then

$$E_\xi^N(\theta-\hat\theta_N)=\frac{1}{c}\int_\Omega(\theta-\hat\theta_N)L_N(\theta)\xi(\theta)\,\mathrm{d}\theta$$

where c is a normalizing constant, $c=\int_\Omega L_N(\theta)\xi(\theta)\,\mathrm{d}\theta$. Multiplying by $(X_N'X_N)$ and the integrating by parts,

$$(X_N' X_N) E_\xi^N(\hat{\theta}_N - \theta) = \frac{1}{c}\int_\Omega \nabla L_N(\theta)\xi(\theta)\,d\theta$$
$$= -\frac{1}{c}\int_\Omega L_N(\theta)\nabla\xi(\theta)\,d\theta = -E_\xi^N\left[\frac{\nabla\xi}{\xi}(\theta)\right]$$

where ∇ denotes gradient with respect to θ. Thus, using Eq. (44),

$$aE_\xi^N(\hat{\theta}_N - \theta) = -a(X_N' X_N)^{-1} E_\xi^N\left[\frac{\nabla\xi}{\xi}(\theta)\right] \to -\eta(\theta)\frac{\nabla\xi}{\xi}(\theta)$$

and

$$aE_\xi(\hat{\theta}_N - \theta) \to \int_\Omega -\eta(\theta)\frac{\nabla\xi}{\xi}(\theta)\xi(\theta)\,d\theta = \int_\Omega \eta^{\#}(\theta)\mathbf{1}\xi(\theta)\,d\theta \tag{46}$$

where the final equality follows from another integration by parts. This completes the proof, because Eq. (46) is equivalent to Eq. (45). □

The proofs of Eqs. (9) and (18) use similar ideas, but the details of the integration by parts are more complicated. Justification for the corrected confidence sets combines these ideas with a Taylor series expansion.

REFERENCES

1. Abramowitz, M.; Stegun, I.A. *Handbook of Mathematical Functions*; US Department of Commerce, 1964.
2. Berger, J.O.; Wolpert, R. *The Likelihood Principle*; Institute of Mathematical Statistics: Haywood, CA, 1984.
3. Betensky, R. An O'Brian-Fleming Sequential Trial for Comparing Three Treatments. Ann. Stat. **1996**, *24*, 1765–1791.
4. Brockwell, P.J.; Davis, R.A. *Time Series*: *Theory and Methods*; Springer: New York, 1991.
5. Coad, D.S. Sequential Allocation with Data-Dependent Allocation and Time Trends. Sequential Anal. **1991**, *10*, 91–97.
6. Coad, D.S. Sequential Allocation Rules for Multi-Armed Clinical Trials. J. Stat. Comput. Simulation **1995**, *52*, 239–251.
7. Coad, D.S.; Woodroofe, M. Approximate Confidence Intervals After a Sequential Clinical Trial Comparing Two Exponential Survival Curves with Censoring. J. Stat. Plann. Inference **1997**, *63*, 79–96.
8. Coad, D.S.; Woodroofe, M. Approximate Bias Calculations for Sequentially Designed Experiments. Sequential Anal. **1998**, *17*, 1–31.

9. Eisele, J.R. The Doubly Adaptive Biased Coin Design for Sequential Clinical Trials. J. Stat. Plann. Inference **1994**, *38*, 249–262.
10. Ford, I.; Silvey, S.D. A Sequentially Constructed Design for Estimating a Nonlinear Parametric Function. Biometrika **1980**, *67*, 381–388.
11. Ford, I.; Titterington, D.M.; Wu, C.F.J. Inference and Sequential Design. Biometrika **1985**, *72*, 545–551.
12. Hayre, L.S.; Gittins, J.C. Sequential Selection of the Larger of Two Normal Means. J. Am. Stat. Assoc. **1981**, *76*, 696–700.
13. Lai, T.L.; Wei, C.Z. Least Squares Estimates in Stochastic Regression Models with Applications to Identification and Control of Dynamic Systems. Ann. Stat. **1982**, *10*, 154–166.
14. Robbins, H.; Siegmund, D. Sequential Tests Involving Two Populations. J. Am. Stat. Assoc. **1974**, *69*, 132–139.
15. Siegmund, D.O. A Sequential Clinical Trial for Comparing Three Treatments. Ann. Stat. **1993**, *21*, 464–483.
16. Woodroofe, M. Very Weak Expansions for Sequential Confidence Levels. Ann. Stat. **1986**, *14*, 1049–1067.
17. Woodroofe, M. Very Weak Expansions for Sequentially Designed Experiments: Linear Models. Ann. Stat. **1989**, *17*, 1087–1102.
18. Woodroofe, M.; Coad, D.S. Corrected Confidence Sets for Sequentially Designed Experiments. Stat. Sinica **1997**, *7*, 53–74.
19. Woodroofe, M.; Keener, R. Asymptotic Expansions in Boundary Crossing Probabilities. Ann. Probab. **1987**, *15*, 102–114.
20. Wu, C.F.J. Asymptotic Inference from a Sequential Design in a Nonlinear Situation. Biometrika **1985**, *72*, 553–558.